Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

Jade Caines 10/30/11

Using Rasch Measurement Theory to Validate the Student Performance Character and Student Moral Character Scales

By

Jade Caines

B.A., Stanford University

M.A.T., Brooklyn College

_____

Yuk Fai Cheong

Advisor

_____

George Engelhard, Jr.

Committee Member

_____

Robert Jensen

Committee Member

Accepted:

_____

Lisa A. Tedesco, Ph.D.

Dean of the James T. Laney School of Graduate Studies

_____

Date

Using Rasch Measurement Theory to Validate the Student Performance Character and Student Moral Character Scales

By

Jade Caines

B.A., Stanford University

M.A.T., Brooklyn College

Advisor: Yuk Fai Cheong, Ph.D

An abstract of

A dissertation submitted to the Faculty of the James T. Laney School of Graduate Studies of Emory University in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Educational Studies

2011

Abstract

According to Bond & Fox (2001), without deliberate, thoughtful, and scientific construction of measures, psychosocial research will progress slowly. In the field of character education, Davidson , Khmelkov, and Lickona (2010) contend that assessment should be a high priority for practitioners. Heeding their advice, this study further investigates the validity and gender invariance of two existing scales that measure two types of character within students: S*tudent Performance Character* (SPC) and *Student Moral Character* (Cornerstone Consulting & Evaluation, 2009). Using data collected on 239 middle-school students through a character education evaluation, this study addresses three research questions:

1. How well do items on the Student Performance Character and Student Moral Character scales measure the two constructs?
2. Do the items on the Student Performance Character and Student Moral Character scales fit the Rasch measurement model framework? If so, how well is the fit?
3. Do any of the items display gender differential item functioning (DIF)? If so, what are their patterns, directions, and magnitudes?

Results indicate mixed results for both scales regarding rating scale diagnostics, model-data fit and item fit to constructs. Overall, the SPC scale shows a better fit to the Rasch model framework than the SMC scale; principal component analyses suggest multidimensionality for the SMC scale. Finally, no gender DIF is evident. Implications of the results for scale revisions and theoretical advancement are discussed.

Using Rasch Measurement Theory to Validate the Student Performance Character and Student Moral Character Scales

By

Jade Caines

B.A., Stanford University

M.A.T., Brooklyn College

Advisor: Yuk Fai Cheong, Ph.D

A dissertation submitted to the Faculty of the James T. Laney School of Graduate Studies of Emory University in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Educational Studies

2011

TABLE OF CONTENTS

"Without deliberate, thoughtful, and scientific construction of measures, psychosocial research will continue to progress slowly" (Bond & Fox, 2001).

CHAPTER I

Researchers have asserted that character plays an important role in academic life. For example, Davidson, Lickona, and Khmelkov (2008) state four important roles for character in academic life: (a) students need initiative, self-discipline, perseverance, etc. in order to do their best academically; (b) students develop their character from working hard and smart on academic work; (c) students need moral character in order to create the relationships that positively contribute to the learning environment, and (d) students develop moral character from their schoolwork (e.g., helping peers through group work, service learning projects, etc.). Essentially, these four roles demonstrate how different types of character can have a transformational impact on students' academic work.

First, students need initiative, self-discipline, and perseverance in order to succeed academically. For example, Berger (2003) claims that when students begin to make discoveries that impress their classmates, solve problems as part of the group, create projects that are admired by others, and produce quality work, a new self-image as a proud student emerges:

> Once a student sees that he or she is capable of excellence, that student is never quite the same. There is a new self-image, a new notion of possibility. There is an appetite for excellence. After students have had a taste of excellence, they're never quite satisfied with less; they're always hungry. (p. 8)

Working hard and smart on academic work fosters this hunger within students. More specifically, when a teacher sets an expectation of excellence, students may have to rewrite an assignment several times before submitting a final draft. Demonstrating this perseverance and

hard work ethic will not only help them obtain a higher grade on the task, but it will carry them throughout their academic career:

> In Mose Durst's junior high school class, learning to write means learning to rewrite. He makes a copy of each student's first draft for every student in the class. Together, the class identifies the strengths of each paper and areas for improvement. He works with students not only on getting grammar and punctuation right but also on style—on varying their sentence structure so as to produce syntactically pleasing sentences. (Lickona, 2004, p. 126)

While students dedicate themselves to the process, they are developing a work ethic that can improve the quality of their work.

Moral character also plays a large role in students' academic lives. Davidson, Lickona, and Khmelkov (2008) claim that students need moral character in order to create the relationships that positively contribute to the learning environment, and develop moral character from their schoolwork. For example, when students engage in service-learning projects, they develop moral character qualities such as empathy, justice, and altruism (Lickona & Davidson, 2005). Also, school communities can help develop these characteristics by emphasizing concern for others. Students *and* teachers can help each other achieve their best by holding one another accountable:

> "Concern in our society means, 'Help me do what *I* want to do. If I am feeling bad, you are supposed to sympathize…' [However, this] isolates us from one another. [I have a responsibility to] say or do what I truly believe is in your best interests, regardless of how it may make you feel about me." (Gauld, 1995, pp. 28-29)

In other words, concern is a moral characteristic and requires, within school communities, students and teachers to challenge one another to achieve their best. This moral principle can have a positive influence on the quality of schoolwork, thereby improving students' academic lives greatly. Essentially, academic development is an integral part of character development; you don't get one without the other (Gauld, 1995).

Additionally, character matters, not just for its potential impact on academic achievement, but because it helps students be better people and live fulfilling lives. Davidson et al. (2008) argue that developing students' character is essential, separate from academic achievement goals:

> The development of character is a worthy pursuit in its own right, not simply for the other desired outcomes it can bring to a school (e.g., academic achievement…). We believe in the importance of character in all phases of life. From this perspective, the most important goal of character education is to prepare all young people to lead a flourishing life. (p. 371-372)

A flourishing life is antithetical to delinquent behaviors. As a result, good character can be a deterrent to negative behaviors that may lead to detrimental lifestyles. According to Search Institute, there are eight categories that promote positive attitudes within youth: support, empowerment, boundaries and expectations, constructive use of time, commitment to learning, positive values, social competencies, and positive identity (Lickona, 2004). Youth with the highest amount of these assets are the least likely to engage in high-risk, negative behaviors (such as illegal drug/alcohol use and violence).

The school community is an ideal place for students to develop these assets. As a result, character education programs and schools that promote a character culture permeate our

country's landscape. Approaches to teaching character, however, whether direct or indirect, vary widely. Dewey (1975) argues that the nature of the school community is a far more important factor in students' moral development than direct moral instruction. Also, Ryan (1986) argues that there is a hidden curriculum that manifests within schools and classrooms: "Very little of the moral education that inevitably occurs in the schools is formally recorded in lesson plans, curriculum guides, or behavioral objectives" (p. 228). Rather, students develop an awareness of morality by examining what rules are or are not enforced and the rituals of daily classrooms. Whether a school community chooses direct or indirect approaches to character education, it is evident that character matters in schools; therefore, a rigorous evaluation of schools that promote character education in necessary.

Since 1998, the Character Education Partnership (CEP) has been recognizing schools and school districts that promote good character in students. In 2011, CEP named 43 schools and 1 school district *National Schools of Character* for excellence in student character development (http://www.character.org/uploads/PDFs/Press_Releases/NSOC/PressRelease_NSOC2011.pdf). These school communities meet the standards of excellence established by the 11 Principles of Effective Character Education (Character Education Partnership, 2010). They have implemented character education programs deemed successful by a national organization and serve as models for how character education can positively influence students' futures (generally) and academic lives (specifically).

Statement of the Problem

In 2005, Berkowitz and Bier identified only 54 character education programs nationwide that had *any* type of research study conducted. Using a scoring rubric, they reported that 78 studies were scientifically acceptable, leaving a pool of only 39 programs with at least one

adequate research study evaluating it and only 33 programs that reported scientific evidence supporting their effectiveness in promoting character development in students: “It is worth repeating that there are likely many more effective character education programs that do not yet have scientific research to demonstrate their effectiveness” (p. 7).

Also, the United States Department of Education’s Office of Safe and Drug-Free-Schools (US DOE) has increased methodological standards for educational research over the past several years. Stricter standards include the development and utilization of instruments that measure program effectiveness. This presents a challenge for the character education field; instruments measuring character elements in K-12 students often lack rigorous/appropriate validity and reliability testing (United States Department of Education Office of Safe and Drug-Free Schools, 2007).

Furthermore, some scholars in the field of character education have expressed concerns that scales and instruments measuring character and reasoning are not valid and reliable across various social groups (Carlo, McGinley, Roesch, & Kaminski, 2008; Lapsley & Narvaez, 2006; Snarey, 1985). Some researchers propose that existing limitations in addressing differences in moral reasoning partly stem from the limited available information concerning the adequacy of measures to use across different cultural groups (Snarey, 1985). In fact, evaluations of character education programs using outcome measures that are not equivalent across social groups could yield misleading or inaccurate information about the effectiveness of interventions (Lapsley & Narvaez, 2006).

Purpose

The purpose of this study is to validate two existing scales that measure two types of character within students. First, using a modern measurement methodology called Rasch

measurement theory (Andrich, 1988; Bond & Fox, 2001). I examine how well items on the *Student Performance Character* (SPC) and *Student Moral Character* (SMC) scales measure the constructs student performance and moral character. Student performance character is defined as qualities needed to realize one's potential for excellence in academics, co-curricular activities, the workplace, or any other area of endeavor. Student moral character is defined as the qualities needed for successful interpersonal relationships and ethical behavior, such as integrity, caring, and respect. It highlights how we treat ourselves, as well as others (Lickona & Davidson, 2005).

Second, I examine how items on the SPC and SMC scales fit the Rasch measurement model framework. The Rasch model provides an additive, mathematical algorithm that expresses the probabilistic expectations of item and person performances. Utilizing key aspects of this model, such as reliability indices and fit statistics, I examine whether or not the SPC and SMC scales fit this measurement framework. Finally, I check whether items perform equivalently across the two gender groups.

Rationale

There are several reasons to validate character education outcome measures more judiciously: (a) new theories regarding character development can be more rigorously tested, (b) the use of the modern measurement framework can lead to better inferences, (c) the effectiveness of character education programs can be more properly assessed, (d) a more accurate assessment of the relationship between character variables and academic achievement can be made, and (e) true item impacts for different social groups (e.g., gender-related differences) can be detected.

First, when an instrument is validated, researchers can better investigate theories and potentially propel research fields into new directions. Of particular interest to this study is Berger's (2003) theory called *ethic of excellence*. This new theory posits that when an ethic of

excellence is at the center of school culture, students do well. It highlights the experiences of excellence that are a central part of human fulfillment. According to Berger (2003), educating for character ought to be about developing ethics *and* excellence. Although the underpinnings of this theory are pragmatic, there have been few empirical investigations that validate these theoretical claims. Moreover, instruments created to measure the performance character construct must be tested in order for theoretical claims to stand up:

> Modern psychological theories of moral behavior and moral development need new, theoretically valid methods of measurement. Unless we have them, we cannot make valid inferences from data on the empirical validity of those theories. The evolution of new and better theories depends on the construction of better research instruments and vice versa. (Lind, 2008, p. 186)

Furthermore, whereas the Student Performance Character and Student Moral Character scales have evidence of validity based upon a classical test theory approach (CTT), the use of a more modern approach, such as item response theory (IRT) may yield more useful inferences. CTT methods represent a traditional approach to data analysis in the social sciences that has been used widely for most of the 20th century (Iramaneerat, Smith Jr., & Smith, 2008). Its basic concept is that ability is expressed by the *true score*, or the expected value of observed performance on the test of interest (Hambleton, Swaminathan, & Rogers, 1991). First introduced by Charles Spearman in 1904, CTT is based on conceptual models where relations among constructs are theorized (Royal, 2008). Generally, CTT is used to examine a group of responses to a test. In other words, an examinee's ability is defined only in terms of a particular test.

The CTT framework has significant limitations. The biggest shortcoming of this approach is that examinee characteristics and test characteristics cannot be separated; both

person ability and test difficulty can only be interpreted in the context of the other (Hambleton, et al., 1991). In other words, measures of persons and items are test and sample dependent. For example, if a student encounters a difficult test, his or her ability measure will be lower, but if the same student encounters an easy test, his or her measure of ability will be higher (Iramaneerat, et al., 2008). Therefore, person ability and item difficulty cannot be generalized to other samples of persons and items. This study seeks to overcome this limitation by applying an item response theory (IRT) approach to establishing validity.

Another limitation of the CTT approach is that researchers can only make score comparisons within a sample if every examinee gives complete responses. Unless a missing data imputation/estimation method is employed, measures of person ability or item difficulty cannot be fairly compared (Iramaneerat, et al., 2008). This is especially problematic within educational settings where students may fail to complete every item. Within character education evaluations, the instruments used must be validated in order to make fair and accurate inferences regarding program success, despite missing data.

A third shortcoming of the CTT approach is that there are limited techniques for validating response patterns. Within educational settings, a student measuring high on a latent variable might answer an easy item wrong because of time constraints, while a student measuring low on that variable might answer a difficult item correctly because of guessing (Iramaneerat, et al., 2008). Kuhn (1961) states that the significant role of measurement in scientific discovery lies in its ability to display serious anomalies to direct scientists where to look for new qualitative phenomena. CTT, however, lacks the methods to detect abnormal response patterns, which can lead to distorted inferences and prohibit new scientific discoveries.

Character education evaluation instruments should provide validity evidence that uses techniques that validate response patterns.

The most significant benefit of an IRT approach is that the performance of an examinee on a test item can be predicted/explained by a latent variable (Hambleton, et al., 1991). As the measurement field advances, it is useful to apply different theoretical approaches to establishing validity. Therefore, an IRT approach, such as the Rasch model, can provide evidence regarding the latent variables student performance character and student moral character: Iramaneerat et al. states that Rasch measurement is a model-based approach in measurement that has become increasingly popular for scale construction in the social and other sciences (2008). As a result, this study applies an item response theoretical approach to establishing validity for the Student Performance Character and Student Moral Character scales:

> IRT is a relatively recent development in psychometric theory that overcomes deficiencies of the classical test theory with a family of models to assess model-data fit and evaluate educational and psychological tests. (Bond & Fox, 2001, p. 231)

In fact, validating an instrument/scale through a Rasch measurement framework allows researchers to revise their instruments, testing one latent variable per scale. It requires the ordering of all items by their difficulties to be the same for all ability levels. As a result, fair and accurate comparisons can be made, allowing for substantive interpretations (Iramaneerat, et al., 2008). For example, schools can benchmark their current state (regarding student effort, frequency of cheating, etc.) and then plan an intervention that addresses needs revealed by the data (Davidson, et al., 2010). Furthermore, the fit statistics allow researchers to examine item responses, as well as scale ordering. These elements make Rasch measurement theory an attractive option for scale development/analysis.

Validating scales, such as the *Student Performance Character* or *Student Moral Character* scales, via a Rasch measurement framework, can be beneficial in helping evaluators and program deliverers identify and interpret evidence on the internal structure of the constructs, as well as the items design. The visual map produced through Rasch analyses, called the Wright map (Bond & Fox, 2001), is a key feature of Rasch modeling and is useful for practitioners who are interested in evaluating a character program's effectiveness. This map is a visual depiction of how persons and items are ordered on the same linear scale. It is easily interpretable and makes the findings intelligible to teachers, administrators and parents (Mentkowski, 1980).

A third reason to validate character education outcome measures more judiciously is that researchers can better evaluate and assess character education programs for accountability and replication purposes. For example, Power (1980) discusses the challenges with evaluating Kohlberg's *just community approach* to moral education. This program was specifically designed to help schools build democratic communities. As a result of this new theoretically-based approach to teaching character, new methods for investigating the effectiveness of moral education programs were also needed:

> The transition from assessing moral competence to moral performance required the sacrifice of many of those "laboratory-like" conditions which made Kohlberg's original structural developmental analysis possible. The new methodology had to be flexible in utilizing all sources of information available about a particular program in order to interpret the meaning of particular events, statements, decisions, and actions. As evaluators we had to become not only pre- and post-test interviewers but participant observers of school functioning and historians charting the progress or decline of the community. (Power, 1980, p. 180)

In other words, as new approaches for moral/character education instruction emerge, it is even more vital that program implementers and researchers design and utilize valid instruments that capture program effectiveness/ineffectiveness.

Also, an empirical investigation of scales and instruments measuring character is necessary in order to determine whether program effects are accurately being captured. A 2008 U.S. Department of Education report analyzed information provided by pilot character education programs in various states over a six-year period. They recommended that "the relative impact on school climate, teacher efficacy, involvement of parents and community, and other components of character education programs should be measured to determine the level of success of each, using both process and outcome evaluation methods and valid, reliable survey instruments" (United States Department of Education Office of Safe and Drug-Free Schools, 2007, p. 10). In other words, the federal government recognized that states implementing school-based character education programs needed to use valid tools in order to accurately capture programmatic effects.

A fourth reason to better validate character education measures is to produce a more accurate assessment of the relationship between character variables and other, school-related variables. Overall, scholars argue that the variety of strategies and activities specific to a character education program, as well as the numerous contextual variables outside of the program, make it especially challenging to isolate the specific intervention that contributes to a student's change in character or a program's success (Berkowitz & Bier, 2005; Corrigan, Chapman, Grove, Walls, & Vincent, 2007; Thomas, 1991).

For example, there is a large body of literature that demonstrates the connection between students' psychological traits (i.e., character attributes) and academic achievement. According to

a study conducted by Duckworth and Seligman (2005) students' self-discipline was a predictive variable of final academic grades. This effect of self-discipline even held when controlling for prior grades, standardized test scores, and IQ. Additionally, Marzano (2001) synthesized research studies that investigated students' interest, thereby effort, on their academic achievement. There is a moderate to strong relationship for all six studies between interest and achievement; the average effect size is approximately .7, demonstrating that the more interest students have in a topic, the more energy and attention they will put into their academic work. In order to examine positive character attributes and its correlation with constructs such as self-discipline, effort, and other youth development outcomes, solid methodology must be used in the instrument development process (Bond & Fox, 2001).

Another research study revealed that students' background characteristics (such as gender, ethnicity, socioeconomic status, and academic achievement) prior to enrolling in high school are much less important in explaining failures than are their behaviors in high school (Allensworth & Easton, 2007). Students' background characteristics explain 7% of the differences in failure rates among students, and test scores explain an additional 5% (12% total), but absences and studying explain an additional 61% beyond test scores and demographic characteristics (73% total). In other words, students' experiences and behavior while in high school are more important for passing courses than their background characteristics. A large body of research has explored academically-related behaviors as an avenue for investigation in improving the nation's abysmal graduation rates (Alexander, Entwisle, & Kabbani, 2001; Allensworth, 2005; Lee & Burkham, 2003; Neild & Balfanz, 2006). This research agenda suggests that character traits matter in the improvement of high school graduation rates. If so, valid instruments must be used in the evaluations of character initiatives.

Identifying the specific factors that influence student character is not only a research design issue, but also a measurement issue as well. Corrigan, et al. (2007) states that hypothesized impacts or possible causal relationships between character education and academic achievement cannot be detected without the collection of numerous variables *and* valid scales that measure change within each variable. The study outlines the character education intervention as the experimental stimuli, student character as one of numerous independent variables, and several dependent variables including student achievement test scores, grade point average, and discipline reports. In their study, they outlined a dimensional model that included these varied components. Researchers carefully chose these variables based on the funder's interest in distal outcome measures. Unreliable measures, however, would only mask the relationships between variables, thereby discrediting inferences: "One must use reliable and (when possible) valid scales or measurements for each category so that quantitative evidence is observable in a meaningful fashion" (p. 109).

Finally, inquiries into gender-related differences in character development require instruments that are invariant across the two gender groups. There are divergent beliefs amongst character education scholars that moral orientations differ according to gender: "…The field of the psychology of moral development and moral education is generally bifurcated between Kohlberg-inspired advocates of an ethic of justice and Gilligan-inspired advocates of an ethic of care" (Siddle-Walker & Snarey, 2004, p. 4). Some scholars believe that males have character orientations that favor rights and responsibilities, while others believe females are oriented towards caring and responsibility (Gilligan, 1982). According to Carlo et al. (2008), gender differences in specific types of moral reasoning have been reported in empirical studies (Carlo, Eisenberg, & Knight, 1992; Carlo, Kohller, Eisenberg, Da Silva, & Frohlich, 1996; Carlo,

Roesch, & Kohller, 1999; Eisenberg, Carlo, Murphy, & Van Court, 1995; Eisenberg, Zhou, & Kohller, 2001; Wyatt & Carlo, 2002). For example, adolescent females more frequently report empathic and internalized forms of prosocial moral reasoning than adolescent boys. Meanwhile, adolescent boys report more approval-oriented prosocial moral reasoning than their female counterparts : "These findings are consistent with gender socialization theory (Gilligan, 1982) that suggests that socialization agents encourage girls to focus on issues of care and nurturance more than boys" (Carlo, et al., 2008, p. 488). According to Gilligan (1982), women process moral dilemmas as a problem of care and responsibility within relationships, a very different orientation than Kohlberg's ethic of justice (see Table 1 and 2). Walker (1991), however, found moral judgment orientations did not consistently differ according to gender.

As a result of the divergent views, and because there are limited studies regarding gender invariance in instruments measuring character (Carlo, et al., 2008), more empirical studies are needed to determine if scale/instrument[1] items function equivalently for males *and* females. Differential item functioning can provide a lens in order to investigate potential item bias. In the context of this study, an investigation to detect differential item functioning (DIF) within items is to see if male and female students have differing probabilities or likelihoods of endorsing an item after conditioning or matching on the extent of the latent trait performance or moral character (Clauser & Mazor, 1998).

In conclusion, the scales/instruments that measure character should be validated within a rigorous framework for numerous reasons: (a) new theories regarding character development can

[1] Throughout this study, the terms "scale" and "instrument" are used interchangeably.

be more rigorously tested , (b) the use of the modern measurement framework can lead to better inferences, (c) the effectiveness of character education programs can be properly assessed, (d) a more accurate assessment of the relationship between character variables and academic achievement can be made, and (e) true item impacts for different social groups (e.g., gender-related differences) can be detected. Therefore, this validation study is an important addition to the field of character education specifically, and education/psychology more broadly because it seeks to address the stricter standards posed by the federal government for scientific inquiry within education evaluations (United States Department of Education Office of Safe and Drug-Free Schools, 2007).

In response, this study addresses three different research questions: (a) how well do items on the Student Performance Character and Student Moral Character scales measure the two constructs; (b) do the items on the Student Performance Character and Student Moral Character scales fit the Rasch measurement model framework, and if so, how well is the fit; (c) do any of the items display gender differential item functioning (DIF) and, if so, what are their patterns, directions, and magnitudes?

CHAPTER II

Theoretical Framework

There are two psychometric theories that undergird this research study. First, a measurement theory called Rasch (Andrich, 1988; Bond & Fox, 2001) provides the foundation necessary to answer research questions one and two. Rasch measurement theory consists of several fundamental principles required for measurement; these principles are described in the following section. The second theoretical framework undergirding this empirical study is a unified model of validity (Messick, 1989). I will examine the historical context of several validity theories in order to understand what it means to "validate" an instrument and utilize a unified model of validity.

Rasch Measurement Theory

Rasch measurement theory provides the framework needed to empirically examine the validity of the Student Performance Character and Student Moral Character scales. It is an important tool in scientific research where phenomena are observed on an ordinal scale and parametric statistics are employed for analysis (Wright & Linacre, 1989). It is the only specifically objective, specific, item- and person-free measurement theory present. It is objective because, if model fit and sufficient targeting are present, it allows for comparisons between items without reference to the respondents and allows the researcher to compare respondents without interference from the items (Mead, 2008). This fundamental principle provides the basis for this research study.

Measurement theory refers to a body of principles, ideas, rules, and techniques for quantifying a variable (Mead, 2008). In order to make inferences from observations, several principles must be followed. According to Rasch measurement theory, the first primary

requirement for fundamental measurement is the reduction of experience to a one dimensional abstraction (Andrich, 1988; Bond & Fox, 2001; Wright & Masters, 1982; Wright & Mok, 2004). In other words, in order to make meaningful estimations of an object, researchers must focus on measuring only one attribute at a time. This is referred to as unidimensionality.

The second requirement for measurement according to the Rasch model is that comparisons can be made along the unidimensional continuum. Using a mathematical algorithm that expresses the probabilistic expectations of item and person performances, the Rasch model turns ordinal data into interval transformations (Andrich, 1988; Bond & Fox, 2001; Wright & Masters, 1982). As a result of the Rasch model, ordering observations begins as a qualitative task, but transforms into a quantitative one in order to represent the estimation of person ability and item difficulty.

The final principle of the Rasch model is the production of fit statistics. This framework provides indicators of how well each item fits with the underlying construct. These fit indices help researchers determine whether the assumption of unidimensionality holds up empirically (Bond & Fox, 2001). These three foundational tenets provide a conceptual framework for the development and analysis of any scale within social science fields. This study will conduct two types of Rasch measurement analyses: Rating Scale and Partial Credit (Wright & Masters, 1982). These models will be compared to determine which one best estimates parameter for the student performance latent variable and the student moral character variable.

Validity Theory

Within the educational measurement field, validity has always been regarded as the most fundamental and important concept (Angoff, 1988). As a result, it has significantly evolved throughout the 20th century. In fact, the theory of validity has changed more than the practice of

validation (Kane, 2011). Much of the early discussions around validity defined it as the evaluation of how well test scores predicted the criterion scores (Bingham, 1937; Cureton, 1951; Guilford, 1946; Kane, 2001; Lindquist, 1942). According to psychometric scholars, validity was determining how well a test did the job it was designed to do (Cureton, 1951). Eventually, however, it was discovered that many criteria used in formal programs were severely inadequate (Angoff, 1988). Therefore, psychometricians began searching for techniques to improve the development of criteria. Cronbach and Meehl (1955) proposed that test developers examine the psychological trait (i.e. construct) the test presumed to measure. The resulting data are then used to validate, reject, or revise the theory underlying the construct. In sum, construct validity became the whole of validity from a scientific point of view (Loevinger, 1957).

With the development of varying types of validity evidence (criterion, content, and construct), validity theorists expressed concern about researchers treating validation methods as a toolkit; different models were used for different assessments. For example, “the criterion model would be used to validate selection and placement decisions, the content model would be used to validate achievement tests, and the construct model would be used for theory-based explanations” (Kane, 2011, slide 20). Furthermore, researchers were sometimes choosing methods for their convenience, rather than for their appropriateness. With the birth of standardized testing, for example, validity was assessed using many different procedures and was called by a variety of names: “The type of evidence adduced to demonstrate test validity varied with the purpose of the test, the theoretical orientation of the test author, and—all too often—with the ready availability of the data” (Anastasi, 1986). Eventually, to combat this phenomenon, a unified model of validity was developed, providing the foundation for this validity study: “Validity is an integrated 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" (Messick, 1989). This definition indicates a major shift from the validation of an instrument to the validation of a proposed interpretation or use of assessment scores. Thus, establishing validity of interpretations and uses for a particular test came to be viewed as a concept that required multiple research studies rather than a single empirical study (Kane, 2011). As a result, researchers could no longer just pick and choose which was most convenient; a variety of validity evidence needs to be presented.

CHAPTER III

Review of Literature

The purpose of this validation study is to examine two scales that measure key qualities and traits of moral and performance character within a school context. This literature review will provide a context for this study in several ways. First, I will define character and character education. Second, I will discuss the major theoretical, pedagogical, and philosophical orientations used in character education programs. Next, I will discuss three general categories of instruments that measure certain traits of moral and/or performance character in students: (a) instruments based primarily on moral development theory, (b) instruments based primarily on psychological theory, and (c) instruments based on both moral and psychological theories. Validation processes of various instruments that are based on moral development and psychological theories will also be discussed. Finally, the Student Performance Character and Student Moral Character scales will be compared to the scales discussed.

Defining Character

Generally, operationalizing the construct of character is challenging because the concept is more ethically reflected upon than empirically studied (Carr, 2008a). Scholars are likely to agree on the general definition of good character, but operationalizing it for empirical study may spark debate. Identifying the inclusionary and exclusionary rules is an important step in defining "character" for empirical study, especially when empirical questions relate to the effectiveness of a character education program. Consequently, to effectively measure the efficacy of character education programs a common understanding of the dimensions being measured is required (Eastman, 2008). Hence, this section will discuss conceptions of character from a philosophical, educational, and moral perspective.

Dating back to Ancient Greek times, essential virtues included wisdom, justice, fortitude, self-control, love, a positive attitude, hard work, integrity, gratitude, and humility (Lickona, 2004). These ten virtues encompass Aristotle's *life of right conduct,* including right conduct in relation to other persons and right conduct in relation to oneself: "In the classical Aristotelian view, a virtue is a commendable state or trait of character apt for a certain mode of conduct… a kind of disposition" (Carr, 2008b, p. 43).

Currently, in the 21st century, those same virtues remain a priority. According to the No Child Left Behind Act of 2001 (NCLB), the Secretary of Education is authorized to award grant money to state and local educational agencies for the design and implementation of character education programs ("No Child Left Behind Act of 2001,"). Since 2004, the Office of Safe and Drug-Free Schools has awarded over 22 million dollars for this purpose (http://www2.ed.gov/programs/charactered/awards.html). As awardees are expected to develop measures of effectiveness in order to evaluate and measure the success of their character education programs, the United States Department of Education has listed topics on or "elements of character" that they are permitted to teach. These include caring, civic virtue and citizenship, justice and fairness, respect, responsibility, trustworthiness, and giving. But even the federal government recognizes that this list is not exhaustive; they permit "any other elements deemed appropriate by the eligible entity" to be covered within the character curriculum (http://www2.ed.gov/policy/elsec/leg/esea02/pg69.html). Therefore, to provide context for this study, key terms and concepts will be defined.

Defining the abstract concept of character can prove to be quite challenging given that researchers and scholars oftentimes use the term interchangeably with other, related terms (e.g., ethics, morals, and values (Cole, 2004; Eastman, 2008). Although, for example, the distinction

between character and morality is not transparent, Kohlberg (1964) and Lind (2008) introduce morality as a predecessor of good character; in other words, they believe that morality is necessary for the existence of good character.

Generally, Kohlberg (1964) defines morality as a set of cultural rules of social action which have been internalized by the individual; thereby moral development is the increase in such internalization of basic cultural rules. More specifically, according to Lind (2008), there are three definitions of morality: conformity, intentions, and competence. Morality as rule-conformity consists of "a list of things that should be done and better be avoided: Don't steal, don't murder, [etc.]" (Lind, 2008, p. 187). Therefore, according to the *conformity* definition, morality is measured by the number of instances in which a person exhibits morally correct behavior and avoids morally wrong behavior. Secondly, morality as *intentions* defines a behavior as morally good if it is based on good intentions (or moral values, motives, attitudes, or principles). Last, morality as *competence* conceptualizes the relationship between moral ideals and moral behavior. Competence (also called ability or capacity) is defined as the capacity to make moral decisions and judgments and to act in accordance with those judgments (Lind, 2008). Morality defined as competence marries the cognitive and affective domains of behavior. Lind (2008) defines this relationship as the dual aspect model, where moral ideals lead to moral competencies, which then lead to moral action. In conclusion, whether it is coined character or morality, the literature suggests great overlap between these terms. This is especially challenging when researchers attempt to measure good character.

Berger (2003) expands Lind's definition of character by developing a theory that birthed a new type of character: performance character. Performance character is defined as qualities needed to realize one's potential for excellence in academics, co-curricular activities, the

workplace, or any other area of endeavor (Lickona & Davidson, 2005). Uniquely distinct from moral character definitions, performance character adds a new facet to character literature.

Defining Character Education

According to Willis (2008), there are four schools of thought that shape the moral education field: character education, moral reasoning, critical literacy, and community of compassion. *Character education* refers to a direct approach rooted in traditional American education. It aims to transmit a specific set of predetermined values to children (Solomon & Watson, 2008). *Moral reasoning* counters this authoritarian school of thought. Rooted in the works of Kant, Dewey, Piaget, and Kohlberg, this democratic, indirect approach focuses on the settings in which interpersonal relationships are formed, thereby helping children to reason about, understand, care about, and act in accordance with moral principles (Solomon & Watson, 2008; Willis, 2008). Kohlberg's theory of moral development (Puka, 1994), for example, often referred to as an *ethic of justice*, offers a hierarchical structure to how a person's moral character should develop. Each stage outlines three levels of moral development: preconventional, conventional, and postconventional. Each level is subdivided into two stages, yielding 6 total stages (see Table 2). Each stage is viewed as cognitively higher than the previous stages. Preconventional moral judgment, for example, is considered egocentric, where people are only concerned with individual needs, while postconventional judgment adopts a reflective perspective on societal values (Gilligan, 1982). According to Kohlberg, when an individual's moral understanding expands from self to a societal point of view, he has achieved a higher stage of moral development.

The third school of moral thought, *critical literacy*, is based on the theories of Freire (1998) and Giroux (1992) and aim to increase the moral treatment of one another. Finally, some

theorists, such as Nel Noddings (2008), seek to increase the compassion within school communities. This approach, called *community of compassion,* applies feminist concepts of care and nurture to the field of moral education.

Although these four traditions are not the only schools of thought regarding moral education (e.g., the nearly obsolete ideology of values clarification; see (Willis, 2008), they are the most prominent: "These four landmarks…are certainly not enough to cover every major approach out there… Still, four labels may be more accurate than the typical two [moral development and character education] and more manageable than twenty or thirty" (Willis, 2008, p. 6).

In summary, character education is simply one approach to moral education. There are several historical perspectives that contextualize character and moral education. For the purposes of this study, however, I will describe three definitions of character education.

According to the now-defunct National Commission on Character Education, character education is "any deliberate approach by which school personnel, often in conjunction with parents and community members, help children and youth become caring, principled and responsible" (Williams & Schnaps, 1999, p. 1). Similarly, Lockwood (1997) believes character education is defined as any school-instituted program, designed in cooperation with other community institutions, to directly and systematically shape the behavior of young people. Finally, Ryan and Bohlin (1999) describe character education as the development of virtues, good habits, and dispositions which lead students to responsible and mature adulthood. Many of the current definitions of character education draw from the direct approach.

In summary, this literature review defines character and character education from a historical, philosophical, theoretical, and governmental lens. In light of these various definitions,

this study will specifically study the character constructs according to Lind's (2008) definition of moral character and Lickona and Davidson's (2005) operational definitions of performance character because the scales under investigation in this study (the Student Performance Character and Student Moral Character scales) follow in these traditions.

Furthermore, while reviewing the literature on scales and instruments, some reoccurring ideologies consistently emerged. First, many scales/instruments were birthed out of the tradition of Kohlberg and moral development theory. Secondly, there were scales/instruments measuring aspects of character that were rooted in psychological theories. Finally, there were hybrid measures that combined moral development and psychological theories. Overall, the historical tradition of character scales and instruments is vast. Therefore, it may be difficult to define types of character and measure them appropriately without careful consideration of the historical traditions that emerge or converge. For the purposes of this study, performance character and moral character are considered two essential aspects of character, with origins from both moral and psychological traditions.

Using this construct definition, this literature review categorizes instruments measuring various types of character into three areas. A description of each scale/instrument is outlined, as well as the validation approaches used within each study. Specifically, the following section will outline five major scales/instruments used to measure character based on a moral development theory *and* report validation studies performed on each**.**

Scales/Instruments Based Primarily on Moral Development Theory (Moral Character-Related)

Moral Judgment Scale. One of the first applications of Kohlberg's theory of moral judgment was the development of the first moral judgment test (Kohlberg, 1984). It consisted of

two forms—A and B—where each form included three hypothetical moral dilemmas. The first dilemma demonstrated a conflict between helping someone improve their quality of life, although it was in violation of the law. The second dilemma showed a conflict between the conscience and retributive justice. Finally, the third dilemma established a conflict between the maintenance of a contract as opposed to the upholding of legitimate authority (Kohlberg, 1984). See Table 3 for a more specific example of a moral judgment situation item. The item describes the criterion being judged (a woman in pain asks her doctor to give her a drug that will kill her) and the critical indicators that scorers should use in evaluating responses (the doctor should not risk his job and breaking the law for a mercy-killing). Also, the scorer has examples of "match" and "guess" examples as possible responses. This example is also matched to the Kohlberg moral stage of development (2A).

Kohlberg (1984) states that his validation strategy included demonstrating perfect invariant sequence. He proposed collecting longitudinal data from subjects tested at three-year intervals over 20 years. Rest (1979), however, disagrees that validation can only occur 20 years after the instrument's development. Kohlberg also reported several statistics that supported high reliability for both forms of the moral judgment test. There was high correlation between two time points for forms A and B (over 90%) and the percent agreement between scores for both time points were within 1/3 stage of each other. Overall there was between 70 and 80% agreement between forms and time periods (Kohlberg, 1984). Test and retest reliabilities were also reported to assess interrater agreement; the correlation between raters one and two was .98.

Construct validity is addressed but not thoroughly examined:

Our discussion of the construct validity of the test has been restricted to its internal characteristics, structural wholeness and invariant sequence. There is a recurring demand to

know about the validity of the test as a predictor to external criteria, particularly moral action or behavior… Our test was a measure of judgmental competence necessary but not sufficient for real-life moral decision making and judgment. (Kohlberg, 1984, p. 34)

Therefore, Kohlberg's interview measure could use some further investigation of its relationship to other psychological constructs. Also, Simpson (1974) states that the Moral Judgment Scale fails to identify principled moral reasoning in Non-Western populations, and thus a major portion of his theory remains invalidated. Furthermore, there were a total of nine dilemmas for test administrators to choose from. According to eight published studies, researchers rarely reported which of the nine dilemmas they used, challenging the assumption of content validity: "[The] lack of a standardized scale content contributes further to the difficulty of evaluating research results because not all of the situations are equally effective for assessing moral reasoning" (Kurtines & Blank-Greif, 1974).

Defining Issues Test. Derived from Kohlberg's work on the Moral Judgment scale, the Defining Issues Test (DIT) was the next step in the evolution of moral judgment testing. Unlike Kohlberg's measure, this instrument is a self-administered, multiple choice questionnaire that uses 6 dilemmas or scenarios. Each scenario has 12 questions that are designed to represent different schemes of fairness (Rest, Thoma, Moon, & Getz, 1986). The candidate determines the importance of each dilemma by using a 4-level scale. As a result of this instrument, the "P-score" (*P* standing for principled moral thinking) helps determine the relative importance students give to items representing Kohlberg's Stages 5 and 6 (see Table 2).

The Defining Issues Test is an example of one of the most commonly used measures for testing ethical reasoning in high-stakes examinations. Numerous scales measuring moral

character have been birthed from this seminal work. There are also countless studies that examine various aspects of the instrument's validity.

First, Rest (1979) extensively outlines how the DIT was constructed. The relationship between the instrument's content and the construct it is intended to measure is clear through the definition of development features and stages of moral judgment. Also, data collection procedures provide a strategy for coding the responses into an outcome space and scoring them. Rest (1979) states:

> If moral judgment is an important construct and refers to something pervasive and influential in human functioning, it should be manifested in many ways. If it could be assessed only by interviewing subjects about particular hypothetical dilemmas [like Kohlberg's measure], then we should question whether the whole field studies a circumscribed, relatively question whether the whole field studies a circumscribed, relatively trivial phenomenon of interviewing behavior. On the other hand, if moral judgment is really "the fundamental structure by which people perceive and make decisions about their rights and responsibilities," it should be manifested in many kinds of responses. A great variety of tasks could be used to generate moral judgment material. (p. 76)

Several options of collecting data are listed such as abstract direct questioning, justifying solutions to moral dilemmas, and comparing acts/actors in stories. Rest (1979) is clear on the varied ways for content to be measured and provides numerous supports that the DIT measures what it intends to measure.

Second, the DIT provides evidence that supports the internal structure. Davison, Robbins, and Swanson (1978) apply scaling and factor analytic techniques to test the stage structure

hypothesis (where moral judgment stages comprise a definite order). Results showed support for the ordering of Stages 2, 3, 4, 5A, and 6 (Davison, 1979). Also, evidence did not support a distinct ordering between 5A and 5B; in other words neither stage represented a higher level of development than the other. Overall, however, good internal structure exists with the DIT.

Third, DIT studies have also yielded evidence that support the construct map. Davison, Robbins, and Swanson (1978) examined the P score index of overall moral development. Alternatives to this index were explored (specifically, a sum of responses to all items and a weighted sum of item responses). Several studies were conducted to compare these different models for deriving a final score. Results showed that the weighted sum provided the most desirable measure of moral development. The mean scale values provide clear support for the hierarchical ordering of Kohlberg's Stages 2, 3, 4, and more ambiguous support for Stages 5 and 6. Therefore, there is ample evidence, using a statistical/measurement model, which supports the construct map.

Fourth, there are numerous studies that provide evidence supporting the item design. Rest (1976) reviewed 22 studies examining how items performed differently across gender; only two studies had a significant difference in scores between males and females. Furthermore, within those two studies, only about 6% of the variance was accounted for by the gender variable and both studies showed females scoring higher (Rest, 1976).

Rest et al. (1986) conducted meta-analyses of 56 studies on sex differences and examined findings of 20 cross-cultural studies. They found that test-retest coefficients on the original DIT tended to be lower in non-Western cultures, but generally differences could be attributed to differences in translated versions of the DIT. Also, Thoma (1984) conducted a meta-analytic procedure that provided precise information on the size and significance of the gender effect (an

improvement from previous empirical studies; (Rest, et al., 1986). Using two measures, Cohen's *d* and *W*, Thoma (1984) found that females actually scored higher on the DIT than males. Furthermore, the *W* values indicated that gender accounted for no more than .9% of the variance in DIT scores.

Other studies, such as Haan, Brewster-Smith, and Block (1968), show females disproportionately scoring within Stage 3 of Kohlberg's moral development pyramid, while males score within Stage 4 (see Table 2 for a full description of each stage). They found that the majority of females were in Stage 3, with the next largest concentration within Stage 4, compared to males. Critics (Holstein, 1976), however, claim that 41% of females in Stage 3 compared to 22% of males, and 39% of females in Stage 4, in contrast to 22% males is not enough evidence to deem the DIT bias against females (Holstein, 1976; Rest, 1979). This dissension creates concern that items are not invariant across various social groups; DIF, therefore, can be a useful analysis to validate a scale or instrument across varying groups.

Finally, there is evidence that supports the relation of ethical reasoning (measured by the DIT) to other variables. According to Rest (1979), there are countless correlation studies that show the DIT's relationship with other psychological constructs, cognitive traits, and other moral judgment measures. Convergent and divergent validity has been demonstrated with evidence from numerous studies. Some glaring results include: (a) measures of liberal-conservative attitudes do not uniformly correlate highly with the DIT; (b) the DIT correlates with measures of general intelligence fairly consistently; (c) the DIT distinctly measures moral judgment, and not simply general logical content; and (d) the DIT correlates modestly with two other measures of moral judgment (Kohlberg's measure and the Comprehension of Moral Issues, Law and Order Orientation, and Political Tolerance).

Moral Judgment Test (MJT). Patterned after Rest's Defining Issues Test, the MJT measures the degree to which respondent's judgments about a moral dilemma are morally consistent (Lind, 1978). The original version was developed by Georg Lind in 1977 and consisted of two dilemmas: a doctor's dilemma, similar to Kohlberg's item (see Table 3) and a worker's dilemma (involves breaking into a firm). It uses the "C-index" to determine the degree to which the respondent let his/her judgment be determined by moral principles rather than by other psychological forces (Lind, 1978).

Lind (1999) states that the MJT is theoretically valid because it was developed based on a solid theory. Also, half a dozen experts of Kohlberg's stages of moral development provided feedback on the adequacy of each moral situation included. Therefore, the MJT provides evidence based on instrument content.

The MJT also provides evidence based on internal structure. According to Lind (1999), the instrument has an intentional structure that follows a specific order: "To my knowledge, all MJT-studies have found… a preference order [and] have a quasi-simplex structure" (p. 12).

Furthermore, the MJT demonstrates evidence based on its relation to other variables. Lind (1999) found a correlation between the instrument and two other aspects of moral judgment behavior. Also, two experiments showed that respondents were able to simulate a score higher than their own on other tests of moral development, but *not* the MJT. This posits that the other instruments measured moral attitudes rather than moral competence, thereby providing convergent evidence of construct validity.

Prosocial Moral Reasoning Scale (PROM). Modeled after Rest's DIT, this scale was developed to measure adolescents' level of prosocial moral reasoning. Several stories are included within the instrument and students must respond to each scenario on a 7-point Likert

scale. Scored responses indicate one of the following developmental levels: hedonistic, needs-oriented, approval-oriented, stereotyped, and internalized. Validity evidence that supports the instrument's relationship to other constructs is provided; higher level, internalized reasoning is significantly, positively related to cognitive variables (e.g., perspective taking and school aptitude) and affective variables (e.g., sympathy; Carlo, 1992).

Ethics Position Questionnaire. Developed by Forsyth (1980), the Ethics Position Questionnaire (EPQ) consists of 20 Likert-scale items, with a 9-point scale. The purpose of this instrument is to measure attitudes about idealism (10 items) and relativism (10 items), and then classify individuals into one of four ethical ideologies: situationism, absolutism, subjectivism, and exceptionism. Evidence supporting the relationship of idealism and relativism as constructs to other constructs is presented through correlations. Idealism on the EPQ is strongly correlated to another measure of idealism, while relativism is moderately correlated to another measure of relativism. Furthermore, results show the EPQ is not correlated with Kohlberg's stages of morality, but is significantly correlated with another measure of ethical attitudes (Hogan's Survey of Ethical Attitudes). Finally, factor analyses are provided and insure that the scales did not have a limited range of application; items were sampled from as many of the factors as possible to insure heterogeneity in content (Forsyth, 1980, p. 178).

Scales/Instruments Based Primarily on Psychological Theories (Performance Character-Related)

In addition to measures based primarily on moral development theory, there are also scales/instruments that measure aspects of character that draw primarily from psychological theories. Therefore, this section will outline measures and their validation approaches from a psychological theoretical lens.

Responsibility: Personal Responsibility Measure (PRM). This measure was specifically designed to measure 4 components of personal responsibility for adolescents: (a) an awareness of, and control over, one's own thoughts and feelings; (b) an awareness of, and control over, choices made regarding behavior, (c) a willingness to hold oneself accountable for the behavior enacted and the resulting outcome, and (d) an awareness of, and concern for, the impact of one's behavior upon others (Mergler, Spencer, & Patton, 2007).

Mergler et al. (2007) explicitly present multiple levels of validation evidence. First, evidence supporting the instrument's content is provided. They conducted a thorough literature review of the personal responsibility construct and other related constructs. Also, they conducted multiple focus groups with teachers (two) and students (four) in order to explore and develop the construct. Sample focus group questions include, "If you broke personal responsibility down into its key parts, what would they be?" and "What in your life do you take personal responsibility for?"

Also, the instrument has adequate content validity because developers chose items from closely related constructs such as the Vocational Locus of Control Scale: Questionnaire (Fournier & Jeanrie, 2003), The Generalized Self-Efficacy Scale (Schwarzer & Jerusalem, 1995), and the Social Responsibility Scale (Flewelling, Paschall, & Ringwalt, 1993; Nedwek, 1987). No evidence, however, was presented showing the relationship between the instrument and these other related constructs. Finally, exploratory and confirmatory factor analyses were conducted to provide evidence for the item structure.

Responsibility: Perceived Responsibility for Learning Scale. This 18-item, 7-point scale was constructed to measure whether respondents perceived the student or the teacher as more responsible for learning tasks or outcomes (Zimmerman & Kitsantas, 2005). An example

item is: “Who is more responsible for a student being interested in school?” Exploratory principal component factor analysis was conducted to provide evidence of item design, but no other validity evidence is provided.

Motivation: Student Motivation Scale (SMS). This scale was developed to measure the state of a student's intrinsic motivation to learn (Richmond, 1990). It consists of 5 Likert scale bi-polar adjectives (e.g., motivated and unmotivated). Rubin, Palmgreen, and Sypher (1994) report coefficients of .94 and considerable evidence of construct validity.

Motivation: Motivated Strategies for Learning Questionnaire (MSLQ). This self-report instrument was created to measure student motivation, cognitive strategy use, metacognitive use, and management of effort (Pintrich & DeGroot, 1990). With a total of 56 items rated on a 7-point Likert scale (l = not at all true of me, 7 =very true of me), this measure highly correlates with the use of cognitive strategies and academic achievement. Five scales along two dimensions emerged from a factor analysis. Related to motivational beliefs, the scales are: (a) Self-efficacy ("I'm certain I can understand the ideas taught in this course"), (b) Intrinsic Value ("I prefer class work that is challenging so I can learn new things"), and (c) Test Anxiety ("I have an uneasy, upset feeling when I take a test"). Related to self-regulated learning strategies, the scales are cognitive use strategies ("When I study for a test, I try to put together the information from class and from the book"), and self-regulation ("I ask myself questions to make sure I know the material I have been studying").

Motivation: Academic Motivation Scale. This scale is composed of seven subscales which assess three types of intrinsic motivation: (a) Intrinsic Motivation to Know, (b) Intrinsic Motivation to Accomplish Things, and (c) Intrinsic Motivation to Experience Stimulation. It also measures three types of extrinsic motivation (external, introjected, and identified regulation).

Vallerand et al. (1993) translated this scale from its original French version into English. Validity evidence provided show identical findings to the French form in terms of internal consistency, temporal stability, and factorial structure. Concurrent and construct validity were adequately evidenced in several ways. Primarily, correlations between the academic motivation subscales and other motivation scales support the theories used to develop the instrument with high positive correlations in predicted ways.

Motivation: School Achievement Motivation Rating Scale. Developed by Chiu (1997), this 15-item, 5-point scale measures the construct achievement motivation. Teachers rate students on 15 behavioral descriptors (e.g., *Chooses to work above and beyond what is expected* and *Does something over again just to get it right*). Validity evidence provided shows items are highly correlated with grade point averages and moderately correlated with standardized tests. Concurrent and construct validity were evidenced with correlations between the scale and other scales that measured need for achievement, self-esteem, among other constructs.

Persistence: Persistence Scale for Children. This instrument specifically measures persistence in children. It includes 40 true/false items and was sampled with Israeli children. Sample items include, “When I take part in an argument, I do not stop until everything is clear” and “I usually give up easily when I do not succeed” (Lufi & Cohen, 1987). Reliability coefficients for the sample were initially .66, but after conducting the test-retest method it increased to .77.

Evidence that supports the item design is presented. First, results are compared across age and gender (means and standard deviations are very similar for boys and girls, across ages). Second, the original sample of respondents (M = 22.71, SD = 4.61) was compared to a group of young male gymnasts who were persistently active in their sport for at least one year (M = 25.06,

SD = 4.43; stat. sig at $p < .01$). Another comparison between the group of young male gymnasts (who were active for two additional years; *M* = 27.10, *SD* = 3.78) was compared to a third group of male gymnasts who eventually dropped out (*M* = 23.80, *SD* = 4.38; stat. sig at $p < .01$).

Also, evidence that supports the construct's relation to other variables is provided. Correlations between Persistence and Locus of Control ($r = -.42$, $p < .001$), Anxiety ($r = -.28$, $p < .01$, and various subscales of Frustration ($r = .58$, $p < .01$, for the need-persistence subscale and $r = -.55$, $p < .01$, for both the extrapunitive and impunitive subscales.

Evidence supporting the item design was mentioned, but not reported due to lack of relevance: "Factor analyses of the persistence scale with 224 children have not provided any meaningful information" (Lufi & Cohen, 1987, p. 182).

Persistence: The Urgency, lack of Premeditation, lack of Perseverance, Sensation seeking impulsive behaviour scale (UPPS). Developed in 2001, this 45-item questionnaire measures respondents on four different psychological processes that lead to impulsive behavior (Whiteside & Lynam, 2001). The fourth factor/subscale of UPPS specifically measures perseverance, or the tendency to stop completing a task due to easy boredom. Sample items include, "I tend to give up easily," and "Once I start a project, I almost always finish it." This scale is comprised of subscales including self-discipline, a subscale measuring persistence of the EASI-III Impulsivity Scales, a subscale of the NEO Personality Inventory, and two subscales of the Sensation Seeking Scale (disinhibition and boredom susceptibility; (Whiteside & Lynam, 2001).

Content validity is established because all of the items came from previous instruments. Also, evidence is presented that supports the relationship of the construct perseverance to other variables. Finally factor analyses were conducted, providing evidence of item selection.

Persistence: The Eysenck I.6 Junior Impulsiveness Subscale. This Impulsivity subscale (i.e., doing and saying things without thinking) was developed specifically children. The 23 yes/no items were divided into three subfactors: impulsiveness, venturesomeness (i.e., being aware of the risks involved, but still willing to chance it), and empathy (Duckworth & Seligman, 2005; Eysenck, Easting, & Pearson, 1984). Factor analyses were conducted, presenting evidence supporting the item design. Also, evidence supporting the relation to other constructs is discussed (correlations): impulsiveness correlated mainly with psychoticism and neuroticism (and somewhat with extraversion), while venturesomeness correlated mainly with extraversion (and somewhat with psychoticism). Finally, evidence supporting item analyses were presented; results were examined holistically, as well as by gender. Eysenck and Zuckerman (1978) found a moderate correlation between the construct *sensation-seeking* (SSS;) and two dimensions of Eysenck's construct impulsivity; Furthermore, a study of American males and females found that the correlation between those two constructs was .52 and .59, respectively, while English males and females showed a correlation of .41 and .43, respectively. Divergent validity evidence is also reported: The lack of relationship between the s*ensation-seeking* and the *N* dimension of the i*mpulsivity* construct is consistent with data using other general anxiety and neuroticism scales (Eysenck, et al., 1984).

Self-control: The Brief Self-Control Scale (BSCS). This scale consists of 13 items designed to measure self-regulatory behaviors amongst five domains: (a) achievement and task performance, (b) impulse control, (c) adjustment, (d) interpersonal relationships, and (e) moral emotions (Tangney, Baumeister, & Boone, 2004). Evidence supporting instrument's content is presented, after an extensive review of empirical studies on aspects of self-control. Item revisions brought the final total from 93 to 36 items, with a 5-point rating scale. Also, evidence

supports the construct's relation to other constructs. SCS scores were substantially negatively related with the dimensions of psychological adjustment from the Million Clinical Multiaxial Inventory - III(MCMI-III), a more detailed measure of psychopathology (Tangney, et al., 2004). This instrument is void of other types of validity evidence. Developers state that items were reduced using both rational and empirical methods, but no measurement model is explicitly discussed.

Self-control: The Self-Control Schedule (SCS). Developed by Rosenbaum (1980), this instrument was specifically developed to measure self-controlling responses cued by an internal response such as anxiety or pain in a clinical population (Peterson & Seligman, 2004). Therefore, self-controlling responses are intended to reduce the interference cause by that internal response. Content validity is established through a literature review of stress-coping methods. Also, the list of 60 potential items was examined by two experts in behavioral, clinical psychology. They were required to evaluate each item against the following questions: (a) is the item comprehensible, (b) does the item describe a situation that could be experienced by a wide range of people, and (c) does the item reflect an effective use of a self controlling response (Rosenbaum, 1980)? This review reduced the items to a total of 44. A sample item follows: "When I find that I have difficulties in concentrating on my reading, I look for ways to increase my concentration."

Evidence is provided supporting this instrument's relation to other constructs. For example, scores on the SCS were negatively and moderately related to dimensions of the Irrational Beliefs Test (Jones, 1968) score. In other words, respondents who reported greater application of self-control methods were generally less likely to have irrational beliefs (Rosenbaum, 1980).

Self-efficacy: The Self-Efficacy Scale. Developed from Bandura's theory of behavioral change (1977), the Self-Efficacy Scale measures general self-efficacy expectancies in a variety of areas, including social skills or vocational competence. The 23-item instrument, with a 5-point Likert rating scale, focus on three central areas: (a) willingness to initiate behavior, (b) willingness to expend effort in completing the behavior, and (c) persistence in the face of adversity (Sherer et al., 1982). This is not the original version (version 2), but it is the first self-efficacy instrument developed and directly supports Bandura's (1977) theory. Sample items include, "I avoid facing difficulties," and "I do not handle myself well in social gatherings" (Sherer et al., 1982). This instrument demonstrates moderate to high reliability with Cronbach alpha coefficients of .86 and .71 for the General Self-efficacy subscale and the Social Self-efficacy, respectively.

Also, evidence supporting the construct map is provided. Factor analysis was used to determine relevant and superfluous items. For example, items were required to load at the .40 level in order to be retained within the instrument; 13 items did not meet this criterion and were, therefore, discarded. Furthermore, evidence supporting the construct's relation to other variables is presented. Scores on this instrument were correlated with measures of several other personality traits, such as personal control (general = -.355 at $p < .0001$ and social = -.132 at $p < .0001$), social desirability (general = .431 at $p < .0001$ and social = .278 at $p < .0001$), interpersonal competency (general = .451 at $p < .0001$ and social = .432 at $p < .0001$), and self-esteem (general = -.510 at $p < .0001$ and social = -.279 at $p < .0001$). The developers of the scales used in this study, Student Performance Character and Student Moral Character, use self-efficacy as a rationale for their domains: "Our rationale for using self-efficacy to assess [our] competencies is

that changes in self-efficacy ('I am able to do this') can be expected to precede and predict changes in motivation and behavior ('I do this')" (Davidson, et al., 2010).

Diligence: The Diligence Inventory. Diligence is defined as an expression or reflection of effort expended toward a balanced or holistic development by the students in their mental, physical, social, and spiritual dimensions of life (Bernard, 1991). The Diligence Inventory includes 55-items that measure five dimensions of diligence: (a) motivation, (b) concentration and Assimilation, (c) discipline, (d) conformity and responsibility, and (e) devotedness and spirituality. Evidence supporting the item design is provided through factorial analyses and a procedure called Known-group Difference. Also, evidence based on the instrument's content is supported by a thorough literature review of several fields, including high school dropout, school reform, and attribution theory. Content validity was also established by the use of expert judges to evaluate the items. Finally, evidence does support the construct's relation to another variable. A correlation coefficient of .32 ($p < .001$, stat. sig) shows a low, yet positive, relationship between diligence and a measure of competence (grade point average; Bernard, 1991).

Scales/Instruments Based on Moral and Performance Character Theory

This final classification of scales/instruments and their validation approaches combine items based on moral development and psychological theories. This new theory marries the concepts of ethics *and* excellence. There are two examples of scales/instruments that demonstrate this hybrid, theoretical approach: Individual and Team Character in Sport Questionnaire (ITSQ; Davidson, 2006) and the Student Performance Character and Student Moral Character Scale (Cornerstone Consulting & Evaluation, 2009).

Individual and Team Character in Sport Questionnaire (ITSQ). This survey is designed to measure character-related outcomes within the context of sports (Davidson,

Khmelkov, & Moran-Miller, 2006). There are 3 scales with a total of 48 items. One of the scales, the Values Rating Scale, directly relates to character development within students. They rate themselves on key sports values, such as sportsmanship (8 items, alpha = .79), personal responsibility (5 items, alpha = .76), and perseverance (3 items, alpha = .82). Confirmatory factor analysis, often used to measure the construct validity of a scale, was the only analysis conducted. Results supported the hypothesized structure.

Student Performance Character and Student Moral Character Scales. According to Lickona and Davidson (2005), "to unlock the power of character is to define it to include the quest for excellence as well as the quest for ethics" (p. 18). Drawing heavily upon Berger's (2003) theory e*thic of excellence*, it highlights the experiences of excellence that are a central part of human fulfillment. During the twenty-first century, the character education field has evolved from the single view that moral character is the only type of character development students should strive to improve. Berger posits that when an ethic of excellence is at the center of school culture, students will do "amazing things" (Lickona & Davidson, 2005, p. 17). It highlights the experiences of excellence that are a central part of human fulfillment: "Character—working hard, persevering—is essential for realizing excellence. Excellence matters, and character matters in our pursuit of excellence. It follows that educating for character ought to be about developing ethics *and* excellence" (Lickona & Davidson, 2005).

> But what does an ethic of excellence look like? Berger (2003) explains:
>
> I think of my life in my small town. The policeman for my town is a former student. I trust him to protect my life; I trust him to work kindly and carefully with the young students in my school, which he does often and does tenderly. The nurse at my medical clinic is my former student. I trust her with my health. The excavator who measured and

> dug the foundation hole for my house, who built my driveway and septic system, is a former student; I built my home on his work. The lifeguard at the town lake is my former student; she watches my grandsons as they swim. There may not be numbers to measure these things but there is a reason I feel so free and thankful trusting my life to these people: They take pride in doing their best. They have an ethic of excellence. (p. 154)

The Student Performance Character and Student Moral Character scales operationalize the construct of character according to this definition. Moral character is defined as the qualities needed for successful interpersonal relationships and ethical behavior. It is described as a relational orientation because it highlights how we treat ourselves, as well as others (Lickona & Davidson, 2005). It follows in the traditions of Kohlberg's (1984) ethic of justice and Gilligan's (1982) ethic of care.

Distinctly different from moral character, performance character is defined as qualities needed to realize one's potential for excellence in academics, co-curricular activities, the workplace, or any other area of endeavor (Lickona & Davidson, 2005). This type of character is called a mastery orientation and is different from the relational nature of moral character. The primary goal is not the outcome of performance, but instead the character strengths and qualities such as effort, diligence, perseverance, a strong work ethic, a positive attitude, self-discipline, etc. that enable students to pursue their personal best (Lickona & Davidson, 2005). Students can display great performance character and still not meet the final goal, as well as succeed without displaying performance character traits. Ultimately, however, performance character does maximize performance because it highlights the qualities and traits that help students challenge themselves to do their personal best in all situations.

These two distinct types of character are operationalized through the lens of three psychological realms. In order to have "good" moral and performance character an awareness of what good ethical behavior and excellence requires is critical. Second, after awareness is established, a caring attitude about ethical behavior and excellence must manifest. Finally, students with the highest levels of moral and performance character must demonstrate actions that strive for ethical behavior and excellence. See Table 4 for a visual breakdown of these three realms.

Lickona and Davidson (2005) report five sources that support the student performance and moral character constructs: (a) research on motivation and talent development, (b) the wisdom of the ages, (c) lives of character, (d) the practices of great character educators, and (e) the voices of high school teachers and students.

First, citing a 5-year longitudinal study of 200 talented teenagers, Lickona and Davidson (2005) report that teens who develop their talent do so through performance character: "The combination of strong performance character, supportive and challenging adults, and the capacity to take pleasure in developing their gifts maximized the likelihood that talented teens fulfilled their potential" (p. 25).

Secondly, Lickona and Davidson (2005) cite quotes from notable, historical figures (including Booker T. Washington, Confucius, and Mother Teresa) to highlight how the wisdom of the ages confirms the necessity of both performance and moral character. Third, they cite researchers who examined the lives of people with strong performance and moral characters, including businessmen, teachers, and social movement leaders: "None of the noble accomplishments of these exemplars would have been possible without the mutually supportive contributions of performance character and moral character" (p. 25).

Another source for the performance and moral character constructs is examining the practices of great character educators, such as John Wooden. As a legendary UCLA basketball coach, Wooden not only held championship records, but was well-known for teaching character principles to his players. Finally, Lickona and Davidson (2005) interviewed high school teachers and students and found that the performance and moral character constructs consistently emerged: "'Performance character' thus gives high school educators a new character language for describing the academic endeavor of teaching and learning that is the focus of their daily work. Of course, good teachers… also pay attention to moral character…" (p. 27).

With these five sources of support for the student performance and student moral character constructs, the scales examined within this study will operationally define character as both moral *and* performance excellence.

Summary

In conclusion, this literature review summarizes how character and character education has been defined and investigated from a variety of perspectives, as well as presents three general categories of instruments: instruments based primarily on moral development theory**,** psychological theory, and both moral and psychological theories. These three categories can be useful angles in which to examine scales/instruments that are related to the Student Performance Character and Student Moral Character scales. Also, there are several similarities, as well as distinct differences, between these scales/instruments that allow for a comparative view. Therefore, this section will consider the validation approaches, theoretical development (stage vs. psychological realms), and internal structure of scales/instruments examined within the literature review.

One similarity among the previously mentioned scales/instruments is the approach to establishing validity. Most developers of these instruments have relied mostly on a classical test theory (CTT) approach. For the SPC and SMC scales specifically, construct validity was established by examining the pattern of relationships between numerous constructs using bivariate correlation analysis based on CTT (Hambleton, et al., 1991). Results show that all relationships between character, character-related experiences, and school/classroom climate scales were positive and statistically significant (Khmelkov & Davidson, 2005-2008). Also, both character-related experiences and Ethical Learning Community scales developed for CREE are much stronger predictors of student performance and moral character, than they are of student academic motivation and learning style. Also, to establish convergent, discriminative, and predictive validity, observed patterns of relationships in the data were compared to the theoretical patterns of relationships between constructs of interest. Results show strong evidence in support of all three types of validity. For example, moral character experiences in school are moderately correlated with moral character (.475, $p < .01$), but show almost no correlation with self-efficacy in math (.005, $p < .01$) and very little correlation with cooperative learning styles (.131, $p < .01$). This pattern of relationships holds for all character-related scales:

Patterns of observed relationships in the data within and between domains, such as performance and moral, as demonstrated by the size of correlation coefficients, correspond completely to the patterns that were expected: performance character-related scales are stronger correlated with each other than they are with moral character-related scales, and the other way around… Somewhat similarly, when two domains of learning style are examined, all character-related scales have positive significant correlations with cooperative learning style, but no significant correlations with competitive learning style. (Khmelkov & Davidson, 2005-2008)

Since the observed patterns of correlation results correspond well to the theoretically expected pattern of relationships, instrument developers propose that appropriate evidence of simultaneous convergent, discriminant, and predictive validity of the CREE instrument was established.

Although the SPC and SMC scales used a classical test theory approach to validation, similar to previous psychological and character scales, some differences do exist. For example, all of the scales/instruments birthed from moral development theory are based on a stage development model (i.e., Kohlberg). The Student Performance Character and Student Moral Character scales, however, are not based on this structure. Instead, the SPC and the SMC scales are based on Lickona and Davidson's (2005) theory of performance and moral character, as well as Berger's (2003) ethic of excellence theory. A stage development theory versus a theory based on psychological realms significantly differentiates the SPC and SMC scales from previously validated character scales.

Furthermore, an instrument's internal structure (the way the items are arranged within the scale/instrument), is a construct rather than an inference for these types of scales/instruments:

Kohlberg's manifest behavior pattern approach contrasts sharply with approaches like classical test theory and Rasch scaling, which regard each response item as an indicator of some unobservable, hypothetical variable or latent entity, and the structure of the individual response pattern only as a sign of measurement error. (Lind, 2008, p. 191)

Lastly, the scales/instruments derived from moral development theory use a complex moral situation to elicit moral judgment competence. This is very different from the scales/instruments birthed from psychological and moral and performance character theories where individual items represent a specific domain. The Student Performance Character and

Student Moral Character scales are related to similar constructs that the aforementioned scales/instruments measure, but it is unique in that it provides an opportunity to further test a new theory (Berger's theory of *Ethic of Excellence*).

In conclusion, there are some similarities between the SPC and SMC scales and the scales mentioned within the moral development and psychological categories, but there are also some major differences that make these scales significant, separate from their historical predecessors. The limited validation studies that have been done for the SPC and SMC scales use a classical test theory approach, similar to other preceding scales, but a more modern approach is necessary to ensure that these scales are measuring student performance character and student moral character accurately. This validation approach will allow for character education evaluations to define and measure outcomes more accurately and make appropriate comparisons between social groups. This is the primary purpose of this study.

CHAPTER IV

Methodology

Instrument

Since 2004, the Collective Responsibility for Excellence and Ethics instrument (CREE) has been used in schools and districts throughout the country. This instrument consists of numerous scales: (a) Student Performance Character, (b) Student Moral Character, (c) The Ethical Learning Community-student Version, (d) Acceptance of Differences in Peers & Attachment to Class/School Community, (e) Collective Responsibility for Class/School Community, (f) Performance Character Experiences in School, (g) Moral Character Experiences in School, (h) Performance Character Experiences at Home, (i) Moral Character Experiences at Home, (j) Intrinsic Interest in Reading, (k) Intrinsic Interest in Writing, (l) Self-Efficacy in Math, (m) Cooperative Learning Style, and (n) Competitive Learning Style.

Various forms of the CREE instrument have been administered to multiple schools that implement federally-funded *Partnerships for Character Education Program* projects (http://www.excellenceandethics.com/assess/cree.php). This study only examines the student form of this instrument: The CREE Student Form. This form was designed to measure (a) school/classroom climate, (b) ethical learning communities, (c) experiences of learning the strengths of character, and (d) student performance and moral character (Khmelkov & Davidson, 2005-2008). The subscales *Student Performance Character* (SPC) and *Student Moral Character* (SMC) are the focus of this study; they measure character according to Lickona and Davidson's (2005) model.

The SPC and SMC scales have three psychological components: awareness, attitude, and action (Lickona & Davidson, 2005). The awareness component is related to cognition. Morally,

this means that students can recognize the ethical dimensions of situations and grapple with important moral questions. Students with performance character can identify “excellence” in a variety of areas and understand the virtues required to pursue “excellence.”

The attitude component is related to “matters of the heart,” or emotions. Morally, this means students care deeply about doing the right thing and have courage of conscience in the face of social pressures. Students with performance character give their best effort and are committed to high-quality work. Finally, the action component is related to behavior and/or habits. Morally, this means students act upon ethical convictions and take a stand for what is right. Students with performance character demonstrate the habits required for excellence and practice these habits in order to improve (see Table 4). In sum, students with moral character know what ethical behavior requires (cognitive), care about ethical behavior (attitude), and strive to act in ethical ways (action). Likewise, students with performance character know what excellence requires (cognitive), care about excellence (attitude), and strive for excellence (action; (Lickona & Davidson, 2005).

Based on these three psychological components, items were used from other scales or developed specifically for the SPC and SMC scales. The SPC scale has 12 items and the SMC scale has 11 items. The SPC scale items include: (a) I can be counted on to do my part for the team/group (CountedOn); (b) I try to get out of doing things that are difficult or boring (AvoidBoring); (c) I spend extra time working to improve my weaknesses (TimetoImprove); (d) I continue trying hard, even when things are not going well (TryHard); (e) I forget to bring what is needed for class (4getMaterials); (f) I work with another student to help him or her do better on an assignment, without letting them copy my work (HelpOthersWk); (g) I forget to do my homework (4getHW); (h) I think about my school work and consider whether I need to work

harder (Reflect); (i) I talk to a teacher to find out if I'm doing well in my school work (AskProgress); (j) I run out of time to do my assignments well (RunOutTime); (k) I give up watching TV or hanging out with friends to study for a test or do an assignment for school (GiveUp4Sch); (l) I am willing to redo a school assignment to make it better (RedoWk). Table 5 displays the items, item numbers (from the actual instrument), item abbreviations, and which items were negatively worded (labeled "reversed"). Also, I categorize these 12 items into the three psychological realms described by Lickona and Davidson (2005) in order to provide a useable framework for conceptualizing items (see Table 6). The Cognitive realm includes 1 item: Reflect. The Attitude realm includes 2 items: TryHard, and RedoWk. Finally, the Action realm includes 9 items: CountedOn, AvoidBoring, TimetoImprove, 4getMaterials, 4getHW, AskProgress, RunOutTime, HelpOthersWk, and GiveUp4Sch.

The SMC scale has 11 items which include: (a) I treat teachers and staff with respect, even if I disagree with them (AdultRespect); (b) I break classroom or school rules (BrkRules); (c) when I see someone having a problem, I offer to help (HelpWithProb); (d) I do the right thing no matter what others might think (DoRight); (e) I help another student choose between doing what is right and what is wrong (HelpOthersRgt); (f) I make fun of someone (Tease); (g) I speak up when someone is bullied (CallOutBully); (h) I cheat on a test or an assignment (Cheat); (i) I think about how my parent, teacher, or coach would act before making an important decision (ThkBoutAdults); (j) I admit if I do something wrong (AdmitWrong); (k) I consider different points of view when making a decision about a moral issue or dilemma (DiffPOVs; see Table 5). Similar to the items on the Student Performance Character Scale, I categorize these 11 items into the three psychological realms described by Lickona and Davidson (2005; see Table 7). The Cognitive realm includes 2 items: ThkBoutAdults and DiffPOVs. The Attitude realm includes 2

items: HelpWithProb and AdmitWrong. Finally, the Action realm includes 7 items: AdultRespect, BrkRules, DoRight, HelpOthersRgt, Tease, CallOutBully, and Cheat.

Khmelkov and Davidson (2005-2008) report that the SPC and SMC scales have reliabilities, or Cronbach alphas, that are moderate to good. The 12 items related to the performance character scale have reliabilities ranging from .71-.75, with .71 for middle school (N = 452), .75 for high school (N = 171), and .72 for middle and high school combined (N = 622). The 11 items for the moral character scale, has an alpha of .76 for middle school (N = 453), high school (N = 171), and combined (N = 625). This indicates that these items have good internal consistency. Some of the items from these scales were used in another instrument (Davidson, 2006).

Finally, evidence supporting the item design was presented through factor analysis. The instrument developers field tested the data using factor analysis. The performance character scale explained 47.5%, 50.8%, and 38.8% of the variance of the items within the middle school, high school, and combined samples, respectively. The moral character scale explained 53.4%, 45.5%, and 52.8% of the variance of the items within the middle school, high school, and combined samples, respectively. This is a classical approach to examining explained variance; this study, however, will assess dimensionality using a Rasch measurement framework.

Participants and Setting

The CREE instrument was used as the measurement tool in a federally-funded character education grant spanning 4 years (personal communication, Joanne Goubourn, September, 2009). As a result, the sample for the original CREE Student Form is vast. Khmelkov and Davidson (2005-2008) report a total sample of 622 students, spanning from middle school to high school. The sample for this study is a subsample of previous analyses and is comprised of

secondary data collected from surveys that were conducted for four different schools at one time point. The first round of data collection was in 2007 and in the Fall of 2008. The two remaining data collection periods were in Spring of 2008 and Spring of 2009. This study focuses on the data collected in Spring 2009 only.

Four different schools were sampled in Spring 2009, and data were collected from 239 students in grades 6 through 8. School A is a 4-year old, medium-sized, public charter school in a large, metropolitan area in the Northeastern region of the United States. It has almost a 100% Black and Hispanic student population and is located in one of the poorest congressional districts in the United States. School B is an 11-year old, large, public charter school in a mid-sized, metropolitan area in the Eastern region of the United States. Serving grades kindergarten through twelve, over 97% of students are African American. Also, 80% of students in grades K-8 eligible to receive free or reduced lunch, while 58% of high school students are eligible. School C is an 11-year old, small, public charter school in a mid-sized, metropolitan area in the Eastern region of the United States. Serving grades Pre-kindergarten through eight, this school is almost 100% African-American. Finally, School D is an 8-year old, medium-sized charter school located in the Northeastern region of the country. Serving grades kindergarten through eight, over 90% of students are African American. About 70% of students receive free or reduced lunch (see Table 8).

Procedures

This section will describe the procedures used in this study. Specifically, it will present five major analytic steps: (a) an investigation of gender and grade differences in the outcomes, (b) a diagnosis of the use of the rating scales, (c) an examination of the indices and fit statistics

associated with Rasch models, (d) a description of the dimensionality principle, and (e) the definition of differential item functioning.

Gender and grade differences. Independent *t*-tests, one-way analysis of variance (ANOVA) tests, and averages will be examined for key variables. Specifically, *t*-tests will show if males or females show a significant difference in their mean scores on performance and/or moral character compared to their counterpart. Additionally, one-way ANOVA testing will determine if there is a significant grade-related difference in the scores for both scales.

Rating scale diagnostics. The two most widely used Rasch models for polytomous data are the Rating Scale model (RSM) and the Partial Credit model (PCM; Iramaneerat, et al., 2008).Therefore, this study applies these two types of Rasch model approaches. More specifically, the Rating Scale model was developed to analyze rating scale data with a fixed number of response categories across a set of items designed to measure a unidimensional construct (Engelhard, 2005). In other words, it assumes equal threshold structures between rating scale categories. Generally, the higher the number on the rating scale, the more evident the latent variable (Andrich, 1988). A rating scale item is usually intended to be governed by the fixed set of rating points with the items. As the same set of rating points is used with each item, it is thought that the relative difficulties of the steps in each item should not vary from item to item (Wright & Masters, 1982).

The Rating Scale model has two parameters: a location parameter (λ_i) and a category parameter (δ_j). The location parameter reflects item difficulty, while the category parameters are equivalent across items (Engelhard, 2005). The additive equation is

$$\pi_{nix} = \frac{exp\sum_{j=0}^{x}[\beta_n-(\delta_i+\tau_j)]}{\sum_{k=0}^{m} exp\sum_{j=0}^{k}[\beta_n-(\delta_i+\tau_j)]} \quad \text{x=0,1,…,m} \qquad (1)$$

where $\tau_0 = 0$ $exp\sum_{j=0}^{0}[\beta_n - (\delta_i + \tau_j)] = 1$.

Wright & Masters (1982) state that when this model is applied to a rating scale analysis a position on the variable β_n is estimated for each person *n*, a scale value δ_i is estimated for each item *i*, and *m* response "thresholds" $\tau_1, \tau_2, \ldots, \tau_m$, are estimated for the $m + 1$ rating categories. Specifically, the rating scale threshold locations correspond to the transition across adjacent categories *k*-1 and *k*. Linacre (2004) described the rating scale categorizations as ways to elicit unambiguous, ordinal indications of the locations of respondents along a latent variable of interest.

For example, the Student Performance and Student Moral Character scales both utilize a Likert scale with 5 categories (Almost Never, Rarely, Sometimes, Often, Almost Always). A student responding to these ordered response categories chooses to complete the "k'th" step; in other words, they choose the "k'th" alternative (e.g., rarely) over the (k-1)'th (e.g., sometimes) in response to the item (Wright & Masters, 1982).

Unlike the Rating Scale model, the Partial Credit model does not assume an equal threshold structure. Instead, it assumes a unique structure for each item (Wright & Masters, 1982). This model allows for the possibility of having differing numbers of steps for different items on the same test (Bond & Fox, 2001). Unlike the commonly held view of "partial credit" on an educational assessment, the Rasch Partial Credit model requires that the "part marks" be awarded in an ordered way, where the increasing value represents an increase in the latent variable (Bond & Fox, 2001). If several, ordered performance levels are identified *within* an

item, the item is a multi-step item. Although the probabilistic formula begins similarly to the Rating Scale model, with the probability of student *n* scoring 1 on an item, it also demonstrates more than one ordered performance level (δ_{ij}). The additive equation is

$$\pi_{nix} = \frac{exp \sum_{j=0}^{x}(\beta_n - \delta_{ij})}{\sum_{k=0}^{m} exp \sum_{j=0}^{k}(\beta_n - \delta_{ij})} \quad x=0,1,\ldots,m \qquad (2)$$

where $\delta_{i0} = 0$ so that $\sum_{j=0}^{0}(\beta_n - \delta_{ij}) = 0$ and $\exp \sum_{j=0}^{0}(\beta_n - \delta_{ij}) = 1$.

This is different from the Rating Scale model in that there is no threshold parameter (or τ) represented, but instead demonstrates more than one item characteristic curve (Wright & Masters, 1982).

An example of how the Partial Credit model can be applied is with the latent variables student performance and moral character. If the threshold structure of each item varies, then the examination of item fit using the PCM may be more accurate.

One of the first steps in a Rasch analysis is to examine how respondents are using the rating scale. Too many response categories can violate model expectations, create noise, and decrease reliability. Also, too few observations in each rating category may yield inaccurate interpretations. Thus, rating scales should reflect careful consideration of the latent construct and convey categories and labels that elicit unambiguous responses (Bond & Fox, 2001). Examining category probability curves, threshold estimates, and scoring category transitions will provide an understanding of how the rating scale is utilized by respondents.

Model data fit. There are several statistical indices and graphical displays that provide information in order to answer this study's research questions. First, it is important to examine the calibrations of the student facet to determine how high or low each rating is. Second, the

standard error for each student verifies how precisely each student has been calibrated along the latent variable. Also, a variable map, also called a Wright map, helps display where students are located on the latent variables student performance character and student moral character.

In addition to calibrations, standard error values, and variable maps, there are also three indices that provide information about the scales: (a) the reliability of separation index, (b) the chi-square statistic, and (c) mean square error statistics (INFIT/ OUTFIT).

Reliability statistics. The reliability of separation index is a useful statistic that provides descriptive information regarding how well the students are separated on a linear continuum in order to reliably define student character. It reflects the true score variance to observed score variance. The reliability of separation index for students, for example, can be calculated as

$$R = \frac{SD^2 - MS}{SD^2} \tag{3}$$

where SD^2 is the observed variance of the students' ratings on the latent variable scale in logits and *MS* is the mean of the calibration error variances for each student's response.

Chi-square statistics. The log-likelihood chi-square statistic is a global fit index that helps to determine the best model to fit the data. Although Rasch analyses typically use other parameter-level indices to fit the data to the model, the log-likelihood ratio test is still useful to choosing the model that fits the data best. In this study, a log-likelihood ratio test allows for the comparison between the Rating Scale and Partial Credit models by examining the difference between the chi-square values and degrees of freedom for both models. In other words, it provides evidence regarding whether the Rating Scale model or the Partial Credit model is the best fit, given the data. It can be calculated as

(Rating Scale model log-likelihood chi-square – Partial Credit model log-likelihood chi-square)

with (Rating Scale model – Partial Credit model) degrees of freedom

Mean square error statistics. The mean square error statistics, also called INFIT/ OUTFIT statistics, are used to evaluate the consistency of student ratings. The OUTFIT statistic, specifically, represents the unweighted mean squared residuals that are particularly sensitive to outlying unexpected ratings. This index shows how closely the data fit the Rasch model. The INFIT values are based on weighted mean squared residuals, and they are less sensitive to outlying unexpected ratings.

For example, the Mean Square Error (*MSE)* statistic for Student *N* can be calculated as

$$MSE_n = \frac{\sum_{i=1}^{I} z_{ni}^2}{(I)} \tag{4}$$

When the data fit the model, the expected value is 1.0. Linacre (2007) interprets values less than 1.0 as indicative of observations that are too predictable and overfit the model, while values greater than 1.0 indicate unpredictability and underfit the model. In other words, *MSE* values greater than 1 demonstrate a student who has more variation in their ratings than the Rasch model predicted. A *MSE* value less than 1 means a student is responding to items holistically and not discriminating enough in their ratings. Generally, the range for good model-data fit is .5 < *MSE* < 1.5 (Linacre, 2007). Many scholars use a range that is a little more restrictive where .6 < *MSE* < 1.5 (Engelhard, 1998; Lunz, 1990); this study will use a slightly more restrictive range where .6 < *MSE* < 1.4. Consequently, an OUTFIT mean square value of 1.60, for example, indicates 60% more variation in the observed data than the Rasch model predicted (1 + .60 = 1.60), while a *MSE* value of .40 indicates 60% less variation (1- .60 = 0.40). In both cases, the data do not fit the model according to the range selected.

Dimensionality analyses. One of the central tenets to the Rating Scale and Partial Credit Rasch models is the assumption of unidimensionality. This fundamental measurement principle assumes that only one dimension or attribute is being measured at a time (Bond, 2001). Some researchers suggest that Rasch analyses should be conducted after determining unidimensionality through factor analysis (Bonk & Ockey, 2003). However, performing this analysis first may yield misleading results because factor analyses use raw scores; it is preferable to use Rasch analyses first in order to investigate dimensionality because it produces a linear continuum (Bonk & Ockey, 2003). The fit statistics will show any aberrations from unidimensionality. Also, applying principal component analyses on the standardized residuals can yield more accurate estimates of the subsequent factors because they will indicate the structure of the underlying dimensions (Wright, 1996).

Differential item functioning (DIF) analyses. A final step in investigating the quality of the SPC and SMC scales is to compare the estimates across two or more distinct groups of interest. DIF detects whether the invariance expected under the model's requirements are substantiated empirically or if some sort of bias exists (Bond & Fox, 2001). In this study, DIF is examined by modeling the invariance of item difficulty estimates by comparing item difficulties across the male and female students.

Limitations

This validation study has several limitations. The primary limitation is that the sample size is not as large as those in many other validation studies. Also, because the survey data were self-reported, there may be inaccuracies or skewed perceptions at play; students may have answered items based on what was socially desirable and not necessarily the basic truth.

Furthermore, the sample was not diverse enough to allow one to study other types of DIF, such as ethnic/racial differences in item functioning.

CHAPTER V

Results

Gender and Grade Differences

The sample for this study consisted of 239 middle grade students. Table 8 shows that 92 of them were males and 134 were females (there were no gender data for 13 students). The grade distribution is as follows: 163 were in sixth grade, 56 in 7th grade, and 9 in 8th grade (eleven students had no grade information). To investigate whether there were gender and grade differences in the two character scales, I performed t-tests and analysis of variances (ANOVA).

First, *t*-test results show that the mean score on the performance character scale for females was significantly different from that of the males ($t_{(222)} = -2.03$, $p < .05$). However, there was no gender difference in the mean scores on the moral character scale ($t_{(223)} = -1.28$, $p > .05$; see Table 9). One-way ANOVA analyses were also conducted and they revealed no significant grade effects for the SPC [$F(2, 219) = 2.32$; $p > .05$] and SMC [$F(2, 220) = 2.89$; $p > .05$] scale (see Tables 10 and 11). In other words, there were no statistically significant differences in the mean scores among any of the three grades (6, 7, and 8) surveyed on student performance character and student moral character measures. Finally, the magnitude of the grade effect for the SPC scale was computed as $R^2 = .02$ and $R^2 = .03$ for the SMC scale.

Rating Scale Diagnostics for SPC Scale

Not all types of data are suitable for Rasch analysis. In order to make valid inferences from this analysis, the data have to meet certain model requirements (Iramaneerat, et al., 2008). Therefore, one of the central research questions of this study is, “Do the items on the Student Performance Character and Student Moral Character scales fit the Rasch measurement model framework and, if so, how well is the fit?

In evaluating how well the polytomous data fit the Rasch model, I follow Linacre's (2004) fundamental guidelines: (a) the orientation of the items, (b) the frequencies of use, (c) the distribution of the observations, (d) the monotonicity in the increase in average measures, (e) the category fit, (f) the monotonicity of thresholds, and (g) the magnitudes of the distances from the thresholds. These guidelines serve as the structure for reporting the following rating scale diagnostic results.

Orientation of items (SPC). All items on the SPC scale have positive point-biserial correlations (see Table 12). Items AvoidBoring (I try to get out of doing things that are difficult or boring), 4getMaterials (I forget to bring what is needed for class), 4getHW (I forget to do my homework), and RunOutTime (I run out of time to do my assignments well) were all negatively worded. If these items were not rescored, and their polarities did not conform to other items, they would skew the rating scale. Since all correlations are positive, results indicate that these negatively-worded items were rescored accurately and properly reflect the rating scale.

Frequency of use (SPC). In order to get accurate threshold calibrations, at least 10 observations are needed within each rating category. Table 13 shows an observed count of 101 observations in Category 1, 328 in Category 2, 967 in Category 3, 813 in Category 4, and 470 observations in Category 5. Therefore, categories are frequently used and add to the rating scale functioning.

Distribution of observations (SPC). Table 13 shows Category 1 (Almost Never) is used the least amount with only 101 observations, while Category 3 (Sometimes) has the highest number of observations (967). Irregularity in the frequency across categories may indicate abnormal category use. Since the observations are not evenly distributed across categories, scale revisions should be made. Combining categories will improve rating scale functioning.

Monotonicity of average measures (SPC). Another element that contributes to rating scale functioning is the monotonicity of the observed average measures. In other words, the observed average logit measures should increase with each increasing rating scale category. If not, then categories show disordering and produce an uncertainty in the meaning of the rating scale. The SPC scale has an observed average of -.19 logits for Category 1, .01 logits for Category 2, .32 logits for Category 3, .73 logits for Category 4, and 1.29 logits for Category 5 (see Table 13). Since each category's average logit measures increase with the increase in categories (from 1 to 5), no disordering exists. Therefore, advancing categories represent more of the latent variable student performance character.

Category fit (SPC). Table 13 shows the OUTFIT Mean Square values for each category. This statistic indicates whether each rating category is used as expected. When the data fit the model, the expected value is 1.0. Categories 2 (Rarely) and 5 (Almost Always) have outfit values of .99, while Category 4 (Often) has an outfit value of .98; this indicates that the data fit the model and categories are being used as expected. Categories 1 (Almost Never) and 3 (Sometimes) have outfit values that do not fit the model (1.28 and .85, respectively). This indicates that Category 1 underfits the model and demonstrates unpredictability in the observations, while Category 3 overfits the model and demonstrates observations within this category are too predictable.

Monotonicity of thresholds (SPC). Table 13 also shows how rating scale category thresholds advance from one category to the next (the column labeled Structure Calibratn). If the rating scale functions properly, each threshold will increase as categories increase, indicating a higher amount of the latent trait (student performance character). Results show the threshold estimate between Categories 1 and 2 is -1.33 logits, between Categories 2 and 3 is -.91 logits,

between Categories 3 and 4 is .71 logits, and between 4 and 5 is 1.53 logits. Therefore, category thresholds all increase monotonically as categories advance.

Distance from the thresholds (SPC). Linacre (2004) states that thresholds should advance by at least 1.0 logits for a five-category rating scale. Results show the distance between Category 2 and Category 3 thresholds is .42, the distance between Category 2 and 3 thresholds is .2, and the distance between Category 4 and Category 5 thresholds is .82. Therefore, although category thresholds all increase monotonically as categories advance, they do not meet Linacre's guideline of advancing by at least 1.0 logits. This indicates that categories are practically inseparable. Combining categories will improve the functioning of the rating scale.

Log-likelihood chi-square (SPC). A log-likelihood ratio test for the Rating Scale model and the Partial Credit model will help identify which model fits the data best. The null hypothesis is that the Partial Credit model fits no better than the Rating Scale model. In order to test this hypothesis, the difference between the chi-square values from both models and the difference between the degrees of freedom must be examined.

Analyses for the SPC scale suggests statistically significant results [RSM: χ_2(2440 df, N = 12) = 6628.16, and PCM: χ_2(2407 df, N = 12) = 6530.45, $p < .05$]. The difference between the Rating Scale model and Partial Credit model log-likelihoods is 97.71 with a difference of 33 degrees of freedom between the two models. The null hypothesis, therefore, is rejected, indicating that the Partial Credit model fits better than the Rating Scale model.

Rating Scale Diagnostics for SMC Scale

Orientation of items (SMC). All items on the SMC scale have positive point-biserial correlations (see Table 14). Items BrkRules (I break classroom or school rules), Tease (I make fun of someone), and Cheat (I cheat on a test or an assignment) were all negatively worded.

Since all correlations are positive, results indicate that these negatively-worded items were rescored accurately and properly reflect the rating scale.

Frequency of use (SMC). Table 15 shows an observed count of 88 in Category 1, 210 in Category 2, 836 in Category 3, 728 in Category 4, and 595 observations in Category 5. Since all 5 categories have more than 10 observations, the categories are used enough to meet the requirements for rating scale functioning.

Distribution of observations (SMC). Table 15 shows Category 1 (Almost Never) is used the least amount with only 88 observations, while Category 3 (Sometimes) has the highest number of observations (836). Low category usage signals a potential problem with the way the rating scale functions. Since the observations are not evenly distributed across categories, combining categories (Almost Never and Rarely, for example) can improve rating scale functioning.

Monotonicity of average measures (SMC). Another factor to consider in examining the rating scale's functioning is the monotonicity of the observed average measures. Scale developers must investigate whether observed average logit measures increase with each increasing rating scale category. The SMC scale has an observed average of -.34 logits for Category 1, -.07 logits for Category 2, .31 logits for Category 3, .96 logits for Category 4, and 2.00 logits for Category 5 (see Table 15). Since each category's average logit measures increase from Categories 1 through 5, no disordering exists. Therefore, advancing categories represent more of the latent variable student moral character.

Category fit (SMC). Table 15 shows the OUTFIT Mean Square statistics for each category. Category 5 (Almost Always) has an outfit value of .98, indicating the data fits the model and categories are being used as expected. Categories 3 (Sometimes) and 4 (Often) have

outfit values of .92 and .86, respectively, indicating that these categories slightly overfit the model and observations within these categories may be marginally too predictable. Finally, Categories 1 (Almost Never) and 2 (Rarely) have outfit values that underfit the model and demonstrate unpredictability in the observations (i.e., 1.19).

Monotonicity of thresholds (SMC). Table 15 also shows how rating scale category thresholds advance from one category to the next. If the rating scale functions properly, each threshold will increase as categories increase, indicating a higher amount of the latent trait (student moral character). If thresholds fail to advance monotonically, threshold disordering has occurred. Results show the threshold estimate between Categories 1 and 2 is -1.18 logits, between Categories 2 and 3 is -1.27 logits, between Categories 3 and 4 is.79 logits, and between 4 and 5 is 1.66 logits. It is evident that category thresholds do not increase monotonically from Category 1 to Category 2. This disordering can degrade the interpretability of resulting measures. Combining Categories 1 (Almost Never) and 2 (Rarely) can improve threshold ordering and how the rating scale functions.

Distance from the thresholds (SMC). This final criterion in evaluating rating scale effectiveness requires that distances between thresholds are not too large or not too small. If distances are less than 1.0 logits or more than 5.0 logits, interpretations about the latent variable can be skewed. Results show the distance between Category 2 and Category 3 thresholds is .09, the distance between Category 2 and 3 thresholds is .48, and the distance between Category 4 and Category 5 thresholds is .87. It is evident that categories do not meet Linacre's (2004) guideline of advancing by at least 1.0 logits. These results indicate that categories are practically inseparable and do not represent a higher amount of student moral character as categories increase.

Log-likelihood chi-square (SMC). The ratio test is used to see whether the Partial Credit model fits no better than the Rating Scale model. The results suggest that the Partial Credit model fits better than the Rating Scale model for the SMC scale [RSM: $\chi 2(2209\ df, N = 11) = 5509.93$, and PCM: $\chi 2(2179\ df, N = 11) = 5392.20$, $p < .05$]. The difference between the Rating Scale model and Partial Credit model log-likelihoods is 117.73 with a difference of 30 degrees of freedom between the two models. The null hypothesis, therefore, is rejected, indicating that the Partial Credit model fits better than the Rating Scale model.

Model Data Fit

In evaluating how well the rating scale functions, there are certain criteria used to examine model data fit for both dichotomous and polytomous data. These include: (a) reliability statistics, (b) mean square error statistics, and (c) the variable maps. Calibration results from both the Partial Credit and Rating Scale models were included for comparison purposes.

Reliability statistics. The Rasch reliability indices for the two models were also compared. The person reliability statistic is similar to Cronbach's alpha for person separability. This shows how well the scale can differentiate students on the latent variable performance character and moral character. The Rating Scale and Partial Credit model report the same person reliabilities of .66 for the SPC. The RSM model reports a .73 person reliability separation for the SMC, while the PCM model reports .74 (see Table 16). The separation of items refers to the ability to define a specific hierarchy of items along the latent construct. Table 17 shows how items separate within the Rating Scale model analysis (.96 and .99 for SPC and SMC respectively) and the Partial Credit model analysis (.96 and .97, for both the SPC and SMC, respectively).

Mean square error statistics. Another central research question of this study is, "How well do items on the Student Performance Character and Student Moral Character scales measure the two constructs?" In order to examine the utility of the Rasch model, it is also necessary to examine item fit using mean square error statistics. Therefore, the patterns of item fit for both scales were examined within the Rating Scale and Partial Credit models.

Tables 18-21 display item fit values for the SPC and SMC scales using both the Rating Scale and Partial Credit models. Rating Scale analyses for both the SPC and SMC scales show all items with OUTFIT statistics falling within the $.6 \leq MSE \leq 1.4$ range (see Tables 18 and 19). Partial Credit analyses for the SPC scale show all items with OUTFIT statistics that fall within the above range. The SMC scale, however, shows that within the Partial Credit model the item CallOutBully falls outside the range with an OUTFIT value of 1.46 (see Tables 20 and 21).

Variable maps. In addition to item fit statistics, Figures 1-4 depict variable maps, also called Wright maps, which provide pictorial information regarding item fit. The Wright maps display visual pictures of the calibrations for the student and item facets. Specifically, it shows the variation in students' responses and the distribution of items on the latent variables student performance character and student moral character scales. Figures 1 and 2 show the distribution of items for the SPC scale (Rating Scale and Partial Credit, respectively), where the item *GiveUp4Sch* has the highest logit score and the item *CountedOn* has the lowest. Figure 3 and 4 show the distribution of items for the SMC scale using the two Rasch analyses. The Rating Scale analysis displays the item *CallOutBully* with the highest logit score and the item *Cheat* with the lowest. The Partial Credit analysis also shows the item *CallOutBully* with the highest logit score, but it shows that items *HelpOthersRgt*, *HelpWithProb*, and *ThkBoutAdults* all have equally high logit scores.

Summary

As there are many variations of Rasch models to choose from within the Rasch models family, deciding which Rasch model is most appropriate is more of a practical decision than a statistical one (Linacre, 2011). In examining rating scale diagnostics and model data fit, it seems that there is great overlap between the Rating Scale and Partial Credit analyses and results. Although the chi-square ratio test indicates the PCM as the better model, it must be considered amongst all other evidential support. For example, most of the item fit values, person and item reliabilities, and variable maps show almost identical results for both models. Although the chi-square ratio test suggests the PCM as the better model, it is the meaning of the measures that motivates the choice of model (www.rasch.org/rmt/rmt1231.htm). The Rating Scale model, compared to the Partial Credit model, is more parsimonious. Therefore, it is more appropriate to choose the Rating Scale analyses and results to answer this study's remaining research questions.

Dimensionality Analyses

Investigation of explained variance. Principal component analyses were conducted in order to investigate variance. Analyses indicate that only about 30% of the raw variance is explained by the SPC Rasch measure and 41% is explained by the SMC measure (See Tables 22 and 23). Investigation of the eigenvalue will help determine if the unexplained variance is a secondary dimension or just random noise and the scree plots (see Figures 5 and 6) provide a visual representation of the log-scale variances of different components.

Examination of residual plots. According to Rasch model simulations, the unexplained variance of the first factor (or contrast) should be no more than 2.0 eigenvalues (Linacre, 2011). First, Table 22 displays an eigenvalue of approximately two for the first and second contrasts on the SPC scale; Contrasts 3, 4, and 5 have an eigenvalue of approximately one. Table 23 displays

an eigenvalue of approximately two for the first and second contrasts on the SMC scale; Contrasts 3, 4, and 5 have an eigenvalue of approximately one. Also, Figures 7-9 display plots of standardized residual principal component analyses (PCA). Figure 7, specifically, shows items 4getMaterials (A) and 4getHW (B) in the upper left quadrant, items RunOutTime (C) and AvoidBoring (D) in the upper right quadrant, items CountedOn (E), TryHard (f), and RedoWk (d) in the lower left quadrant, and all remaining items in the lower right quadrant [Reflect (F), TimetoImprove (a), AskProgress (b), HelpOthersWk (c), and GiveUp4Sch (e)].

Standardized residual PCA plots for the *Student Moral Character* scale for Contrast 1 show most items loading in the lower right quadrant [HelpWithProb (F), DiffPOVs (a), HelpOthersRgt (b), CallOutBully (c), AdmitWrong (d), and ThkBoutAdults (e)]. The upper left quadrant has the second most residuals [BrkRules (A), Tease (B), Cheat (C)], while the upper right quadrant has only one residual [DoRight (D)], and the lower left quadrant has virtually none. Item AdultRespect (E) borders the upper and lower left quadrants (see Figure 8).

The PCA residual plots for Contrast 2 of the SMC scale is evenly distributed between the upper left [BrkRules (A), Tease (B), and AdultRespect (E)], upper right [DiffPOVs (a) and ThkBoutAdults (e)], and the lower right quadrant s [HelpOthersRgt (b), HelpWithProb (F), AdmitWrong (d), DoRight (D), and CallOutBully (c)]. There is only one item residual in the lower left quadrant [Cheat (C); see Figure 9].

Differential Item Functioning (DIF) Analysis

The final research question examines whether any of the items on the SPC and SMC scales display gender differential item functioning (DIF) and, if so, whether there is a particular pattern, direction, and/or magnitude to evaluate further? In order to investigate whether item difficulties differed significantly for males versus females, descriptive analyses were first

performed. As previously mentioned, Table 9 displays the mean ratings on the SPC and SMC scales for males and females. Results were analyzed using an independent-samples *t* test. This analysis revealed that the mean performance character for females is significantly different (and higher) than males ($t(222) = -2.03$, $p < .05$). The mean moral character for females, however, is not significantly different than males ($t(223) = -1.28$, $p > .05$). Also, Winstep analyses show that items do not function differentially for male versus female students. After applying Bonferroni's correction (necessary due to multiple comparisons being made), item difficulty estimates are not found to be significantly different for males and females; therefore, there is no evidence suggesting that DIF exists (see Tables 24-25).

Chapter VI

Discussion

This chapter outlines each research question and examines the following results: (a) person and item reliabilities, point-biserial correlations, item difficulty order (research question 1); (b) item fit and dimensionality (research question 2); (c) gender DIF (research question 3); and (d) implications and scale revisions. I will address each research question posed in the beginning of the study using this outline.

Research Question 1

The first research question is, “How well do items on the Student Performance Character and Student Moral Character scales measure the two constructs?”

To discuss this question, I will examine person reliability, point-biserial correlations, and the psychological orientation of items (cognitive, attitude, or action). Each analysis will provide valuable information regarding the degree to which items measured the desired constructs.

Utilizing the Rating Scale model (RSM), the person reliability statistic for the SPC and SMC scales are .66 and .73 respectively. This statistic is an estimate of how well one can differentiate persons on the latent variable; it replicates person placement across items measuring the same construct. Results suggest that both scales moderately differentiate persons along the latent constructs. The item reliabilities for the SPC and SMC scales are .96 and .99, respectively (RSM).This statistic indicates the replicability of item placements along the latent construct, if the same items were given to another sample with comparable ability levels. Results show extremely high reliabilities. Therefore, we can infer that items were developed with a solid order, where some items are easier to endorse and others are more difficult to endorse, placing

confidence in the consistency of those inferences. Finally, point-biserial correlations are all positive and suggest sufficient homogeneity within the constructs.

Finally, I labeled items as cognitive, attitudinal, or action, according to Lickona and Davidson's (2005) psychological framework. Wilson (1997) argues that the central idea in mapping items to a construct is to examine whether there is a qualitative order of levels inherent in the construct and they demonstrate a continuum from more to less. Examining these psychological realms, along with each items' logit scores, can help to assess the validity of the two constructs under study.

First, the SPC item with the highest logit score in the RSM analyses is *GiveUp4Sch* (I give up watching TV or hanging out with friends to study for a test or do an assignment for school) at .74. This action item may have been the most challenging for middle school students to answer because perhaps they could not distinguish between "giving up watching TV" and "hanging out with friends." This item confounds two very different concepts and could easily confuse adolescents. Also, students may be confused about which scenario they should conceptualize: "study for a test" or "do an assignment for school." Since this item produces responses that are unpredictable, it shows a response pattern that raises concerns. In order to yield student responses that accurately represent the latent construct, this item may need to be reworded or omitted.

The item with the lowest logit score is *CountedOn* (I can be counted on to do my part for the team/group). As a veteran teacher with over 10 years of classroom experience, I can infer that this action item may have been easier for students to endorse because collaborative grouping is often a preferred method of learning for adolescents because it allows them to socialize with their peers. Therefore, this item meets the expectations and measures the SPC construct well.

The item with the second highest logit score is *AvoidBoring* (I try to get out of doing things that are difficult or boring)) at .46. This action item may be difficult to endorse than others because it confounds the constructs “difficult” and “boring.” This item could be confusing for students to interpret because it seems to be asking two separate questions. Also, this item has a high OUTFIT value (1.37) and almost underfits the model. In other words, Item *AvoidBoring* produced responses that indicated unpredictability.

The remaining SPC items fall in the middle of the two extremes. An argument can be made that these items more directly relate to a psychological trait (e.g., perseverance, responsibility, etc.) than those items that underfit the model (i.e., *AvoidBoring).* For example, Item *TimetoImprove* (I spend extra time working to improve my weaknesses), Item *TryHard* (I continue trying hard, even when things are not going well), and Item *RedoWk* (I am willing to redo a school assignment to make it better) can be operationalized actions of diligence. On the other hand, certain items may fall within the .6 to 1.4 range because they are worded in a way where students can discriminate between the categories in a more even distribution. For example, Item *HelpOthersWk* (I work with another student to help him or her do better on an assignment, without letting them copy my work) specifically states a phrase (“without letting them copy my work”) that may resonate with kids as a negative thing to do; therefore, these items measure the SPC construct quite well. There is no disordering of categories, thus students are able to evenly discriminate between categories.

Also, the literature review in this study briefly describes the dual aspect model, where moral ideals lead to moral competencies, which then lead to moral action (Lind, 2008). In reviewing the SPC items, the action items range from easier to endorse (e.g., *CountedOn*) to less endorsability (e.g., Item *GiveUp4Sch*). Similarly, SMC items also show action items as easier to

endorse (e.g., Item *Cheat*) and more difficult to endorse (e.g., Item *CallOutBully*). This does not follow theoretical expectations, according to Lind, where awareness precedes attitude, which precedes action.

Specifically, investigating item fit for the SMC scale, the variable maps (see Figures 3 and 4) indicate that *Cheat* (I cheat on a test or an assignment) is the easiest item to endorse. This may be because students are aware that cheating is bad and may respond in the most socially desirable way. Similar to items on the SPC scale, Item *Cheat* specifically states a phrase that may resonate with kids as a negative thing to do. As a result, they may not respond as honestly due to hypothesized administrative action.

Examining item fit for the SMC scale further, the hardest item to endorse falls within the action psychological realm: *CallOutBully* (I speak up when someone is bullied). There are several hypotheses for why this item was extremely difficult for adolescents to endorse. One explanation for this aberrant pattern is that constructs, such as bullying, are newer to the character field and not yet thoroughly defined within this context. On the other hand, students may not have a clear definition of what it means to "bully." More recently, school districts around the United States have been educating teachers, students, and parents around "bullying" issues; many districts have adopted a "zero tolerance" policy regarding bullying. Nevertheless, there are still many grey areas when it comes to defining and identifying bullying acts (e.g., cyber-bullying is a relatively new and ever-increasing form of bullying). Finally, speaking up against bullying requires a lot of courage and bravery. Adolescents, who tend to have an extremely strong desire to please their peer group, may not have the courage it takes to address a bullying act. If a moral injustice, such as bullying, occurs amongst adolescents, they may be less

likely to speak up and more likely to support the status quo of fellow peers through silence. This could explain why this item was difficult for adolescents to endorse.

In sum, similar to the SPC scale, item endorsability for the SMC scale did not follow many of the theoretical expectations; some of the items with higher endorsability were action items, while some of the items with the lowest endorsability were also action items. This does *not* follow the Lind's dual aspect model theory very well. There was also some item disordering between the remaining psychological realms (cognitive and attitudinal) for both scales. In conclusion, item revisions, as well as adding more items to the cognitive and attitude realms, could be useful.

Research Question 2

Do the items on the Student Performance Character and Student Moral Character scales fit the Rasch measurement model framework, and if so, how well is the fit?

Student Performance Character item fit. All of the SPC items, using the Rating Scale model, show mean square error statistics (OUTFIT values) that fall within the acceptable, restrictive range of .6 to 1.40. The item with the highest logit score is *GiveUp4Sch* (I give up watching TV or hanging out with friends to study for a test or do an assignment for school) at .74. According to the variable map (see Figure 1 and 2) this is the least difficult item to endorse on this scale. This may be because the item confounds two very different concepts and could easily confuse adolescents.

Student Moral Character item fit. All of the SMC items show mean square error statistics (OUTFIT values) that fall within the acceptable, restrictive range of .6 to 1.40, except for Item *CallOutBully* (I speak up when someone is bullied). Using the Partial Credit analyses to calibrate ratings, Item *CallOutBully* has a logit score of .62 and an OUTFIT value of 1.46 (the

Rating Scale model indicates a logit score of .94 and a 1.36 OUTFIT value). This means that Item *CallOutBully* underfits the model for both Rating Scale and Partial Credit analyses, indicating unpredictability in its measurement of the student moral character construct.

Dimensionality. More specifically, dimensionality results provide information regarding item fit to the Rasch measurement framework. Principal component analyses show that the variance explained by the SPC and SMC measures was 30% and 41%, for the performance and moral character scales, respectively. In other words, 70% of the variance explained in the SPC scale is by other constructs, while 59% of the variance is unexplained in the SMC scale. This is not as high as scale developers may anticipate. To determine if the unexplained variance is random noise or a secondary dimension, an investigation of the eigenvalues and residual plots are necessary.

The eigenvalues for the first contrasts of the SPC and SMC scales are 2.2 and 2.1 , respectively. This may suggest that the scales are unidimensional since they are only marginally over the standard 2.0 cutoff. Figure 7, the SPC Contrast 1 residual plot, shows item residuals scattered randomly, without any clear clustering; this suggests random noise (Linacre, 2011). The SMC Contrast 1 residual plot (Figure 8), however, demonstrates a pattern (most of the residuals are concentrated in the lower right quadrant), indicating the presence of a secondary dimension. Therefore, the data for the SPC construct can be modeled as a unidimensional construct, but the data for the SMC construct should be modeled as a multidimensional construct.

Although some of the PCA results may suggest that the SMC scale is unidimensional, other, more salient aspects (the residual plot) point toward a multidimensional model. This conflicting interpretation indicates that although the data for the SPC scale fit the Rasch model to an acceptable extent, the data for the SMC scale do not because the assumption of

unidimensionality was not met. Further research is needed on the dimensionality of the SMC scale.

This study also examines Rasch statistics and fit indices. After reviewing rating scale diagnostics and model-data fit for both SPC and SMC scales, it was determined that the Rasch Rating Scale model, by definition, is more parsimonious to use than the Rasch Partial Credit model. Although the chi-square statistics suggest otherwise, the reliability indices, patterns of item fit, scoring category transitions, and small sample size all indicate that the Rating Scale model is a better model choice to fit the data.

Overall, the majority of the items on both scales fit the Rasch measurement model framework in terms of person and item reliabilities, but principal component analyses suggest multidimensionality for the SMC scale. Variation between schools sampled, item wording, or confusion about the latent construct are all potential sources of this multidimensionality. Nevertheless, results suggest that the items on the SPC scale fit the Rasch measurement model framework, but the items on the SMC scale do *not*. In order to improve item fit, more research on scale revisions is needed.

Research Question 3

Do any of the items display gender differential item functioning (DIF)? If so, what are their patterns, directions, and magnitudes?

The last research question investigates whether or not any of the items display gender differential item functioning (DIF). DIF analyses show that items do *not* function differentially for male versus female students; item difficulty estimates are the same for males and females. Also, although large sample sizes are generally preferred for DIF analyses, useful information can be obtained from samples less than 200 when the Rasch model is used (Lai, Teresi, &

Gershon, 2005; Wright, 1977). In conclusion, no DIF exists between males and females on the SPC and SMC scales.

Implications

This study has several implications. First, results from the Student Performance Character scale provide support for Berger's (2003) ethic of excellence theory. Therefore, the construct underlying the SPC scale matches the theoretical understandings used to develop the construct. Second, analyses suggest multidimensionality within the SMC scale. This is problematic due to the fundamental idea that measurement must be a latent variable on a unidimensional continuum. Although the evaluators of this character education grant conducted descriptive statistics and factor analyses, they were based on a classical test theory approach. This study sought to apply an item response theory approach, where the performance of an examinee on a test item can be predicted/explained by a latent variable (Hambleton, et al., 1991). This new approach presents a new way to empirically validate psychological attributes on measurement instruments, thereby advancing the field of psychological measurements. If evaluators continue to use the SMC scale as constructed, inaccurate inferences about students' moral and performance characters will continue to be made.

Another implication of this study is the notion that scale revisions can improve how the rating scale functions and add more information about the latent variables. The SMC scale is the primary scale that needs revision, given the multidimensionality results. After collapsing Categories 1 (Almost Never) and 2 (Rarely), results show that the orientation of items is all positive. Also, there are at least 10 observations in each category; observations range from 298 to 836. The distribution of observations, however, is quite large. Despite collapsing Category 1 and 2 (Almost Never and Rarely), they still represent the least counts (298 observations), while

Category 3 (Sometimes) has the highest number of observations (836). This irregularity in the frequency across categories indicates that more scale revisions are needed to improve rating scale functioning.

Additionally, the monotonicity of the observed average measures and rating scale category thresholds do increase with each rating scale category. In other words, as categories advance they represent more of the latent variable student moral character. Also, category fit shows that three out of five categories slightly overfit the model (Category 3 at .92, Category 4 at .92, and Category 5 at .95), while Categories 1 and 2 combined only marginally underfit the model at 1.20. Finally, the distance from the thresholds between the collapsed 1 and 2 category and Category 3 is between the recommended 1.0 and 5.0 logits, while the threshold distance between Category 3 and 4 is less than 1.0 (.94). This indicates that categories may be difficult to distinguish. A more in-depth, qualitative analysis can possibly explain why the sample did not use the rating scale as intended.

Finally, the person and item reliability for the scale revisions were the same as the original scales (.73, and .99, respectively). To improve person reliabilities, further scale revisions should include a person sample with a large ability range or should add a significant number of items to the scale that represent varying aspects of the cognitive, attitude, and action psychological realms. Finally, a multidimensional Rasch model (Adams, Wilson, & Wu, 1997; Adams, Wilson, & Wang, 1997) should be explored as another model to fit the data to, given the lack of unidimensionality for the SMC scale. Although these preliminary scale revisions show promising results, it is clear that more research is necessary to establish the unidimensionality of the SMC construct.

Recommendations

There are several ways to extend this empirical study. First, it is essential to vet performance and moral character items through a panel of character education experts in order to investigate assessment format, item clarity, and item importance. Additionally, this panel can create performance level descriptors to differentiate levels of performance and moral character. Next, the revised scales should be piloted to a larger sample of respondents; this sample should be representative and include varying types of schools, geographic regions, and student groups. Also, focus groups and one-on-one interviews with students would qualitatively highlight students with aberrant response patterns. Furthermore, residual analyses can be used to explore item fit and person fit. Finally, a multidimensional Rasch model should be applied if the Student Performance and/or Student Moral Character scales suggest multidimensionality.

Summary

According to Messick (1980), "Validation is a continuing, indeed, unending process that begins early in the test development process" (p. 1019). Anastasi (1986) agrees that validity should be built into the test from the outset, rather than being limited to the last stages of the test development. Given the importance of validating instruments from the onset of instrument development, the purpose of this study was to validate two existing scales that measure two types of character within students. Further research studies, however, are needed to validate scales and instruments that measure student character in order for statements to be made about the success of a character education program.

In conclusion, character matters, not just for its potential impact on academic achievement, but because it helps students be better people and live fulfilling lives. Further research directions include possible further analysis using a multidimensional IRT framework. Also, scale modifications can be made using Wilson’s (2005) constructing measures framework.

This framework allows stakeholders, in conjunction with evaluators, to be intimately involved in the scale development process. Although it may take an iterative process to reach consensus, it provides a more valid approach to operationalizing a construct. Finally, more research can be conducted on the theoretical underpinnings of these scales. Modern psychological theories need new and theoretically valid methods of measurement; the evolution of better theories, however, depends on the construction of better research instruments (Lind, 2008). This means that the advancement of research will be hindered without validated measures:

> In the end, identifying promising practices is as much art as science. Ultimately, it is an effort to make judgments that are informed by what research tells us. [Hopefully] identification of these promising practices will be followed by further research on their effectiveness. (Lickona & Davidson, 2005, p. xxiii)

In order for promising character education practices to be uncovered and its effectiveness to be accurately captured, scales/instruments that measure character must undergo a stricter investigation and provide substantial validity evidence.

Table 1

Gilligan's Theory of Moral Development

Ethic of Survival	Stage 1: Caring for Self
Ethic of Conventional Goodness	Stage 2: Caring for others
Ethic of Care	Stage 3: Caring for both self and other

Table 2

Kohlberg's Theory of Moral Development

Level One: Preconventional Morality	Stage 1: Punishment-Obedience Orientation	Obedience and punishment driven
	Stage 2: Instrumental Relativist Orientation	Self-interest driven
Level Two: Conventional Morality	Stage 3: Good Boy-Nice Girl Orientation	Driven by desire to conform to societal expectations
	Stage 4: Law and Order Orientation	Driven by desire to obey laws and social conventions
Level Three: PostConventional Morality	Stage 5: Social Contract Orientation	Driven by democratic notion that laws are not hard and fast rules
	Stage 6: Universal Ethical Principle Orientation	Driven by abstract reasoning using ethical principles (e.g. civil disobedience to violate law)

Table 3

Example of Kohlberg's Moral Judgment Manual Item

Stage & Substage: 2A
Criterion Judgment The doctor should not give the woman the drug because he would risk losing his job or going to jail.
Stage Structure Not killing the woman is justified because it involves a risk (rather than certainty) of punishment. Punishment is seen as something to be instrumentally avoided. The risk of punishment overrides the recognition of the pragmatic reasonableness from the woman's point of view of giving her the drug.
Critical Indicators One of the following must be used as the central justification for not killing the woman: (a) punishment as possible or probable, a risk to be weighed in the decision; OR (b) other disadvantageous consequences to the doctor (he might lose his job, etc.)

Note. Adapted from Kohlberg (1984).

Table 4

Psychological Realms of Moral and Performance Character

	Awareness (cognitive)	Attitude (emotional/valuing)	Action (behavior/habits)
MORAL	Students - can recognize the ethical dimensions of situations - can grapple with important moral questions	Students - care deeply about doing the right thing - have the courage of conscience in the face of social pressure	Students - act upon ethical convictions - take a stand for what is right
PERFORMANCE	Students - identify excellence in many areas of endeavor - understand the performance virtues required to pursue excellence	Students - are strongly motivated to give best effort - are committed to high-quality work	Students - demonstrate the skills/habits required for excellence - practice in order to improve in the skills and habits required for excellence

Table 5

Student Performance Character and Student Moral Character Items

Performance Character	Moral Character
A1. I can be counted on to do my part for the team/group (CountedOn)	A13. I treat teachers and staff with respect, even if I disagree with them (AdultRespect)
A2. I try to get out of doing things that are difficult or boring (reversed; AvoidBoring)	A14. I break classroom or school rules (reversed; BrkRules)
A3. I spend extra time working to improve my weaknesses (TimetoImprove)	A15. When I see someone having a problem, I offer to help (HelpWithProb)
A4. I continue trying hard, even when things are not going well (TryHard)	A16. (DoRight)
A5. I forget to bring what is needed for class (reversed; 4getMaterials)	A17. I help another student choose between doing what is right and what is wrong (HelpOthersRgt)
A6. I work with another student to help him or her do better on an assignment, without letting them copy my work (HelpOthersWk)	A18. I make fun of someone (reversed; Tease)
A7. I forget to do my homework (reversed; 4getHW)	A19. I speak up when someone is bullied (CallOutBully)
A8. I think about my school work and consider whether I need to work harder (Reflect)	A20. I cheat on a test or an assignment (reversed; Cheat)
A9. I talk to a teacher to find out if I'm doing	A21. I think about how my parent, teacher, or

well in my school work (AskProgress)	coach would act before making an important decision (ThkBoutAdults)
A10. I run out of time to do my assignments well (reversed; RunOutTime)	A22. I admit if I do something wrong (AdmitWrong)
A11. I give up watching TV or hanging out with friends to study for a test or do an assignment for school (GiveUp4Sch)	A23. I consider different points of view when making a decision about a moral issue or dilemma (DiffPOVs)
A12. I am willing to redo a school assignment to make it better (RedoWk)	

Table 6

Student Performance Character Items by Psychological Realm

COGNITIVE (AWARENESS)	ATTITUDE (EMOTIONAL/VALUING)	ACTION (BEHAVIOR/HABITS)
Reflect A8. I think about my school work and consider whether I need to work harder (Reflect)	TryHard A4. I continue trying hard, even when things are not going well (TryHard)	CountedOn A1. I can be counted on to do my part for the team/group (CountedOn)
	RedoWk A12. I am willing to redo a school assignment to make it better (RedoWk)	AvoidBoring A2. I try to get out of doing things that are difficult or boring (reversed; AvoidBoring)
		TimetoImprove A3. I spend extra time working to improve my weaknesses (TimetoImprove)
		4getMaterials A5. I forget to bring what is needed for class (reversed; 4getMaterials)
		4getHW

		A7. I forget to do my homework (reversed; 4getHW)
		AskProgress A9. I talk to a teacher to find out if I'm doing well in my school work. (AskProgress)
		RunOutTime A10. I run out of time to do my assignments well (reversed; RunOutTime)
		GiveUp4Sch A11. I give up watching TV or hanging out with friends to study for a test or do an assignment for school (GiveUp4Sch)
		HelpOthersWk A16. I work with another student to help him or her do better on an assignment, without letting them copy my work (HelpOthersWk)

Table 7

Student Moral Character Items by Psychological Realm

COGNITIVE (AWARENESS)	ATTITUDE (EMOTIONAL/VALUING)	ACTION (BEHAVIOR/HABITS)
ThkBoutAdults A21. I think about how my parent, teacher, or coach would act before making an important decision (ThkBoutAdults)	HelpWithProb A15. When I see someone having a problem, I offer to help (HelpWithProb)	AdultRespect A13. I treat teachers and staff with respect, even if I disagree with them (AdultRespect)
DiffPOVs A23. I consider different points of view when making a decision about a moral issue or dilemma (DiffPOVs)	AdmitWrong A22. I admit if I do something wrong (AdmitWrong)	BrkRules A14. I break classroom or school rules (reversed; BrkRules)
		DoRight A16. I do the right thing no matter what others might think (DoRight)
		HelpOthersRgt A17. I help another student choose

		between doing what is right and what is wrong (HelpOthersRgt)
		Tease A18. I make fun of someone (reversed; Tease)
		CallOutBully A19. I speak up when someone is bullied (CallOutBully)
		Cheat A20. I cheat on a test or an assignment (reversed; Cheat)

Table 8

Summary Statistics of Study Sample

Characteristic	Spring 2009	Percentage
Total	239	
Gender		
Male	92	39%
Female	134	56%
Unknown	13	5%
Ethnicity		
Black	180	75%
Hispanic/Latino/a	40	17%
Native American	1	< 1%
White	1	< 1%
Other/Unknown	17	7%
Grade		
Grade 6	163	68%
Grade 7	56	23%
Grade 8	9	4%
Missing	11	5%
School		
School A	77	32%

School B	53	22%
School C	57	24%
School D	52	22%

Table 9

Means for Gender Groups on SPC and SMC

Gender	N	Mean	Std Dev
Student Performance Character			
Male	92	3.38*	.47
Female	132	3.51*	.48
Student Moral Character			
Male	92	3.57	.52
Female	133	3.66	.50

Note. * indicates $p < .05$

Table 10

Means for Moral Character and Grades

Grade	N	Mean	Std Dev
6	158	3.58	.49
7	56	3.75	.49
8	9	3.46	.75

Note. * indicates $p < .05$

Table 11

Means for Performance Character and Grades

Grade	N	Mean	Std Dev
6	158	3.42	.48
7	56	3.57	.40
8	8	3.38	.82

Note. * indicates $p < .05$

Table 12

Student Performance Character Item Fit Statistics

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
2	696	224	.46	.07	1.36	3.6	1.37	3.7	A .36	.50	33.9	43.6	AvoidBoring
10	762	222	.05	.08	1.27	2.7	1.30	3.0	B .32	.48	45.9	43.3	RunOutTime
7	836	223	-.39	.08	1.16	1.7	1.26	2.7	C .48	.46	43.5	42.7	4getHW
5	794	223	-.12	.08	1.24	2.5	1.23	2.4	D .52	.47	40.8	42.3	4GetMaterials
9	731	222	.22	.08	1.11	1.2	1.13	1.4	E .41	.49	45.0	43.4	AskProgress
11	645	224	.74	.07	1.12	1.3	1.11	1.3	F .47	.50	46.9	41.9	GiveUp4Sch
6	756	224	.12	.08	.83	-1.9	.84	-1.9	f .58	.49	51.8	43.3	HelpOthersWk
3	735	224	.24	.08	.82	-2.1	.83	-2.0	e .46	.49	41.1	43.8	TimetoImprove
8	761	222	.05	.08	.81	-2.2	.82	-2.1	d .53	.48	54.1	43.3	Reflect
12	807	223	-.20	.08	.79	-2.4	.78	-2.6	c .55	.47	48.9	42.8	RedoWk
1	913	225	-.85	.08	.79	-2.5	.77	-2.6	b .49	.43	49.3	43.7	CountedOn
4	824	223	-.31	.08	.67	-4.1	.68	-3.9	a .63	.46	52.5	42.7	TryHard
MEAN	771.7	223.3	.00	.08	1.00	-.2	1.01	-.1			46.1	43.1	
S.D.	67.0	.9	.40	.00	.22	2.5	.24	2.6			5.5	.6	

Note. OBS% represents the percent of data points which are within .05 score points of their expected values. EXP% is the percent of data points that are predicted to be within .05 score points of their expected values.

Table 13

Student Performance Character Category Statistics

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	STRUCTURE CALIBRATN	CATEGORY MEASURE	
1	1	101	4	-.19	-.32	1.17	1.28	NONE	(-2.75)	1
2	2	328	12	.01	.01	.99	.99	-1.33	-1.25	2
3	3	967	36	.32	.34	.87	.85	-.91	-.05	3
4	4	813	30	.73	.74	.99	.98	.71	1.23	4
5	5	470	18	1.29	1.25	.99	.99	1.53	(2.87)	5
MISSING		21	1	.24						

Table 14

Student Moral Character Item Fit Statistics

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MODEL MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
8	1028	223	-1.86	.12	1.44	3.4	1.18	1.2	A .35	.35	64.0	65.1	Cheat
7	666	222	.94	.08	1.31	3.2	1.36	3.6	B .36	.57	45.7	42.9	CallOutBully
1	898	223	-.58	.09	1.27	2.8	1.21	2.1	C .49	.48	41.9	45.8	AdultRespect
2	916	223	-.73	.09	1.13	1.4	1.13	1.2	D .48	.47	50.0	46.8	BrkRules
5	730	225	.61	.08	1.03	.4	1.03	.4	E .53	.55	50.0	46.4	HelpOthersRgt
9	738	224	.53	.08	1.01	.2	1.03	.3	F .56	.55	47.5	46.8	ThkBoutAdults
6	843	225	-.13	.08	.98	-.1	.97	-.3	e .57	.51	45.1	45.4	Tease
3	724	222	.59	.08	.82	-2.0	.81	-2.1	d .61	.55	57.5	46.4	HelpWithProb
10	789	224	.21	.08	.79	-2.3	.78	-2.4	c .54	.53	52.9	46.0	AdmitWrong
11	757	222	.36	.08	.78	-2.4	.78	-2.4	b .59	.54	50.7	46.3	DiffPOVs
4	814	224	.04	.08	.75	-2.8	.75	-2.8	a .59	.52	55.2	46.2	DoRight
MEAN	809.4	223.4	.00	.09	1.03	.1	1.00	-.1			50.9	47.7	
S.D.	100.3	1.1	.76	.01	.22	2.2	.19	2.0			6.0	5.6	

Note. OBS% represents the percent of data points which are within .05 score points of their expected values. EXP% is the percent of data points that are predicted to be within .05 score points of their expected values.

Table 15

Student Moral Character Category Statistics

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	STRUCTURE CALIBRATN	CATEGORY MEASURE	
1	1	88	4	-.34	-.50	1.15	1.19	NONE	(-2.73)	1
2	2	210	9	-.07	-.12	1.11	1.19	-1.18	-1.34	2
3	3	836	34	.31	.35	.90	.92	-1.27	-.12	3
4	4	728	30	.96	.99	.94	.86	.79	1.30	4
5	5	595	24	2.00	1.95	.98	.98	1.66	(2.99)	5
MISSING		18	1	.59						

Table 16

Person Reliabilities

Rating Scale Model		Partial Credit Model	
SPC (*N*=225)	SMC (*N*=224)	SPC (*N*=225)	SMC (*N*=224)
.66	.73	.66	.74

Table 17

Item Reliabilities

Rating Scale Model		Partial Credit Model	
SPC (*N*=12)	SMC (*N*=11)	SPC (*N*=12)	SMC (*N*=226)
.96	.99	.96	.97

Table 18

Student Performance Character Item Fit Values (RSM)

Item	Measure	Measure SE	Outfit Mean Square
CountedOn	-.85	.08	.77
AvoidBoring	.46	.07	1.37
TimetoImprove	.24	.08	.83
TryHard	-.31	.08	.68
4getMaterials	-.12	.08	1.23
HelpOthersWk	.12	.08	.84
4getHW	-.39	.08	1.26
Reflect	.05	.08	.82
AskProgress	.22	.08	1.13
RunOutTime	.05	.08	1.30
GiveUp4Sch	.74	.07	1.11
RedoWk	-.20	.08	.78

Note. Range of acceptability .6 < MSE < 1.4.

Table 19

Student Moral Character Item Fit Values (RSM)

Item	Measure	Measure SE	Outfit Mean Square
AdultRespect	-.58	.09	1.21
BrkRules	-.73	.09	1.13
HelpWithProb	.59	.08	.81
DoRight	.04	.08	.75
HelpOthersRgt	.61	.08	1.03
Tease	-.13	.08	.97
CallOutBully	.94	.08	1.36
Cheat	-1.86	.12	1.18
ThkBoutAdults	.53	.08	1.03
AdmitWrong	.21	.08	.78
DiffPOVs	.36	.08	.78

Note. Range of acceptability .6 < MSE < 1.4.

Table 20

Student Performance Character Item Fit Values (PCM)

Item	Measure	Measure SE	Outfit Mean Square
CountedOn	-.83	.09	.90
AvoidBoring	.46	.07	1.26
TimetoImprove	.22	.08	1.00
TryHard	-.42	.08	.79
4getMaterials	.00	.07	.98
HelpOthersWk	.16	.08	.87
4getHW	-.33	.07	1.15
Reflect	.07	.08	.93
AskProgress	.19	.08	1.13
RunOutTime	.21	.08	1.29
GiveUp4Sch	.64	.07	1.08
RedoWk	-.36	.08	.88

Note. Range of acceptability .6 < MSE < 1.4.

Table 21

Student Moral Character Item Fit Values (PCM)

Item	Measure	Measure SE	Outfit Mean Square
AdultRespect	-.37	.08	1.00
BrkRules	-.46	.08	1.05
HelpWithProb	.58	.08	.89
DoRight	-.42	.09	.87
HelpOthersRgt	.53	.08	1.04
Tease	-.16	.08	.91
CallOutBully	.62	.08	1.46
Cheat	-1.15	.10	.93
ThkBoutAdults	.55	.08	.99
AdmitWrong	.04	.09	.95
DiffPOVs	.24	.09	.90

Note. Range of acceptability .6 < MSE < 1.4.

Table 22

Student Performance Character Residual Variance (in Eigenvalue units)

	Empirical	Modeled
Total raw variance in observations	17.0	100%
Raw variance explained by measures	5.0	29.5%
Raw variance explained by persons	1.4	8.4%
Raw variance explained by items	3.6	21.1%
Raw unexplained variance (total)	12.0	70.5%
Unexplained variance in 1st contrast	2.2	13.1%
Unexplained variance in 2nd contrast	1.5	8.7%

Table 23

Student Moral Character Residual Variance (in Eigenvalue units)

	Empirical	Modeled
Total raw variance in observations	18.7	100%
Raw variance explained by measures	7.7	41.2%
Raw variance explained by persons	2.6	14.0%
Raw variance explained by items	5.1	27.3%
Raw unexplained variance (total)	11.0	58.8%
Unexplained variance in 1st contrast	2.1	11.5%
Unexplained variance in 2nd contrast	1.6	8.4%

Table 24

Student Performance Character DIF Analyses

PERSON	DIF	DIF	PERSON	DIF	DIF	DIF	JOINT		Welch		MantelHanzl	ITEM	
CLASS	MEASURE	S.E.	CLASS	MEASURE	S.E.	CONTRAST	S.E.	t	d.f.	Prob.	Prob.	Size	Number	Name
1	-.77	.13	2	-.91	.11	.14	.17	.83	209	.4095	.1761	.25	1	CountedOn
1	.34	.12	2	.54	.10	-.20	.15	-1.32	206	.1875	.8983	.14	2	AvoidBoring
1	.35	.12	2	.15	.10	.20	.15	1.33	207	.1848	.0636	.67	3	TimetoImprove
1	-.31	.12	2	-.31	.10	.00	.16	.00	207	1.000	.8163	-.21	4	TryHard
1	-.09	.12	2	-.15	.10	.06	.16	.39	207	.6994	.9734	.22	5	4GetMaterials
1	.29	.12	2	.00	.10	.29	.15	1.87	207	.0628	.1170	.44	6	HelpOthersWk
1	-.47	.12	2	-.33	.10	-.14	.16	-.85	206	.3963	.3424	.14	7	4getHW
1	-.05	.12	2	.12	.10	-.17	.16	-1.10	204	.2732	.1894	-.08	8	Reflect
1	.25	.12	2	.20	.10	.05	.15	.32	204	.7506	.8861	-.10	9	AskProgress
1	-.12	.12	2	.16	.10	-.28	.16	-1.79	205	.0749	.2166	-.06	10	RunOutTime
1	.74	.12	2	.74	.10	.00	.15	.00	206	1.000	.9114	-.08	11	GiveUp4Sch
1	-.15	.12	2	-.24	.10	.09	.16	.57	205	.5694	.7552	-.45	12	RedoWk
2	-.91	.11	1	-.77	.13	-.14	.17	-.83	209	.4095	.1761	-.25	1	CountedOn
2	.54	.10	1	.34	.12	.20	.15	1.32	206	.1875	.8983	-.14	2	AvoidBoring
2	.15	.10	1	.35	.12	-.20	.15	-1.33	207	.1848	.0636	-.67	3	TimetoImprove
2	-.31	.10	1	-.31	.12	.00	.16	.00	207	1.000	.8163	.21	4	TryHard
2	-.15	.10	1	-.09	.12	-.06	.16	-.39	207	.6994	.9734	-.22	5	4GetMaterials
2	.00	.10	1	.29	.12	-.29	.15	-1.87	207	.0628	.1170	-.44	6	HelpOthersWk
2	-.33	.10	1	-.47	.12	.14	.16	.85	206	.3963	.3424	-.14	7	4getHW
2	.12	.10	1	-.05	.12	.17	.16	1.10	204	.2732	.1894	.08	8	Reflect
2	.20	.10	1	.25	.12	-.05	.15	-.32	204	.7506	.8861	.10	9	AskProgress
2	.16	.10	1	-.12	.12	.28	.16	1.79	205	.0749	.2166	.06	10	RunOutTime
2	.74	.10	1	.74	.12	.00	.15	.00	206	1.000	.9114	.08	11	GiveUp4Sch
2	-.24	.10	1	-.15	.12	-.09	.16	-.57	205	.5694	.7552	.45	12	RedoWk

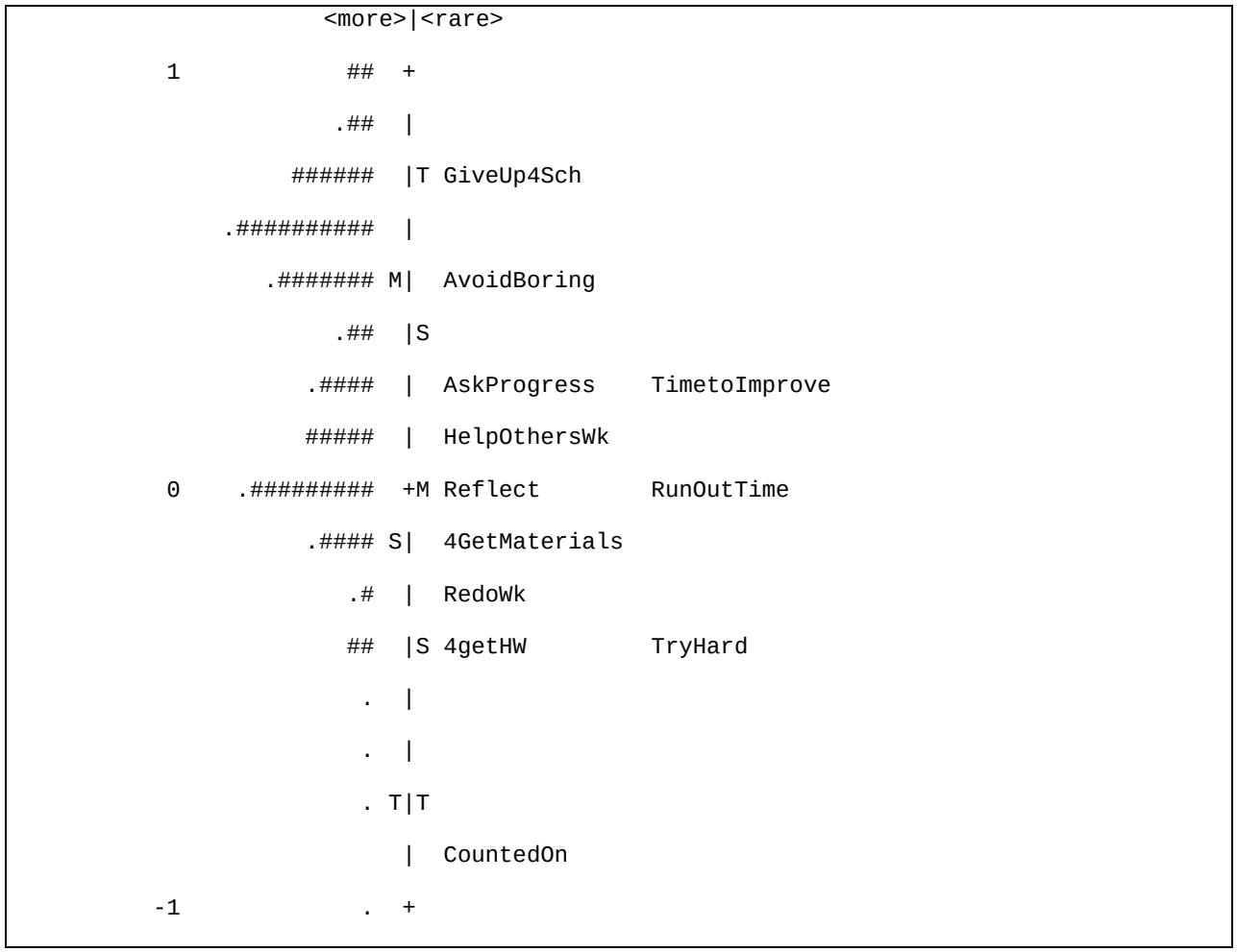
Table 25

Student Moral Character DIF Analyses

PERSON CLASS	DIF MEASURE	DIF S.E.	PERSON CLASS	DIF MEASURE	DIF S.E.	DIF CONTRAST	JOINT S.E.	Welch t	Welch d.f.	Welch Prob.	MantelHanzl Prob.	ITEM Size	ITEM Number	ITEM Name
1	-.58	.14	2	-.58	.12	.00	.18	.00	205	1.000	.4265	-.28	1	AdultRespect
1	-.59	.14	2	-.84	.12	.25	.18	1.38	207	.1687	.0769	.42	2	BrkRules
1	.59	.12	2	.59	.10	.00	.16	.00	203	1.000	.5822	-.33	3	HelpWithProb
1	.13	.13	2	-.02	.11	.16	.17	.95	204	.3453	.0363	.62	4	DoRight
1	.57	.12	2	.63	.10	-.05	.16	-.32	205	.7480	.4455	.33	5	HelpOthersRgt
1	-.17	.13	2	-.11	.11	-.06	.17	-.36	205	.7159	.8587	.21	6	Tease
1	1.03	.12	2	.88	.10	.15	.16	.97	203	.3351	.2087	.02	7	CallOutBully
1	-1.98	.18	2	-1.78	.15	-.20	.24	-.82	203	.4132	.7936	-.04	8	Cheat
1	.33	.12	2	.67	.10	-.34	.16	-2.10	204	.0367	.0149	.01	9	ThkBoutAdults
1	.30	.12	2	.15	.11	.15	.16	.91	206	.3648	.4987	.23	10	AdmitWrong
1	.30	.12	2	.41	.10	-.11	.16	-.65	204	.5182	.1691	-.19	11	DiffPOVs
2	-.58	.12	1	-.58	.14	.00	.18	.00	205	1.000	.4265	.28	1	AdultRespect
2	-.84	.12	1	-.59	.14	-.25	.18	-1.38	207	.1687	.0769	-.42	2	BrkRules
2	.59	.10	1	.59	.12	.00	.16	.00	203	1.000	.5822	.33	3	HelpWithProb
2	-.02	.11	1	.13	.13	-.16	.17	-.95	204	.3453	.0363	-.62	4	DoRight
2	.63	.10	1	.57	.12	.05	.16	.32	205	.7480	.4455	-.33	5	HelpOthersRgt
2	-.11	.11	1	-.17	.13	.06	.17	.36	205	.7159	.8587	-.21	6	Tease
2	.88	.10	1	1.03	.12	-.15	.16	-.97	203	.3351	.2087	-.02	7	CallOutBully
2	-1.78	.15	1	-1.98	.18	.20	.24	.82	203	.4132	.7936	.04	8	Cheat
2	.67	.10	1	.33	.12	.34	.16	2.10	204	.0367	.0149	-.01	9	ThkBoutAdults
2	.15	.11	1	.30	.12	-.15	.16	-.91	206	.3648	.4987	-.23	10	AdmitWrong
2	.41	.10	1	.30	.12	.11	.16	.65	204	.5182	.1691	.19	11	DiffPOVs

Figure 1

Student Performance Character Wright Map (RSM)

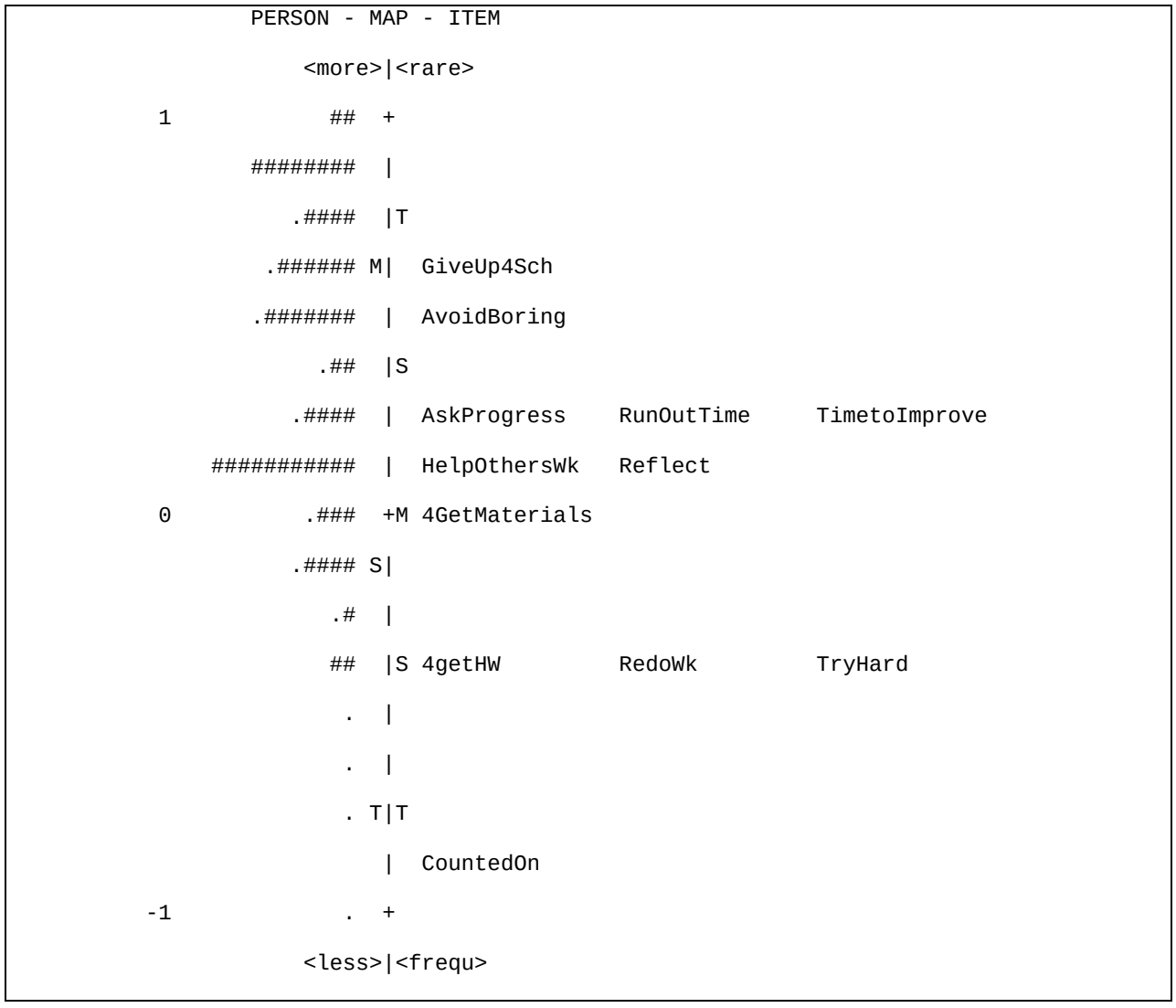


Note. Each "#" is 3.

Note. Each "." is 1 to 2.

Figure 2

Student Performance Character Wright Map (PCM)

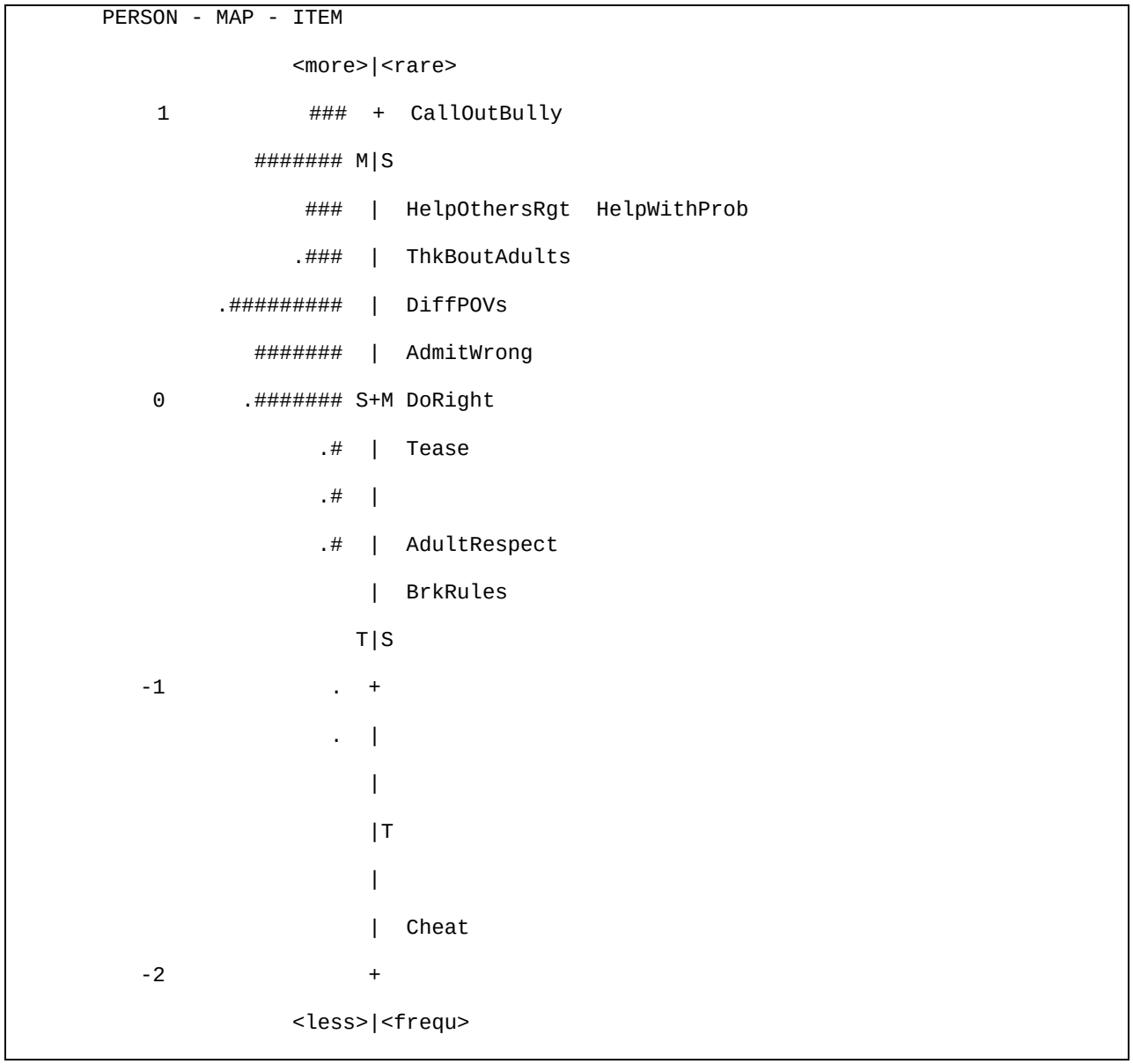


Note. Each “#” is 3.

Note. Each “.” is 1 to 2.

Figure 3

Student Moral Character Wright Map (RSM)

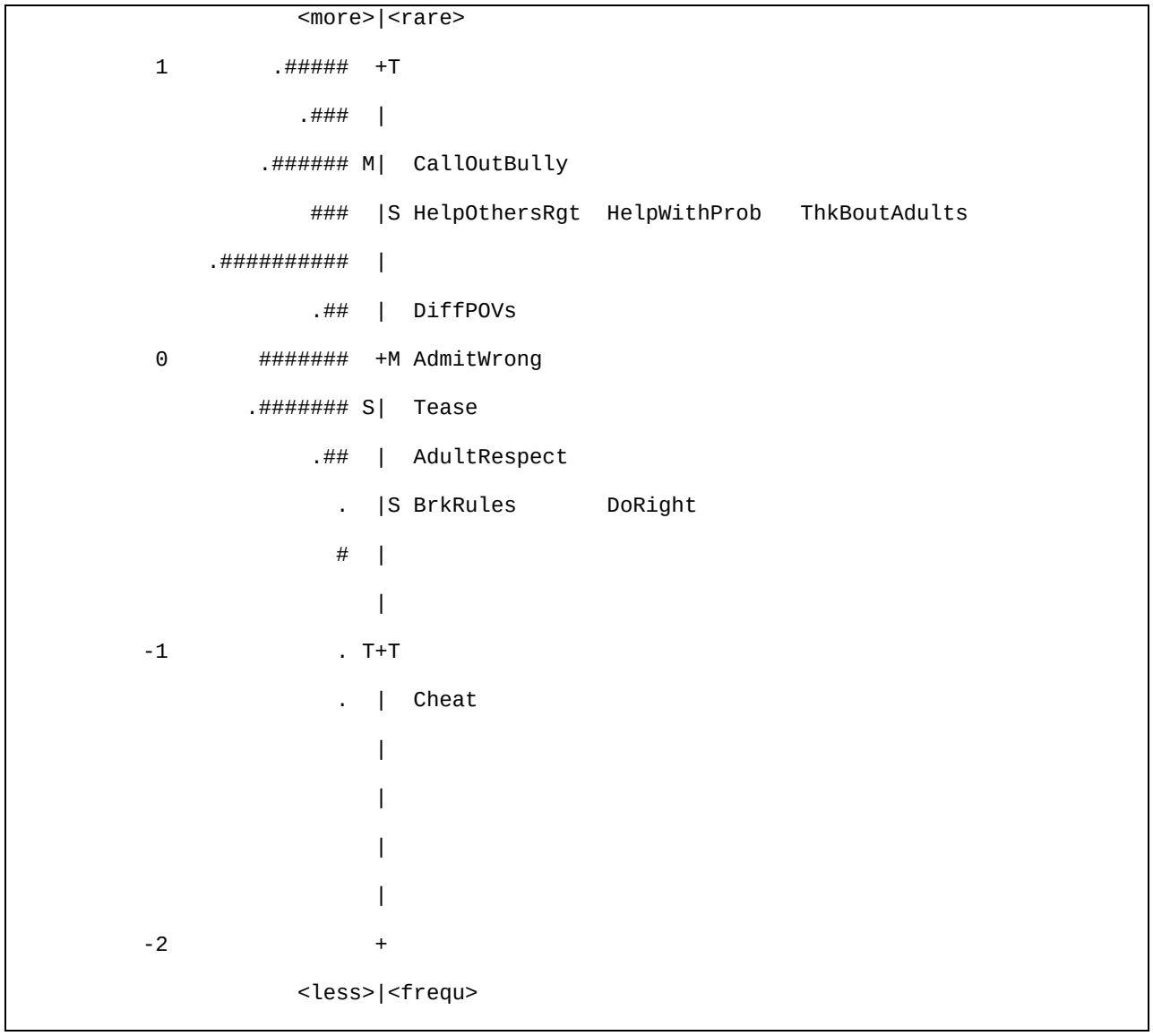


Note. Each "#" is 3.

Note. Each "." is 1 to 2.

Figure 4

Student Moral Character Wright Map (PCM)

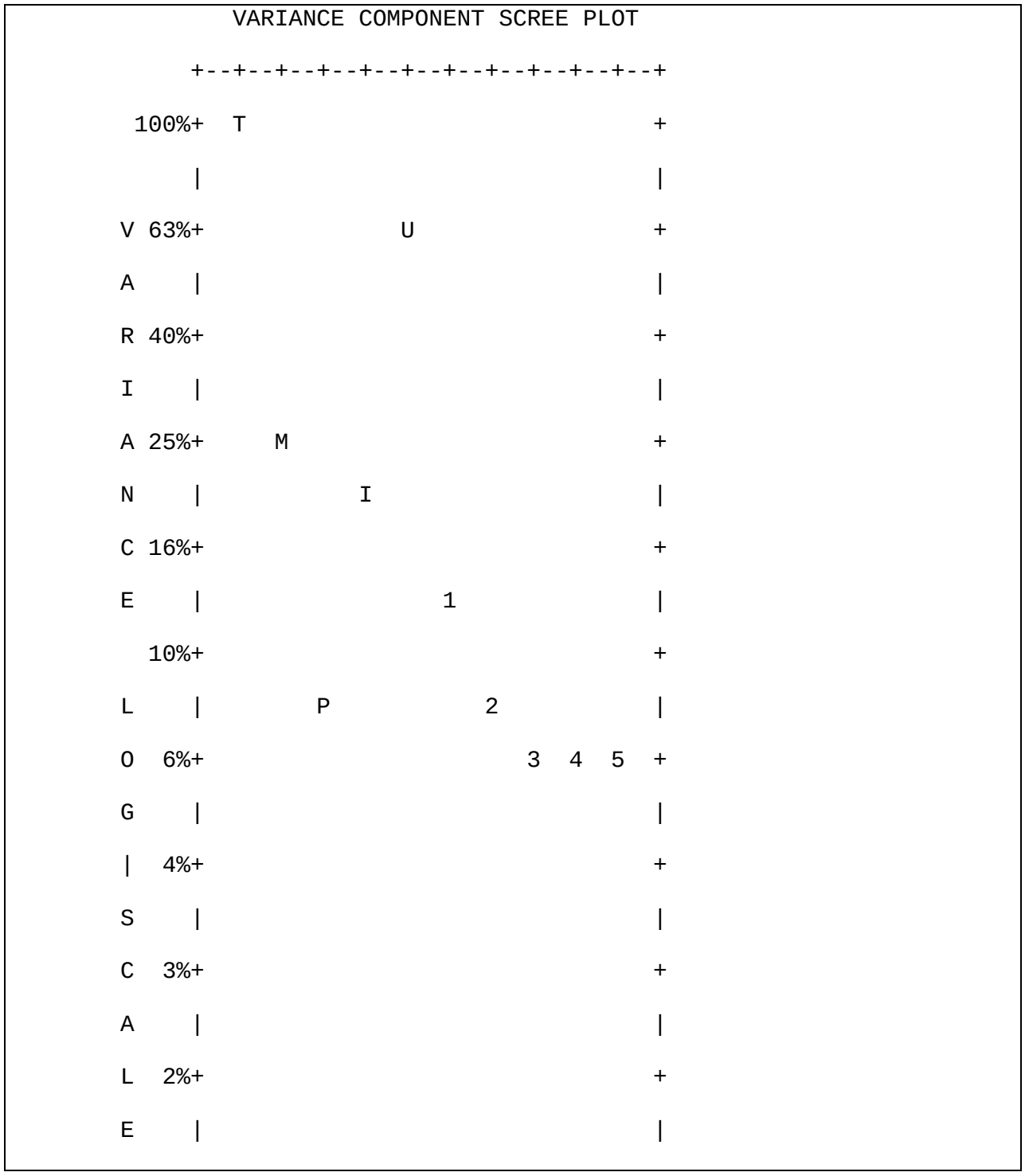


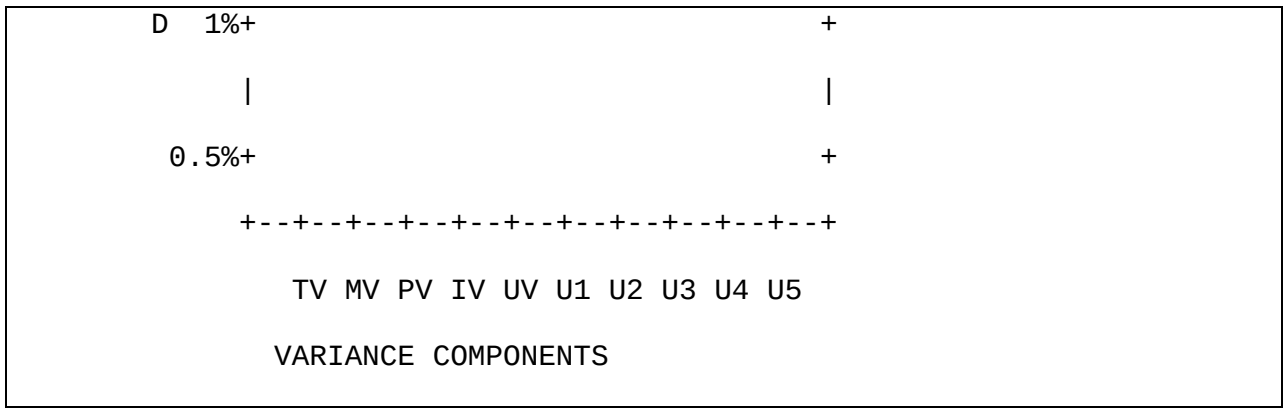
Note. Each "#" is 3.

Note. Each "." is 1 to 2.

Figure 5

Student Performance Character Residual Variance Scree Plot

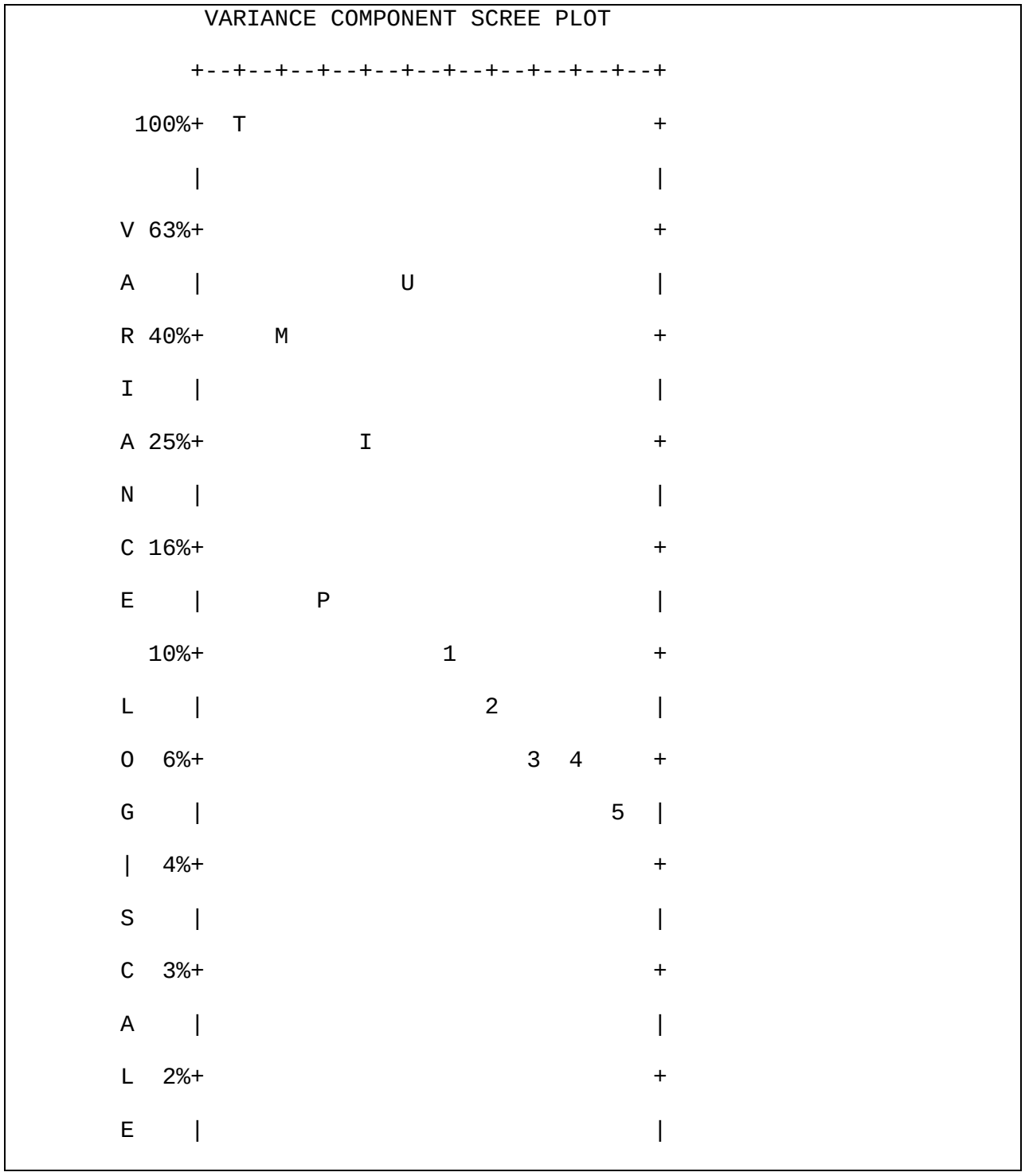


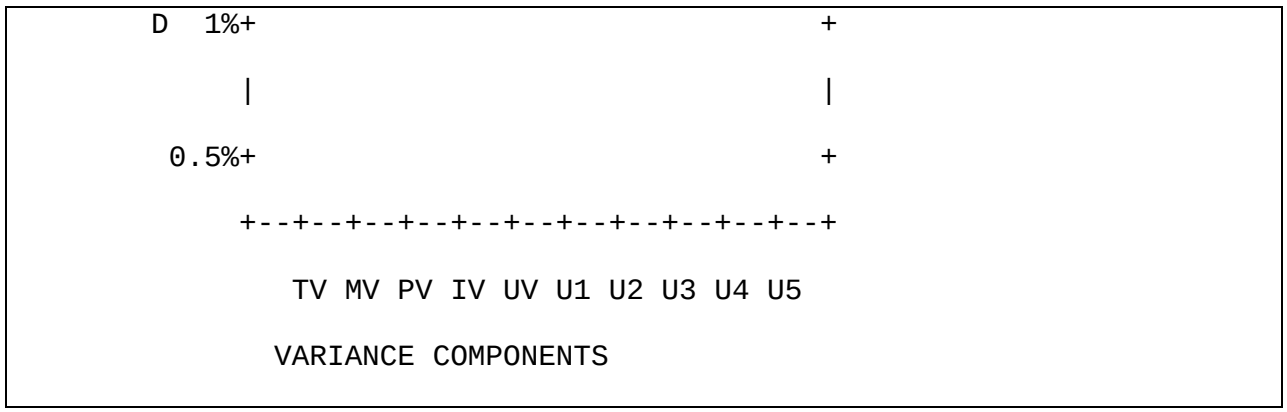


Note. *T* stands for Total Raw Variance, *U* stands for Unexplained Variance, *M* stands for Measure, *P* stands for Person, and *I* stands for Item.

Figure 6

Student Moral Character Residual Variance Scree Plot

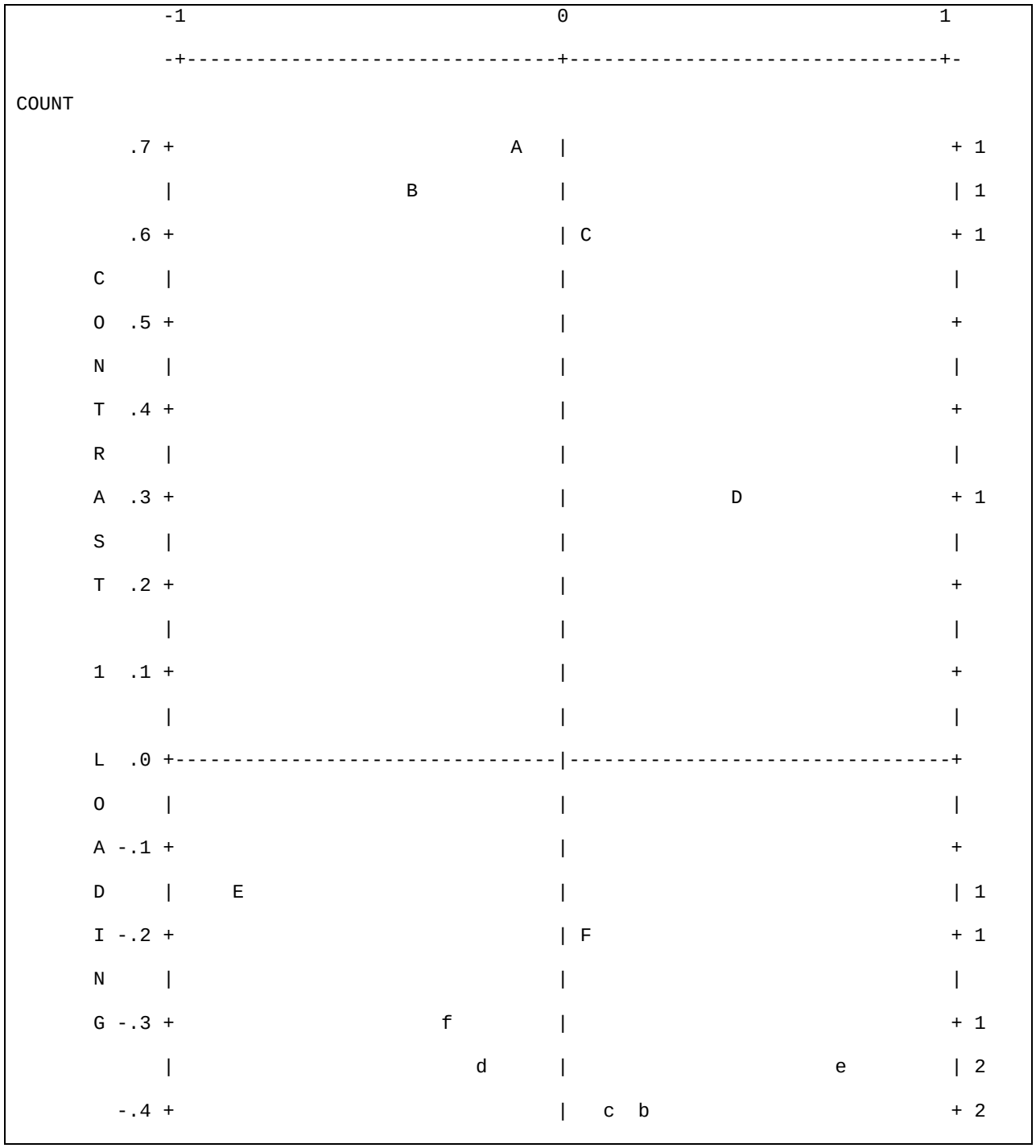




Note. *T* stands for Total Raw Variance, *U* stands for Unexplained Variance, *M* stands for Measure, *P* stands for Person, and *I* stands for Item.

Figure 7

Student Performance Character Plot of Standardized Residual PCA (Contrast1)



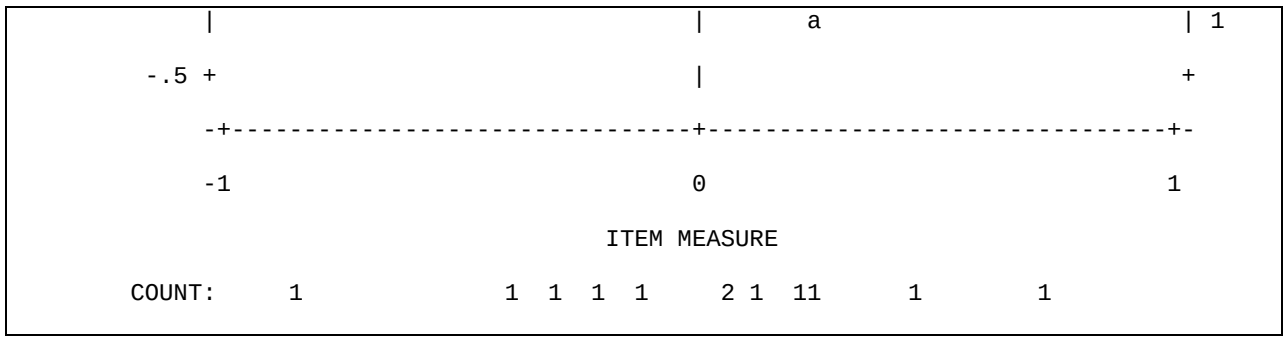
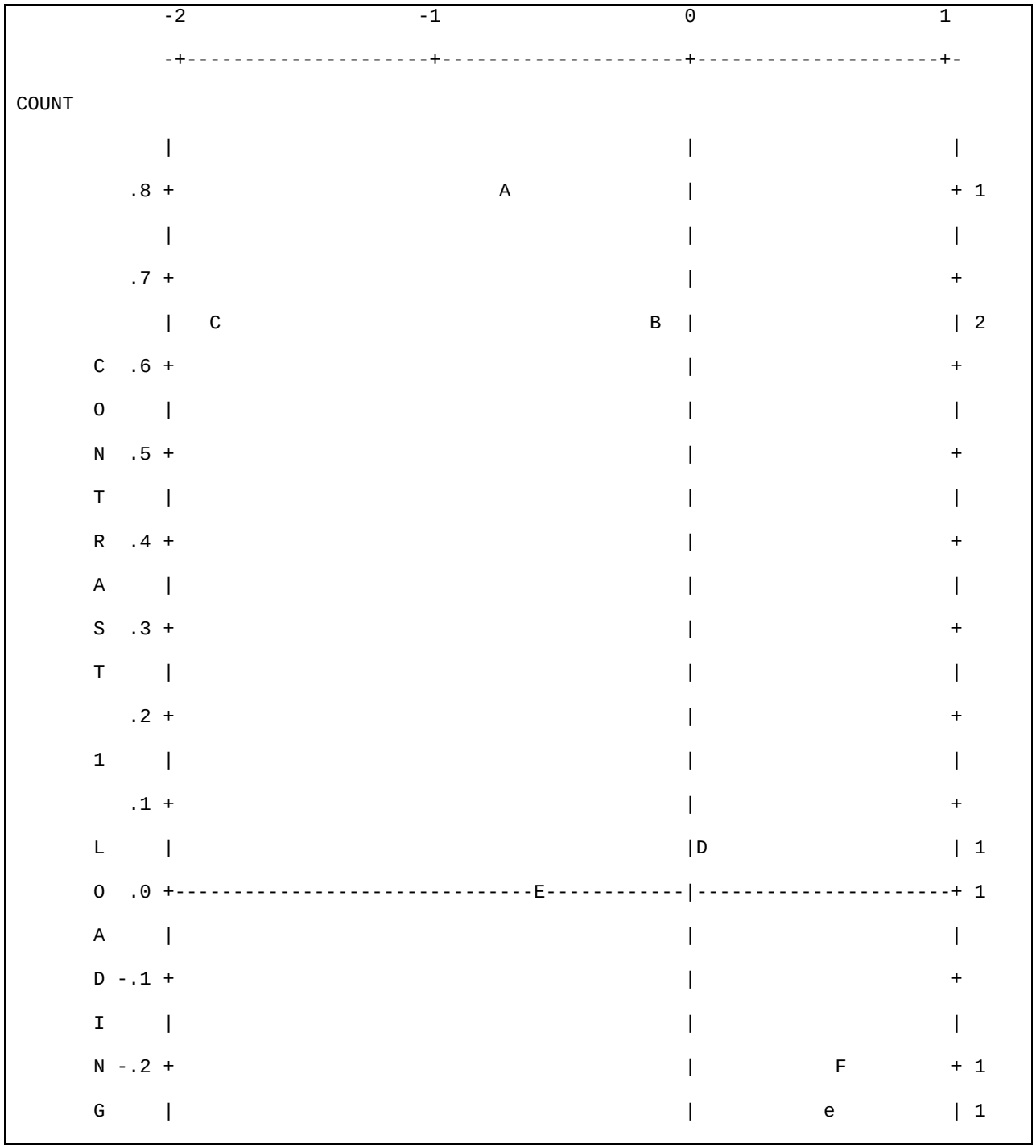

|
|
a
| 1
-.5 +
|
+
-+----------+----------+
-1
0
1
ITEM MEASURE
COUNT: 1 1 1 1 1 2 1 11 1 1

Figure 8

Student Moral Character Plot of Standardized Residual PCA (Contrast 1)



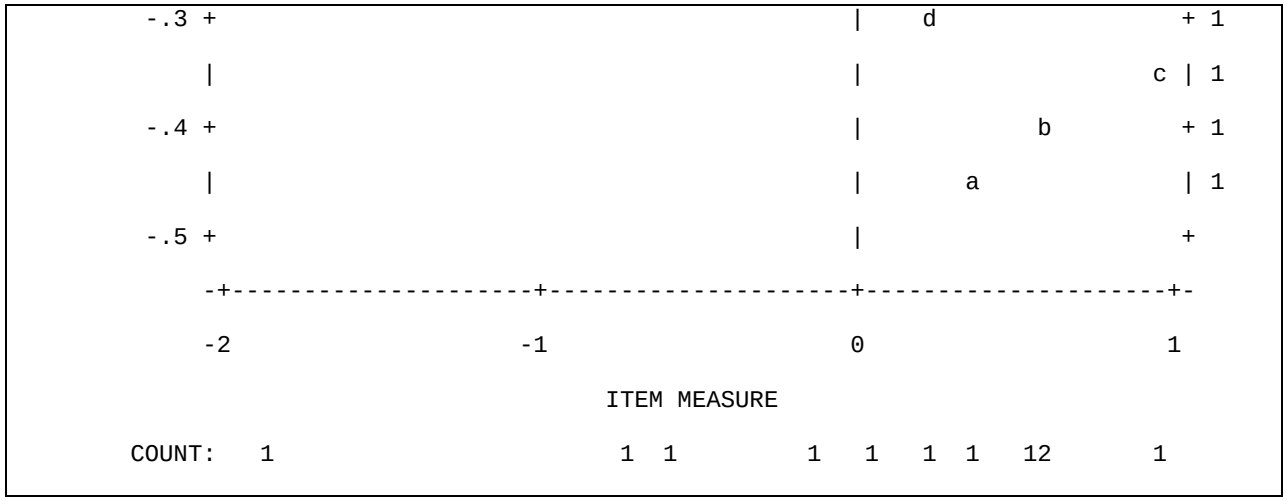
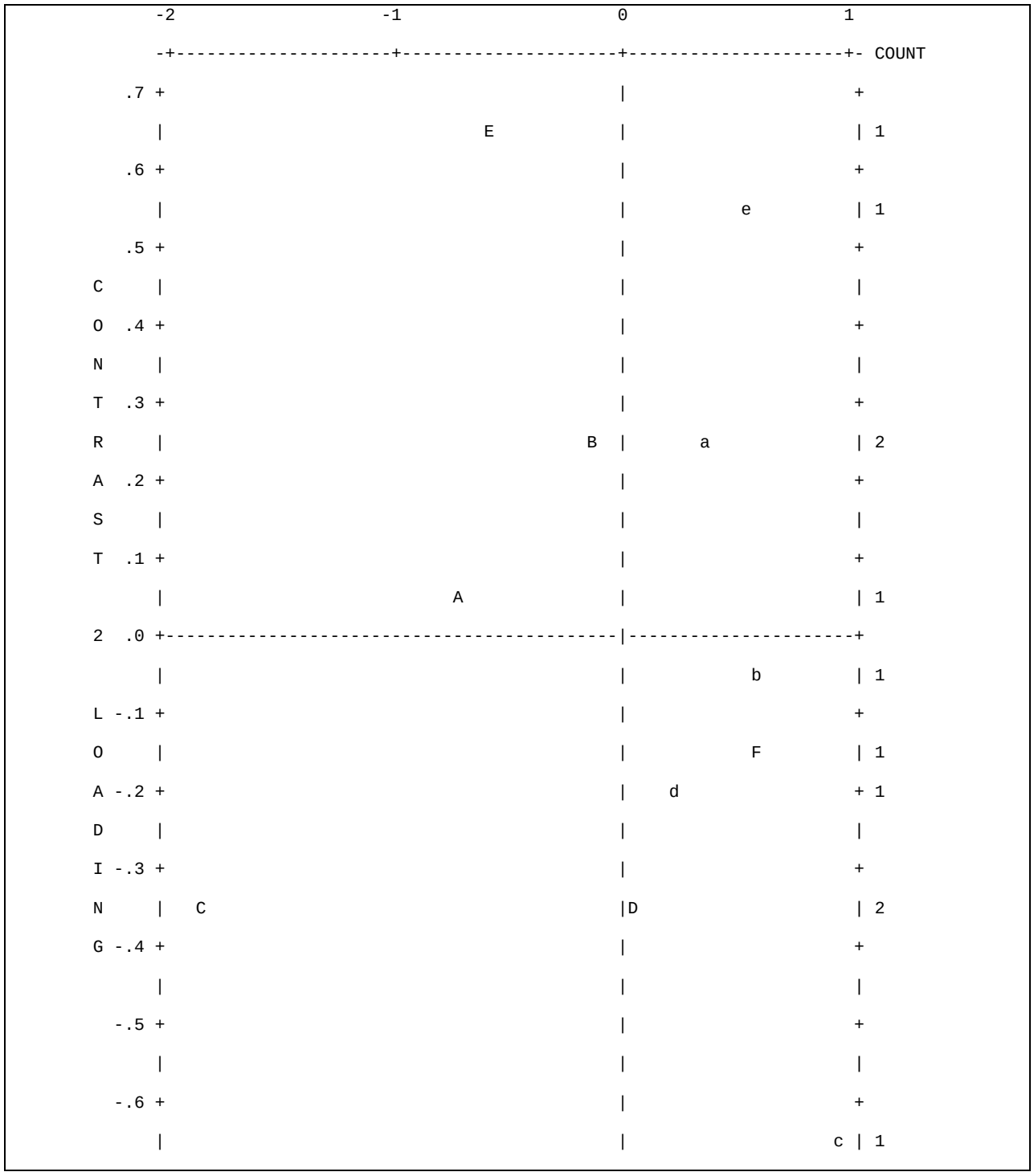
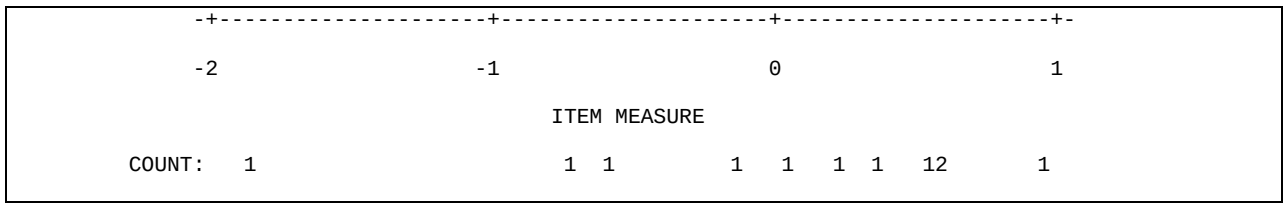

-.3 +
|
-.4 +
|
-.5 +
-2
-1
0
1
d
c
b
a
ITEM MEASURE
COUNT: 1 1 1 1 1 1 12 1

Figure 9

Student Moral Character Plot of Standardized Residual PCA (Contrast 2)




-+----------+----------+----------+-
-2 -1 0 1
ITEM MEASURE
COUNT: 1 1 1 1 1 1 1 12 1

References

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational & Behavioral Statistics, 22*, 47-76.

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

Alexander, K. L., Entwisle, D. R., & Kabbani, N. S. (2001). The dropout process in life course perspective: Early risk factors at home and school. *Teachers College Record, 103*(5), 760-822.

Allensworth, E. (2005). *Graduation and dropout trends in Chicago: A look at cohorts of students from 1991 through 2004*. Chicago: Consortium on Chicago School Research.

Allensworth, E. M., & Easton, J. Q. (2007). *What matters for staying on-track and graduating in Chicago public high schools: A close look at course grades, failures, and attendance in the freshman year*. Chicago: Consortium on Chicago School Research.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Reviews of Psychology, 37*, 1-15.

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 19-32). Hillsdale, NJ: Lawrence Erlbaum.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191-215.

Berger, R. (2003). *An ethic of excellence: Building a culture of craftsmanship with students*. Portsmouth, NH: Heinemann.

Berkowitz, M. W., & Bier, M. C. (2005). *What works in character education: A research-driven guide for educators*. University of Missouri- St. Louis: Character Education Partnership.

Bernard, H. (1991). *Development and application of a diligence-ability regression model for explaining and predicting competence among juniors and seniors in selected Michigan high schools*. Unpublished doctoral dissertation, Andrews University. Berrien Springs, Michigan.

Bingham, W. V. D. (1937). *Aptitudes and aptitude testing*. Oxford, England: Harpers.

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

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20*(1), 89-110.

Carlo, G., Eisenberg, N., & Knight, G. P. (1992). An objective measure of adolescents' prosocial moral reasoning. *Journal of research on Adolescence, 2*(4), 331-349.

Carlo, G., Kohller, S. H., Eisenberg, N., Da Silva, M. S., & Frohlich, C. B. (1996). A cross-national study on the relations among prosocial moral reasoning, gender role orientations and prosocial behaviors. *Developmental Psychology, 32*(2), 231-240.

Carlo, G., McGinley, M., Roesch, S. C., & Kaminski, J. W. (2008). Measurement invariance in a measure of prosocial moral reasoning to use with adolescents from the USA and Brazil. *Journal of Moral Education, 37*(4), 485-502.

Carlo, G., Roesch, S. C., & Kohller, S. H. (1999). Cross-national and gender similarities and differences in prosocial moral reasoning between Brazilian and European-American college students. *Interamerican Journal of Psychology, 33*(1), 151-172.

Carr, D. (2008a). Character education as the cultivation of virtue. In L. P. Nucci & D. Narvaez (Eds.), *Handbook of moral and character education* (pp. 99-116). New York: Routledge.

Carr, D. (2008b). Virtue ethics and the influence of Aristotle. In J. D. Fasko & W. Willis (Eds.), *Contemporary philosophical and psychological perspectives on moral development and education* (pp. 41-59). Cresskill, NJ: Hampton Press, Inc.

Chiu, L. H. (1997). Development and validation of the school achievement motivation rating scale. *Educational and Psychological Measurement, 57*(2), 292-305.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31-44.

Cole, C. (2004). Character development as an outcome of the Ohio Northern University educational experience. *Journal of College and Character, 5*(1), 1-25.

Cornerstone Consulting & Evaluation, L. (2009). *Collective responsibility for excellence and ethics, Hyde Bronx: Student survey item response frequencies and means v3.1, Fall 2008-Spring 2009*. Cortland, NY: Cornerstone Consulting & Evaluation, LLC

Corrigan, M. W., Chapman, P., Grove, D., Walls, R. T., & Vincent, P. F. (2007). The importance of multidimensional baseline measurements to assessment of integrated character education models. *Journal of Research in Character Education, 5*(2), 103-129.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (1rst ed., pp. 621-694). Washington, D.C.: American Council of Education.

Davidson, M., Khmelkov, V. T., & Lickona, T. (2010). The power of character: Needed for, and developed from, teaching and learning. In T. Lovat, R. Toomey & N. Clement (Eds.),

International research handbook on values education and student wellbeing (pp. 427-454). Dordrecht, The Netherlands: Springer.

Davidson, M., Lickona, T., & Khmelkov, V. (2008). Smart & good schools: A new paradigm for high school character education. In L. P. Nucci & D. Narvaez (Eds.), *Handbook of moral and character education* (pp. 370-390). New York: Routledge.

Davidson, M. L., Khmelkov, V. T., & Moran-Miller, K. E. (2006). *Individual & team character in sport questionnaire, v1.2*. Cortland, NY: Cornerstone Consulting & Evaluation, LLC.

Davison, M. L. (1979). The internal structure and the psychometric properties of the Defining Issues Test. In J. R. Rest (Ed.), *Development in judging moral issues* (pp. 223-245). Minneapolis: University of Minnesota Press.

Davison, M. L., Robbins, S., & Swanson, D. B. (1978). Stage structure in objective moral judgments. *Developmental Psychology, 14*(2), 137-146.

Dewey, J. (1975). *Moral principles in education*. Carbondale: Southern Illinois University Press.

Duckworth, A. L., & Seligman, M. E. P. (2005). Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science, 16*(12), 939-944.

Eastman, D. (2008). *Investigating moral and performance character utilizing a student and parent self-reported survey measure*. Unpublished doctoral dissertation, Claremont Graduate University. Claremont, California.

Eisenberg, N., Carlo, G., Murphy, B., & Van Court, P. (1995). Prosocial development in late adolescence: A longitudinal study. *Child Development, 66*(4), 1179-1197.

Eisenberg, N., Zhou, Q., & Kohller, S. H. (2001). Brazilian adolescents' prosocial moral judgment and behavior: Relations to sympathy, perspective taking, gender-role orientation and demographic characteristics. *Child Development, 72*(2), 518-534.

Engelhard, G. (2005). Item response theory (IRT) models or rating scale data. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 2, pp. 995-1003). Chichester: John Wiley & Sons.

Eysenck, S., Easting, G., & Pearson, P. R. (1984). Age norms for impulsiveness, venturesomeness and empathy in children. *Personality and Individual Differences, 5*(3), 315-321.

Eysenck, S., & Zuckerman, M. (1978). The relationship between sensation-seeking and Eysenck's dimensions of personality. *British Journal of Psychology, 69*, 483-487.

Flewelling, R. L., Paschall, M. J., & Ringwalt, C. L. (1993). *SAGE baseline survey*. Research Triangle Institute. Research Triangle Park, NC.

Forsyth, D. (1980). Ethics position questionnaire (EPQ). *Journal of Personality and Social Psychology, 39*, 175-184.

Fournier, G., & Jeanrie, C. (2003). Locus of control: Back to basics. In S. J. Lopez & C. R. Snyder (Eds.), *Positive psychological assessment: A handbook of models and measures* (pp. 139-154). Washington, DC: American Psychological Association.

Freire, P. (1998). *Pedagogy of freedom: Ethics, democracy, and civic courage*. Lanham, MD: Rowman & Littlefield.

Gauld, J. W. (1995). *Character first: The Hyde School difference*. Rocklin, CA: Prima.

Gilligan, C. (1982). *In a different voice: Psychological theory and women's development*. Cambridge, MA: Harvard University Press.

Giroux, H. A. (1992). *Border crossings: Cultural workers and the politics of education*. New York: Routledge.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6*, 427-438.

Haan, N., Brewster-Smith, M., & Block, J. (1968). Moral reasoning of young adults: Political-social behavior, family background, and personality correlates. *Journal of Personality and Social Psychology, 10*(3), 183-201.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Holstein, C. B. (1976). Irreversible, stepwise sequence in the development of moral judgment: A longitudinal study of males and females. *Child Development, 47*(1), 51-61.

Iramaneerat, C., Smith Jr., E. V., & Smith, R. M. (2008). An introduction to Rasch measurement. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 50-70). Los Angeles: Sage.

Jones, R. G. (1968). *A factored measure of Ellis' irrational belief systems*. Wichita, KS: Test Systems, Inc.

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

Kane, M. (2011). *Evolution of our models of validity*. [PowerPoint slides]. *Validity: A General Framework and Prototypes*. University of Georgia workshop. Educational Testing Service.

Khmelkov, V. T., & Davidson, M. L. (2005-2008). *Collective responsibility for excellence and ethics (CREE). Reliability and validity (ver. 2.5)*. Cortland, NY: Cornerstone Consulting & Evaluation, LLC.

Kohlberg, L. (1964). Development of moral character and moral ideology. In M. L. Hoffman & L. W. Hoffman (Eds.), *Review of child development research* (Vol. 1, pp. 383-432). New York: Russell Sage Foundation.

Kohlberg, L. (1984). The meaning and measurement of moral development. In B. Puka (Ed.), *The psychology of moral development: The nature and validity of moral stages* (Vol. 2, pp. 395-425). New York: Harper & Row.

Kuhn, T. S. (1961). The Function of Measurement in Modern Physical Science. *Isis, 52*(2), 161-193.

Kurtines, W. M., & Blank-Greif, E. (1974). The development of moral thought: Review and evaluation of Kohlberg's approach. *Psychological Bulletin, 81*(8), 453-470.

Lai, J.-S., Teresi, J., & Gershon, R. (2005). Procedures for the analysis of differential item functioning (DIF) for small sample sizes. *Evaluation & The Health Professions 28*(3), 283-294.

Lapsley, D. K., & Narvaez, D. (2006) Character education. In W. Damon & R. M. Lerner (Series Eds.) & K. A. Renninger & I. E. Sigel (Vol. Eds.), *Handbook of child psychology: Vol. 4. Child psychology in practice* (pp. 248-296). Hoboken, NJ: John Wiley & Sons.

Lee, V. E., & Burkham, D. T. (2003). Dropping out of high school: The role of school organization and structure. *American Educational Research Journal, 40*(2), 353-393.

Lickona, T. (2004). *Character matters: How to help our children develop good judgment, integrity, and other essential virtues*. New York: Touchstone.

Lickona, T., & Davidson, M. (2005). *Smart & good high schools: Integrating excellence and ethics for success in school, work, and beyond*. Cortland, NY and Washington, DC:

Center for the 4th and 5th R's (Respect & Responsibility) and Character Education Partnership.

Linacre, J. M. (2004). Optimizing rating scale category effectiveness. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models, and applications* (pp. 258-278). Maple Grove, MN: JAM Press.

Linacre, J. M. (2007). *A user's guide to FACETS Rasch-model computer programs*.

Linacre, J. M. (2011). *A user's guide to WINSTEPS MINISTEP Rasch-model computer programs*. Chicago: Mesa Press.

Lind, G. (1978). How does one measure moral judgment? Problems and alternative ways of measuring a complex construct. [German: Wie miβt man moralisches Urteil? Probleme und alternative Möglichkeiten der Messung eines komplexen Konstrukts.] *Sozialisation und Moral*, 171-201.

Lind, G. (1999). *An introduction to the Moral Judgment Test (MJT)*. Konstanz, Germany: University of Konstanz.

Lind, G. (2008). The meaning and measurement of moral judgment competence: A dual-aspect model. In J. D. Fasko & W. Willis (Eds.), *Contemporary philosophical and psychological perspectives on moral development and education* (pp. 185-220). Cresskill, NJ: Hampton Press, Inc.

Lindquist, E. F. (1942). *A first course in statistics, their use and interpretation in education and psychology*. Boston: Houghton Mifflin.

Lockwood, A. T. (1997). *Character education: Controversy and consensus*. Thousand Oaks, CA: Corwin Press.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635-694.

Lufi, D., & Cohen, A. (1987). A scale for measuring persistence in children. *Journal of Personality Assessment, 51*, 178-185.

Marzano, R. J. (2001). *A new era of school reform: Going where the research takes us*. Washington, DC: United States Department of Education.

Mead, R. J. (2008). *A Rasch primer: The measurement theory of Georg Rasch. Psychometrics services research memorandum 2008-001*. Maple Grove, MN: Data Recognition Corporation.

Mentkowski, M. (1980). Creating a mindset for evaluating a liberal arts curriculum where valuing is a major outcome. In L. Kuhmerker, M. Mentkowski & V. L. Erickson (Eds.), *Evaluating moral development and educational programs that have a value dimension* Schenectady, NY: Character Research Press.

Mergler, A. G., Spencer, F. H., & Patton, W. A. (2007). Development of a measure of personal responsibility for adolescents. Retrieved from QUT Digital Repository: http://eprints.qut.edu.au/

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*(11), 1012-1027.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

Nedwek, B. P. (1987). Political socialization and policy evaluation: The case of youth employment and training program. *Evaluation and Program Planning, 10*, 35-42.

Neild, R. C., & Balfanz, R. (2006). *Unfulfilled promise: The dimensions and characteristic's of Philadelphia's dropout crisis, 2000-2005*. Baltimore: Center for Social Organization of Schools, Johns Hopkins University.

No Child Left Behind Act of 2001, 20 U.S.C § 7247 *et seq.*

Noddings, N. (2008). Caring and moral education. In L. P. Nucci & D. Narvaez (Eds.), *Handbook of moral and character education* (pp. 161-174). New York: Routledge.

Character Education Partnership. (2010). A framework for school success: Eleven principles of effective character education. Washington, DC: Character Education Partnership.

Peterson, C., & Seligman, M. E. P. (2004). *Character strengths and virtues.* . Oxford: Oxford UP & American Psychological Association.

Pintrich, P., & DeGroot, E. (1990). Motivated strategies for learning questionnaire (MSLQ). *Journal of Educational Psychology, 82*(1), 33-40.

Power, C. (1980). Evaluating just communities: Toward a method for assessing the moral atmosphere of the school. In L. Kuhmerker, M. Mentkowski & V. L. Erickson (Eds.), *Evaluating moral development and educational programs that have a value dimension* (pp. 177-191). Schenectady, NY: Character Research Press.

Puka, B. (1994). *Kohlberg's original study of moral development* (Vol. 3). New York: Garland.

Rest, J., Thoma, S. J., Moon, Y. L., & Getz, I. (1986). Different cultures, sexes, and religions. In J. R. Rest (Ed.), *Moral development: Advances in research and theory* (pp. 89-132). New York: Praeger.

Rest, J. R. (1976). *Moral judgment related to sample characteristics. Technical report No. 2*. University of Minnesota.

Rest, J. R. (1979). *Development in judging moral issues*. Minneapolis: University of Minnesota.

Richmond, V. P. (1990). Communication in the classroom: Power and motivation. *Communication Education, 39*(3), 181-195.

Rosenbaum, M. (1980). A schedule for assessing self-control behaviors: Preliminary findings. *Behavior Therapy, 11*, 109-121.

Royal, K. D. (2008). *Using item response theory and the Rasch measurement model to investigate faculty perceptions of instructional goals*. Unpublished doctoral dissertation, University of Kentucky, Lexington, KY.

Rubin, R. B., Palmgreen, & Sypher, H. (1994). *Communication research measures: A sourcebook.* New York: Guildford Press.

Ryan, K. (1986). The new moral education. *Phi Delta Kappan, November*, 228-233.

Ryan, K., & Bohlin, K. E. (1999). *Building character in schools: Practical ways to bring moral instruction to life.* San Francisco: Jossey-Bass.

Schwarzer, R., & Jerusalem, M. (1995). Generalized self-efficacy scale. In J. Weinman, S. Wright & M. Johnston (Eds.), *Measures in health psychology: A user's portfolio. Causal and control beliefs* (pp. 35-37). Windsor, UK: NFER-NELSON.

Sherer, M., Maddux, J. E., Mercandante, B., Prentice-Dunn, S., Jacobs, B., & Rogers, R. W. (1982). The self-efficacy scale: Construction and validation. *Psychological Reports, 51*, 663-671.

Siddle-Walker, V., & Snarey, J. R. (2004). Race matters in moral formation. In V. Siddle-Walker & J. R. Snarey (Eds.), *Race-ing moral formation: African American perspectives on care and justice* (pp. 1-14). New York: Teachers College Press.

Simpson, E. L. (1974). Moral development research: A case study of scientific cultural bias. *Human Development, 17,* 81-106.

Snarey, J. R. (1985). Cross-cultural universality of social-moral development: A critical review of Kohlbergian research. *Psychological Bulletin, 97*(2), 202-232.

Solomon, D., & Watson, M. S. (2008). Moral education approaches and their psychological foundations. In J. D. Fasko & W. Willis (Eds.), *Contemporary philosophical and psychological perspectives on moral development and education* (pp. 113-142). Cresskill, NJ: Hampton Press, Inc.

Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality, 72*, 271-322.

Thoma, S. (1984). Do moral education programs facilitate moral judgment? A meta-analysis of studies using the defining issues test. *Moral Education Forum, 94*(4), 20-25.

Thomas, R. S. (1991). Assessing character education: Paradigms, problems, and potentials. *The Clearinghouse, 65*(1), 51-55.

United States Department of Education Office of Safe and Drug-Free Schools. (2007). *Mobilizing for evidence-based character education*. Washington, DC.

Vallerand, R., Pelletier, L., Blais, M., Briere, N., Senecal, S. E., & Vallier, F. (1993). Academic motivation scale. *Educational and Psychological Measurement, 53*(1), 159-172.

Walker, L. J. (1991). Sex differences in moral reasoning. In W. M. Kurtines & J. L. Gewirtz (Eds.), *Handbook of moral behavior and development* (pp. 333-364). Hillsdale, NJ: Lawrence Erlbaum.

Whiteside, S. P., & Lynam, D. R. (2001). The five factor model and impulsivity: Using a structural model of personality to understand impulsivity. *Personality and Individual Differences, 30*(4), 669-689.

Williams, M., & Schnaps, E. (Eds.). (1999). *Character education: The foundation for teacher education*. Washington, DC: Character Education Partnership.

Willis, W. (2008). Introduction to Part one: Philosophical foundations of moral education. In J. D. Fasko & W. Willis (Eds.), *Contemporary philosophical and psychological perspectives on moral development and education* (pp. 1-7). Cresskill, NJ: Hampton Press, Inc.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrencence Erlbaum.

Wright, B. D. (1977). Misunderstanding the Rasch model. *Journal of Educational Measurement, 14*(3), 219-225.

Wright, B. D. (1996). Local dependency, correlations and principal components. *Rasch Measurement Transactions, 10*, 509-511.

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; measurement, however, must be interval. *Archives of Physical Medicine and Rehabilitation, 70*, 857-860.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: Mesa Press.

Wright, B. D., & Mok, M. M. C. (2004). An overview of the family of Rasch measurement models. In J. E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models, and applications* (pp. 1-24). Maple Grove, MN: JAM Press.

Wyatt, J. M., & Carlo, G. (2002). What will my parents think? Relations among adolescents' expected parental reactions, prosocial moral reasoning and prosocial and antisocial behaviors. *Journal of Adolescent Research, 17*(6), 646-666.

Zimmerman, B. J., & Kitsantas, A. (2005). Homework practices and academic achievement: The mediating role of self-efficacy and perceived responsibility beliefs. *Contemporary Educational Psychology*, *30*, 397-417.