2017

DIMENSIONALITY ANALYSIS OF THE PALS CLASSROOM GOAL ORIENTATION SCALES

Angela K. Tombari
University of Kentucky, angela.kristi.tombari@gmail.com
Digital Object Identifier: https://doi.org/10.13023/ETD.2017.230

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Angela K. Tombari, Student

Dr. Michael D. Toland, Major Professor

Dr. Kenneth Tyler, Director of Graduate Studies

DIMENSIONALITY ANALYSIS OF THE PALS CLASSROOM GOAL ORIENTATION SCALES

DISSERTATION

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the College of Education at the University of Kentucky

By

Angela K. Tombari

Lexington, KY

Director: Dr. Michael D. Toland, Professor of Educational, School, & Counseling Psychology

Lexington, KY

2017

ABSTRACT OF DISSERTATION

DIMENSIONALITY ANALYSIS OF THE PALS CLASSROOM GOAL ORIENTATION SCALES

Achievement goal theory is one of the most broadly accepted theoretical paradigms in educational psychology with over 35 years of influencing research and educational practice. The longstanding use of this construct has led to two consequences of importance for this research: 1) many different dimensionality representations have been debated, and 2) methods used to confirm dimensionality of the scales have been supplanted from best practice. A further issue is that goal orientations are used to inform classroom practice, whereas most measurement studies focus on the structure of the personal goal orientation scales rather than the classroom level structure. This study aims to provide an updated understanding of one classroom goal orientation scale using the modern psychometric techniques of multidimensional item response theory and bifactor analysis. The most commonly used scale with K-12 students is the Patterns of Adaptive Learning Scales (PALS); thus, the PALS classroom goal orientation scales will be the subject of this study.

KEYWORDS: Bifactor, PALS, Classroom Goal Orientations, Multidimensional Item Response Theory, Psychometrics

Angela K. Tombari

6/8/2017

DIMENSIONALITY ANALYSIS OF THE PALS CLASSROOM GOAL ORIENTATION SCALES

By

Angela Kristi Tombari

Dr. Michael D. Toland
Director of Dissertation

Dr. Kenneth Tyler
Director of Graduate Studies

6/8/2017

ACKNOWLEDGMENTS

The following dissertation benefited greatly and would not have been possible without my committee chair, Dr. Michael D. Toland, and the support of my amazing committee, notably Dr. Fred Danner, Dr. Kenneth Tyler, and Dr. Joan Mazur. They assured me that I have some modicum of intelligence even when I doubted my capacity, supported me in my literal darkest days (winter), and provided allowances for my intermittent eccentricities (to put it mildly). I also want to thank Dr. Bradley who was on my committee throughout the majority of my time as a graduate student, assuring me that I would eventually finish and encouraging me to seek help when necessary. To all of these individuals, I recognize that being a professor is a harsh and demanding task, and yet you manage to balance the difficulty of your rigorous publication and teaching schedules with an unprecedented amount of empathy and student support. I want to thank you for providing me with an experience of which many graduate students can only dream.

This dissertation would not have been completed without the support and encouragement of many University of Kentucky staff who have become my friends over the years. Particularly, Fred Blackford who always greeted me with a hug, encouragement, and sometimes a quarter for the meter. Additionally, I want to thank Tammy Medaris and Stu Reedy. There are really no words to describe my appreciation for the support and love that these awesome individuals have surrounded me with throughout this journey.

Of equal importance, this dissertation would not have been completed without the extensive support, both emotionally and financially, and the occasional harassment of my family. Without the good-natured ribbing at all family gatherings, this dissertation may have only existed in potentia. My husband, Brandon, made sure to extend the harassment/encouragement to a daily occurrence, forcing me to complete this dissertation if only to eliminate this source of conversation. He is also an amazing and loving man who I feel incredibly lucky to have in my life. I can write this with impunity because I am positive he will never read this full document.

I do want to offer heartfelt thanks to my family. All commentary on family is necessarily abbreviated because I could not detail the full extent of their support throughout this process. My father, Michael Tombari, encourages a love of knowledge and a striving for truth. He finds the time and energy to rationally discuss oppositional points of view and to provide emotional and financial help. My mother, Patty Tombari, has dedicated her life to the happiness of her family and the fostering of upstanding behavior. She continually provides support to both her family and strangers. My sister, Tracy Tombari, who attacks life with a doggedness only softened by her acute wit, has been an unbending ear throughout this process. She is a lifeline to me and has guided me since day one. A final thanks to my Nonno, Aldo Tombari, is necessary. He was a man who never stopped embracing life and knowledge. There is no better example of how to live life than that embodied by these amazing people. I could not have been surrounded by a more supportive group.

And finally, my dog, Athena, has found a way to cuddle with me while allowing me the ability to type: for that, she deserves applause. She has also managed to exhibit remarkable patience as I announce, "just one more page, pup, then we will go outside." Another integral component of this dissertation was the constant flow of coffee, often provided by Third Street Stuff. Their outdoor seating area allowed me to write while feeling as though life was not passing me by, even though it was literally happening around me while I sat in one place.

Overall, although the writing may be mine, the credit for any laudable aspects of this dissertation belongs with others. All errors, of course, are my own.

TABLE OF CONTENTS

List of Tables

List of Figures

DIMENSIONALITY ANALYSIS OF THE PALS CLASSROOM GOAL ORIENTATION SCALES

Chapter One: Introduction

Teachers are tasked with the job of imparting knowledge just as students are tasked with the job of assimilating knowledge. If this were a simple task, this dissertation would end here; unfortunately, the abyss between teacher and student necessitates the bridge of learning, a process often requiring a great deal of effort and persistence, even in the face of failure. The initial failures encountered during early learning may be forgotten once an individual achieves great success, but even the greatest trumpet player once struggled to play a single note. Journeying from novice to trumpet virtuoso does not occur overnight; rather, this process requires hitting untold milestones along the way. The choice of milestones for which to aim and the cultivation of persistence in the face of difficulty are key components of moving from the imagining of an outcome to the accomplishment of an outcome. The drive to persist towards a desired goal is often attributed anecdotally to possessing sufficient motivation.

What is Motivation?

Although a drive to persist describes motivation, it does not define the term. No cohesive definition has attained universal acceptance, so one can resort to the roots of the actual word: motivation derives from a Latin verb which literally translates as the infinitive, to move (Schunk, Pintrich, & Meece, 2008). As such, motivation is not a tangible or even an intangible object; rather, the term motivation refers to the process of being driven to act. This course of action inherently involves both behavior directed towards some sort of goal as well as the maintenance of this behavior in the face of adversity. One useful distinction to draw in understanding motivation concerns the difference between motivation and motive. A motive answers the *why* of engagement, whether it is for money, fame, or perceived accolades, while motivation is the *process* of attaining the desired goal.

Motivation as a construct has been a known topic of study stretching as far back as the writings of Plato; however modern technology and mathematical techniques have allowed for the refinement of theories through an understanding of the properties of the items used to assess a given motivational construct (Jones & Thissen, 2007). These techniques are frequently grouped under the label of psychometrics: psychometrics can be more easily understood as the application of measurement theory to psychological constructs, a burgeoning movement of the twentieth century solidified in purpose with the formation of the Psychometric Society in 1935 (Jones & Thissen, 2007). As psychometric methods advanced and motivational constructs were refined and tested, theorists operating from different paradigms developed divergent perspectives concerning motivation, its origins, and methods of intervention to increase student motivation. One approach which has been used to drive both classroom-level and school-level interventions is achievement goal theory (e.g., Ames, 1992). Since achievement goal theory has been a popular framework for classroom research and interventions for the last quarter decade, the accuracy with which this construct is being measured is of paramount importance.

What are its Correlates?

Achievement goal theory initially parsed achievement into two different types, performance and mastery motivation, as will be discussed in greater detail in Chapter 2 (e.g., Elliot, 1999). These two types of motivation were posited to differ in the underlying outcome which fueled the individual to pursue a task, with mastery orientation denoting pursuit with the intent of mastering the task while performance motivation denoting pursuit with the intention of outperforming others at the task. Using this type of division for motivation, mastery orientation was generally concluded to be the desired form of motivation (see Urdan, 1997, for review). Mastery goal orientation has been found to be positively correlated with a host of desirable cognitive, affective, and behavioral outcomes including: the desire to pursue a task for the task itself and increased recruitment of self-regulatory strategies (e.g., Ames & Archer, 1988; Duda & Nicholls, 1992); greater persistence in the face of failure (e.g., Elliot & Dweck, 1988); and positive emotions, such as pride and satisfaction, in light of success on a task (see Ames, 1992, for review). In contrast, a performance goal orientation was perceived as a potential liability, found to be correlated to a bevy of both positive and negative outcomes, including: attributing failure to perceived lack of ability, devoting less time to a task when confronted with failure, and selecting overly easy or overly difficult tasks (see Ames, 1992 for review); increased academic self-efficacy and higher grade point average (Roeser, Midgley, & Urdan, 1996); and increased mastery goal orientation (e.g., Roeser, Midgley, & Urdan, 1996). The confusion in outcomes related to performance goal orientation led to a further splintering of the construct in an attempt to clarify theoretical relationships by bifurcating performance goal orientation into an approach and an avoid dimension. This shift was based on an in-depth analysis by Elliot (1999) who included both a theoretical review of the literature and a classification of prior studies based on: whether the participants were oriented towards an approach form of performance goal orientation (in cases of goal manipulation), and whether the survey instrument included solely performance approach items or both performance approach and avoid items. This review found that studies using only performance approach items or manipulations showed positive correlations with more optimal outcomes such as intrinsic motivation whereas mixed instruments showed the traditional pattern of mixed valence correlated outcomes (Elliot, 1999). Since the bifurcation, both performance approach and mastery approach motivation have been shown to be related to positive outcomes with performance approach generally relating to improved performance metrics such as higher grades and mastery goal orientation generally relating to metrics conducive for success such as increased interest in the subject matter and increased positive affect (Harackiewicz, Barron, Pintrich, Elliot, & Thrash, 2002). Despite the perceived benefit of bifurcating the performance dimension, mixed results still occur suggesting the possibility of measurement issues (Anderman & Midgley, 2002). Discriminant and convergent validity evidence has been used to determine the factor structure of the achievement goal framework, but models designed to tease out methods effects such as negative item phrasing or similar item content have not yet been implemented to examine impact on dimensionality (Elliot, 1999; Reise, Morizot, & Hayes, 2007). Usage of models, such as the bifactor model (Holzinger & Harman, 1938), would enable a better investigation of the dimensional structure underlying scales designed to assess goal orientations (see Reise, Morizot, & Hayes, 2007; Toland, Sulis, Giambona, Porcu, & Campbell, 2017, for examples).

Methods of Study

Since the origin of the scientific method, replicable observations have become the gold standard of scientific inquiry. Whether these observations support or refute a hypothesis falls within the purview of both math and logic, generating entire fields of research and inquiry dedicated to the efficacy of the methods used to analyze recorded observations. Almost synchronous to the adoption of the scientific method was the realization that the process of measuring observations introduces error (Jones & Thissen, 2007), with early developments in error theory originating from research by the astronomer Bessel. His attention to the discrepancies in observations of the time at which major astronomical events occurred led to the revelation that the precision of a measurement instrument inversely relates to the amount of error in the measurements derived from the aforementioned instrument (Jones & Thissen, 2007). While measurement precision is a relatively easy concept to grasp in the realm of tangible objects, attempting to quantify measurement precision in the measurement of intangibles is a much more difficult issue.

One particular issue when measuring intangible constructs such as motivation is the heavy reliance on Likert-type response scales (e.g., Strongly Disagree to Strong Agree). The scales themselves are not the issue, but the methods used to analyze data produced using Likert-type response scales can be. Likert-type response scales have historically been treated as continuous measures despite the knowledge that data produced using these scales is truly categorical in nature. Treating categorical data as though it is continuous or essentially linear has well-studied and acknowledged shortcomings; however the mistreatment of categorical data continues relatively unabated. Historically, computational limitations necessitated this treatment, but technological progress has expanded the purview of implementing improved psychometric methods, designed to handle coarsely categorized measures, into the realm of the average applied researcher. Models designed to correctly treat polytomous data emerged with Samejima's (1969) graded response model, resulting in the beginning of a time period characterized by explosive growth in psychometric methods (Jones & Thissen, 2007). The result is that researchers are at a juncture where constructs measured via instruments consisting of several items used to tap said construct should be examined using more powerful and exact psychometric methods, which increase the accuracy and precision of dimensionality assessment when applied to polytomous data. The intention of this study is to use modern psychometric techniques designed to handle coarsely categorized measures to elucidate the dimensionality of the classroom goal orientation scales from the Patterns of Adaptive Learning Scales (PALS; Midgley, Maehr, Hruda, Anderman, Anderman, Freeman, et al., 2000). More specifically, the literature review will discuss several distinct formulations of the theoretical structure of the personal goal orientation scales which were extended by proxy to the theoretical structure of the classroom goal orientation scales. This study will use multi-dimensional item response theory [MIRT], a full-information technique, to compare model-data fit for each of the hypothesized models. The bifactor model, which will be introduced in Chapter 2 will provide additional information concerning the plausibility of each of these hypothesized models. Once the best fitting model has been selected for this data, item characteristics of unidimensional subscales will be analyzed in greater depth to elucidate areas of shortcoming or strength for this popular instrument.

Chapter Two: Literature Review

Motivation can be perceived as one of the most important qualities that an individual may possess; however, defining what motivation actually consists of can present difficulty. As mentioned in Chapter 1, motivation entails not just the drive to act, but the ability of this drive to persevere. One can imagine motivation balancing against difficulties, with the desire to achieve optimally outweighing the consequent difficulties emergent in the process of goal attainment. This raises a question concerning motivation: is motivation balanced against hindrances such that the sheer quantity of motivation is the pertinent factor or is the quality of motivation the pivotal piece in determining one's reaction to setbacks? Furthermore, if quality of motivation is important, what determines motivation quality? These questions became central upon the shift in focus to social-cognitive theories as explanatory in the motivation literature. Goal orientation theories, one branch of social-cognitive theorizing on motivation, arose from pondering apparent paradoxes, with key researchers questioning why students of equivalent confidence and sufficient ability to accomplish the task exhibited variable responses to failure. In order to understand goal theory, one must first understand the historical context in which this theory was developed.

Theoretical Framework

From the early 20th century until the mid-nineteen seventies, behaviorism was the primary theoretical orientation driving psychological research. Behaviorism focused on behaviors and environmental influences, with a strict removal of hypotheses concerning the internal mechanisms of the mind. Although the behaviorist approach allowed for much advancement in psychology as a whole, abandonment of behaviorism as a central tenant of psychology expanded the scope of psychological research to include the effects that individual cognitive construal of the outside world had on actions (Schunk, Pintrich, & Meece, 2008).

Behaviorism originated with the idea of classical conditioning as advanced by animal studies in which innate needs/drives were easily identifiable. Early works relied heavily on a simplification of life processes, resulting in a theory of motivation initially reliant on using basic life needs, such as food, water, and sleep, to promote the reoccurrence of desired behaviors (Schunk, Pintrich, & Meece, 2008). When attempting to generalize these principles to human motivation, the human ability to subvert basic genetic drives, or life needs, in lieu of principles created problems for behaviorism. Examples of subverting basic needs to promote a moral principle can be seen in the occurrence of hunger strikes or monk self-immolation for a cause. To handle the increased mental complexity of humans while allowing for the base simplicity demanded in behaviorist models, behaviorism was expanded to include operant conditioning. Operant conditioning extended the concept of reinforcement outside of fundamental life-supporting components to include desired responses from others. This extension allowed for additional reinforcers, such as stickers for a kindergarten-aged student or social approval for an adolescent, while maintaining the ability to theorize without including the components of cognition within the body of the theoretical framework; however, the concept of reinforcers is the primary weakness of behaviorism, in that the definition of reinforcement is inherently tautological. Essentially a reinforcer is anything that increases the occurrence of a behavior; similarly, if a behavior begins occurring more frequently, the response to the behavior must have been a reinforce (Chomsky, 1959; Skinner, 1953). With this definition, the predictive utility of this theory is undermined as a reinforcer useful for one individual may be useless for another individual. Despite the disparity in interpretation of reinforcers across people, the essential premise of behaviorism is unaffected, leaving the theory intact. When intensive time can be spent with an individual to ascertain what actions/behaviors can serve as reinforcement for this particular person, behaviorism can be very effective; however, with the ambiguity of reinforcement across people, widespread application of this theory, as might be seen in a classroom, would be difficult if not impossible. Theories involving cognition emerged to remedy this issue by including the cognitions of the participant as an integral component of the theory. The basis of cognitive models is often rooted in the behaviorist framework such that reinforcement for a particular behavior is posited to exist; however, from the cognitive approach, the focus is on the cognitions attached to the reinforcement rather than the reinforcement itself (Brophy, 1999).

Overview of Intermediary Theories

Between the pinnacle of the behaviorist era and the advent of the social-cognitive meteoric rise to prominence, exists a time of intermediary theories designed to modify behaviorism to better apply to and be predictive of human actions. Two methods of modifying behaviorism existed, both espoused as an outright rejection of the behaviorist foundations: humanistic responses rejected the notion that humans were driven by base drives in favor of a more nuanced perspective of human needs, whereas achievement goal and achievement motive theorists expressed an early version of later cognitive theories by shifting the focus from reinforcements to the cognitions related to the anticipation of reinforcement.

The first modification, humanism, was espoused as an outright rejection of the behaviorist premises and resulted in the effort to define and classify the essential human needs as undertaken by Murray, Maslow, and McClelland although under different auspices. As the classification of needs continued, certain tautological problems emerged such that theorizing tended to occur in the fashion that individual A would engage in action B for need C, and our knowledge that need C existed was due to the fact that individual A engaged in activity B (Brophy, 1999). Need theory also struggled with balancing the level of the need and the different aspects of need expression. Needs were typically perceived as a dispositional trait such that some individuals had higher “needs for achievement” than other individuals. Despite this classification, individuals with the same dispositional needs often reacted differently in specific situations, a variability unexplained by traditional need theories. As a result of these tautological and situational shortcomings, goal theory, a subset of social cognitive theory, emerged to capture the situational context of behavior.

The second reactionary theory development involved a shift to the cognitions involved in rewards, the ignored black box of behaviorism. Early attempts took two main forms: achievement motive such as expectancy-value theory and achievement attribution theory (for review, see Weiner, 2010). Both of these theories were deemed to fall short with the former over-emphasizing dispositions to the neglect of cognitions while the latter emphasized a drive towards achieving competence while failing to sufficiently expound upon the origins of this drive for the human condition (Elliot 2005). Achievement goal theory, falling under the social-cognitive framework, attempted to remedy these perceived faults (Elliot, 2005; Urdan, 1997).

Overview of social cognitive theory. Social cognitive theory arose to counter the theoretical void created when behaviorism began to decline in popular psychology. This theory offered intuitive appeal by treating the human mind as an integral component, useful for study, when explaining human actions and reactions. One theory, which has been broadly studied in public school systems and falls under the social cognitive paradigm, is achievement goal theory (Schunk, Pintrich, & Meece, 2008).

Development of goal structures. Goal orientations were partially created to remedy the previously discussed gap in the attributional literature, namely the tautological basis of need theorizing and the failure of need theories, which were centered on individuals possessing different levels of these needs, to explain intra-individual differences in goal directed behavior across scenarios (Elliot, 2005). The question left unresolved through these theories was: If an individual has a particular level of need for achievement as an attribute, why does this individual display apparent differences in achievement behavior across different scenarios? If a student redoubles his effort when encountering a difficult problem in math class, why does this same student abandon their work when encountering difficulty in history class? In order to remedy this gap, goal orientations were conceived as a construct that exists at multiple levels such that the individual can have a tendency to select a particular goal orientation when approaching tasks which can be modified by the perceived goal orientation structure of the teacher, classroom, or school (e.g. Ames & Archer, 1988). Ironically enough, with the express purpose of addressing this perceived deficit in the attributional theories, most of the theorizing related to goal orientations remained focused on the level of the individual (e.g., Elliot & McGregor, 2001).

Mastery and Performance Motivation

Original formulation. Goal structures emerged as a social-cognitive theory intended to assess: individual understanding of a situation, personal construal of this scenario, and the individual aims that drive the individual to seek success in the endeavor (Dweck, 1986). The synchronous emergence of achievement goals derived primarily from two individuals: Dweck (1986) posited two different learning patterns, one maladaptive and one adaptive, explained by adoption of either performance or learning goals in an achievement scenario (also see Dweck & Leggett, 1988), while Nicholls (1984) suggested differing approaches to pursuing a goal based on whether the child had a differentiated or undifferentiated conception of ability. These two conceptions of ability were theorized to differ in the perceived relationship between effort and ability, with the former positing an inverse relationship while the latter posited a direct relationship (Urdan, 1997). The intricacies of these different approaches will be discussed more as we examine the theoretical origins of goal orientations in more depth.

Founding origins. Goal orientation theory developed from two distinct main sources while incorporating the thoughts and ideas of many different prestigious motivation researchers of the time. The two individuals, Dweck and Nicholls, who pursued the development of this theory with fervor, each went on to eventually articulate separate divergent and nuanced theoretical frameworks for the same basic constructs of achievement goal orientation (e.g., Dweck, 1986; Nicholls, 1984); however, these two theories also possess many commonalities due to shared premises, generated through the collaborative effort of many motivation experts. Of note, a seminar series occurring at the University of Illinois in the late 1970s, including Carol Ames, Carol Dweck, Marty Maehr, and John Nicholls amongst others, engendered discussions pivotal to the subsequent synchronistic development of the achievement goal framework (Elliot, 2005).

Similarities and differences between the originating theories. The nascent premise of goal orientation theory had both similarities and differences when considering these two main fonts of theoretical development. Both theorists: (1) sought to expand the achievement motivation literature by addressing different ways to conceptualize ability; (2) simplified the cognitions related to achievement into one of two main categories; and (3) conceived of this framework as functional at both the dispositional and situational levels (e.g., Dweck, 1986; Nicholls, 1984). An understanding of the similarities between the theories is a necessary precursor to exploring the subtle differences between them.

Conceptualizing ability and the superordinate goals of goal theory. Both Dweck and Nicholls focused on how individuals conceive of ability as the root of their theorizing; however, their conceptions diverged in qualitative focus. Dweck began by pondering paradoxical results from prior studies in which students of equivalent ability exhibited disparate responses to failure, with some students redoubling their efforts while others abandoned the task (Diener & Dweck, 1978). In response, Dweck developed the superordinate purposes of goal orientation theory such that achievement behavior is believed to originate from one of two purposes: 1) to try and accomplish a task, or 2) to show competence or avoid showing incompetence. This second purpose was proposed as linked in Dweck's early theorizing even though further discussion will address the current conceptualization of this as bifurcated by an approach-avoid dimension (Dweck, 1986; Elliot, 2005).

Nicholls (1984) approached the conceptualization of goal orientations from a different perspective, yet came to similar conclusions as to the utility and division of the overarching goal orientations that influence achievement behaviors and subsequent outcomes. Unlike Dweck's focus on response to failure, Nicholl's began his theorizing through work with conceptions of ability. Nicholl's noted two predominant views on the relation between ability and effort which emerged around the beginning of adolescence. He posited that individuals could conceptualize of ability as being either directly or inversely related to effort such that individuals who possess an undifferentiated conception of ability believe effort and ability to be directly related while individuals with a differentiated conception of ability believe effort and ability to be inversely related. This premise led Nicholl's to posit a slightly different root to achievement behaviors than was espoused by Dweck (1986): Nicholl's hypothesized that achievement behaviors are designed to either show ability or avoid showing lack of ability (essentially bifurcating the second purpose proposed by Dweck).

Dweck (1986) worked backwards, based on her observations, from her theory that individuals demonstrate different overarching goals in any given achievement situation to formulate a plausible cognitive explanation as to why these individuals selected different approaches. Her original research, in which she observed individual students displaying disparate responses to failure, included a measure of confidence in one's ability which she expected to act as a predictor of persistence when confronted with failure; however, her findings did not support this initial supposition (Diener & Dweck, 1978). The cognitive solution that Dweck and associates arrived at was that individual students have an implicit conception of ability (later and more familiarly known as intelligence) in which ability can be perceived in one of two ways: malleable (also termed incremental) or fixed (also termed entity) (Dweck, 1986; Dweck & Leggett, 1988). Individuals who perceive ability as malleable are more likely to persist in the face of failure due to a belief that their ability can be improved, promising the possibility of future success. In contrast, individuals who perceive ability as fixed are more likely to abandon a task in the face of failure, believing that their ability is not equal to the task and seeking to avoid further demonstrations of inadequacy. Dweck (1986) further postulates that beliefs concerning ability lead the student to select a referent for determining success, with incremental ability theorists using themselves as a referent when assessing ability while fixed ability theorists use others as a referent for determining success.

By approaching the problem from the opposite direction, beginning with conceptions of ability and moving towards differential outcomes when placed in an achievement situation, Nicholls (1984) proposed the reverse relationship from Dweck (1986): namely, because we are all seeking to demonstrate ability, the difference in motivational frameworks relies on how we determine ability. The individual chooses either to self-reference ability at a particular task or to other-reference performance which leads to different salient conceptions of the relationship between ability and effort. The differentiated conception of ability/effort is a consequence of other-referenced ability assessments while the undifferentiated conception of ability/effort is a consequence of self-referenced assessments. Nicholls (1984) posited that when the individual perceives ability as other-referenced, task difficulty is assessed by success rate of others in accomplishing the task while ability assessment is rooted in an assumption of equal effort towards the task at hand (see Figure 2.1). These two features of other-referenced ability assessment, when considered in tandem, suggest that effort and ability are inversely related, rendering the individual susceptible to certain maladaptive conclusions. One such conclusion is that in order to show equal ability, one must succeed with the same or less effort than others use at the same task. Taken to the extreme, individuals may engage in behaviors often deemed as self-handicapping in order to prevent a negative assessment of ability by themselves or others. For example, an individual may avoid studying for a test to provide a default attribution of lack of effort in case of failure. From the differentiated conception, if one were to study hard for a test and then fail, a lack of ability would seem an inevitable conclusion (Nicholls, 1984).

In contrast, the undifferentiated conception of effort/ability is rooted in self-referential analysis. One compares his/her ability to accomplish the task to the task requirements at hand. An implicit assumption of this conception is that effort and ability are directly related. One implication of this assumption is that tasks are perceived as difficult or easy by comparing the task to one's current assessment of one's ability; furthermore and more importantly, 'difficult' tasks are viewed as providing a greater opportunity to show ability due to the possibility of greater improvement. The ultimate implication is that an individual holding a differentiated conception of ability sees a difficult task as proffering a strong possibility for showing a lack of ability, while an individual holding an undifferentiated conception of ability will see a difficult task as proffering a strong possibility for showing the presence of ability. This is the crux of Nicholl's (1984) achievement goal theory and the explanation of different reactions to difficult tasks across individuals. Interestingly enough, even though Dweck (1986) and Nicholls (1984) perceive the first two variables in opposite order, they come to the same conclusions about the type of superordinate goals that may be pursued in any given academic situation.

Simplifying the goal framework to two main patterns. The final outcome for both Dweck (1986) and Nicholls (1984) was the belief that two different superordinate goals could exist for a student in any given achievement situation. Dweck (1986) characterized these two goal orientations as learning versus performance goals and ascribed particular behaviors, cognitions, and affective responses to each of these goal orientations. In contrast, Nicholls (1984) termed the two possible goal orientations as task-involvement and ego-involvement to reflect his belief in the driving nature of the referent, self- versus other- respectively, in determining the conception of ability pertinent in a given situation. Similar to Dweck, Nicholls conceived of each of these goal orientations as reflecting a superordinate goal striving, the 'why' of achievement behavior, associated with its own constellation of behaviors, cognitions, and affective responses. Despite their similarities, Nicholls (1984) posited that ego-involvement exerted a potentially positive influence under certain circumstances in which the student perceived of initial ability as high and did not encounter failure. These two perspectives on goal orientation were still more similar than different, portending their future hegemony by Ames and Archer (1987, 1988).

Addressing both the situational and dispositional levels. Both Dweck and Nicholls attempted to compensate for perceived shortcomings in prior achievement attribution literature, which focused on dispositions as explanatory factors for behavior, by developing a theory with situational applications. The situation focus of early goal-orientation research is apparent from both the use of situationally induced goal orientations as well through perusing writings of these early theorists which detailed the potential situational applications of the theory (e.g., Ames 1984; Elliot & Dweck, 1988; Elliot & Harackiewicz, 1996; Nicholls, 1975; Nicholls, 1984). Experimental inductions were intended to influence the prominence of one goal orientation versus another through methods such as: providing directions that suggest a task is an ability assessment versus a fun activity; providing feedback after an initial task to orient the student towards either evaluation or skill acquisition (Elliot & Dweck, 1988; Elliot & Harackiewicz, 1996); or instituting a competitive reward structure for some individuals (e.g. Ames, 1984; Nicholls, 1975). Along with the experimental manipulations, these theorists explicitly detailed components of situations that would promote the adoption of an ego-involvement, or performance goal, orientation (e.g. Dweck, 1986; Nicholls, 1984).

The convergence of the three. As can be seen from the above discussion, the use of multiple terms across theorists can generate unnecessary complexity, as the mechanisms and constructs detailed show far more similarities than differences (Elliot, 2005). With one fell stroke, Ames and Archer (1987) ended the proliferation of terminology in the goal orientation literature by declaring: "because the conceptual relation between task and learning goals or ego and performance goals is convergent, these perspectives have been integrated, and hereafter are identified as mastery and performance goals, respectively" (p. 409). Ames and Archer (1987, 1988) additionally summarized mastery goals as reflecting a desire to improve on a task, rendering effort a positive component of the process; in contrast, performance goals were characterized as reflecting a desire for normative success with as little effort as possible (to reflect greater ability). This synopsis effectively eliminated the saliency of the most prominent differences between these two theories: whether an individual *first* selects a referent for comparison *and then* determines the pertinent conceptualization of ability or whether the individual selects a referent for comparison *based upon* personal conceptualization of ability. Rather than be concerned as to which variable precedes which, the focus shifted to the perceived relationship of effort to success in each of these two goal orientations (Ames & Archer, 1987; Ames & Archer, 1988).

Further developments. Two main changes have been suggested since the integration of goal orientation theory: (1) the addition of the approach-avoid distinction, and (2) the controversy between proponents of the mastery goals perspective and proponents of the multiple goals perspective. The approach-avoid dimension of achievement goals was suggested as an important theoretical component intended to clarify mixed findings regarding the effects of ascribing to a performance goal orientation (Elliot, 1994). The apparent utility of this distinction garnered strong support for the bifurcation of performance goals into performance approach and performance avoid dimensions (e.g., Elliot & Church, 1997; Middleton & Midgley, 1997; Urdan & Midgley, 2001; Vandewalle, 1997) leading to the creation of instruments intended to assess this bifurcated construct (e.g., Elliot & Church, 1997, Vandewalle, 1997) as well as the expansion of current instruments to include assessment of the performance avoid construct (e.g., Midgley et al., 2000). The division of performance goals into approach and avoid dimensions led to the concomitant fission of achievement goal researchers into two different theoretical camps differing on two key ideas: 1) the relative utility of performance approach and mastery goals, and 2) whether goal adoption consists of adopting a singular goal or multiple goals simultaneously. Considered in tandem, opinions on these two ideas are represented by the mastery goal perspective and the multiple goals perspective (Harackiewicz et al., 2002).

Ways to Measure these Constructs

Although Ames and Archer (1987) successfully stymied the unnecessary expansion of goal orientation terminology, the problem of a proliferation of self-report goal orientation instruments continued unabated. Within one of the two articles lauded as creating a convergence of goal orientation theory, Ames and Archer (1988) contributed to the plethora of instruments by creating yet another measure to assess perceptions of classroom goal structures. This suffusion of instruments for assessing goal orientations is characteristic of the nascent stage of theorizing on any construct; however, as a construct continues to develop, standardization of instrumentation across studies allows for increased comparability and generalizability of research. Currently, several instruments have emerged as prominent in the literature.

The two most popular instruments for assessing goal orientations are the Achievement Goal Questionnaire (AGQ; Elliot & Church, 1997; Elliot & McGregor, 2001) and the PALS (Midgley et al., 2000). In an effort to determine whether these popular measures actually assess the same constructs, a meta-analysis of 243 studies published prior to the year 2007 was undertaken by Hulleman, Shrager, Bodmann, and Harackiewicz (2010). The crux of this study was an investigation into whether the representation of goal orientation theory as a unitary concept was supported empirically both through cross-instrument study of the qualitative content of the measures as well as an examination of the quantitative relationships found between constructs and outcome measures. To assess this, the authors undertook an extensive review of the theoretical literature to define the potential components of each of the popular constructs followed by a review of the actual content of the questions used to measure these constructs across instruments/studies. One key conclusion from this analysis was that the two most popular instruments for assessing achievement goal orientations in the educational literature differ substantively in the operationalization of the major constructs. This suggests that the instruments are assessing different constructs despite the common name. A further difference between these two instruments concerns the age group to which the instrument is traditionally applied. Although not specifically designed for different age groups, in practice the PALS have been traditionally used with elementary- and middle school- aged students while the AGQ has been used with postsecondary-aged students (Hulleman et al., 2010). Since the majority of broad-scale implementations of achievement goal orientations have been intended for kindergarten through secondary teaching environments, this study will focus on the instrument most commonly used with this group, the PALS.

Properties of the PALS scale

Dimensionality assessment of the personal goal orientation scales. Instruments created to measure goal orientations (commonly termed goal structures when applied to the classroom) at the classroom level reflect the implicit assumption that dimensionality of goal orientations remains immutable across levels of specificity. The PALS scales for assessing perceived classroom goal structures was both created and later revised to reflect the current theoretical perception of goal dimensionality at the individual level, initially encompassing solely master and performance dimensions and later modified so that the performance scale reflected the newly accepted bifurcated performance dimension (Midgley et al., 2000). When the broadly accepted division of goal orientation was solely between mastery and performance at the individual level, the classroom goal structures scale reflected the same division. Upon revision of the PALS personal goal orientation scales in 1999 to reflect an added distinction between performance approach and performance avoid, the classroom goal structures scale was modified to reflect this alteration simultaneously (Midgley et al., 2000). Due to the combination of minimal empirical research on the dimensionality of classroom goal structures and the general assumption of comparability in dimensionality across the constructs of goal orientations and structures, a review of dimensionality for goal structures begins with a review of the dimensionality of personal goal orientations. As previously discussed, the theoretical dimensionality of personal goal orientations has evolved over time, gradually including a third possible dimension created by the bifurcation of performance goals into an approach and avoid dimension (Elliot, 1999). Although the additional bifurcation of mastery goal orientation into approach and avoid dimensions has been proposed (Elliot, 1999; Pintrich, 2000a), the mastery avoid construct has failed to see widespread adoption thus far outside of research involving the AGQ, which was created by one of the researchers who originally proposed the extended theoretical framework and a colleague (Elliot & McGregor, 2001). With limited research addressing the mastery avoid construct as well as the general practice of treating the operationalization of mastery goal orientations solely in an approach fashion, the 2x2 proposed dimensionality will be excluded from further review (Elliot & Church, 1997; Hulleman et al., 2010).

Initial formulation of Goal Orientations. The notion that an individual was either mastery-oriented or performance-oriented implicitly suggested that the personal goal orientation construct be conceptualized as one-dimensional with each of the two posited goal orientations as an endpoint on the construct. Although this was a popular representation in the nascent stages of goal orientation theorizing evidenced by research designs which treated the two orientations as distinct categories based on a given student's predominant framework (e.g., Dweck, 1986), this conceptualization was explicitly rejected by both Dweck and Nicholls in favor of the idea that each of these orientations is a distinct dimension that should be treated separately (as cited in Elliot, 2005, p. 56).

Two-factor formulation. The two-factor perspective became the primary representation of personal goal orientations for the next decade, representing mastery and performance goal orientations as distinct factors with the performance goal orientations scale including items assessing both a desire to out-perform others, the future approach vector, and items assessing a desire to avoid performing worse than others, the future avoid vector (e.g., Anderman, Griesinger, & Westerfield, 1998; Anderman, Urdan, & Roeser, 2003; Pintrich, 2000b). The two-factor perspective allowed for any given individual to have quantifiable scores for both performance and mastery orientations simultaneously. Pressure to shift to a trichotomous framework began to emerge towards the close of the twentieth century, mostly deriving from work by Elliot (1999; Elliot & Church, 1997; Elliot & Harackiewicz, 1996).

In examining previous dimensionality analyses of goal orientation measures using the 2 factor structure, it is important to note that reporting practices have changed across the years. Researchers are now encouraged to report pattern loadings, structure coefficients, estimation methods and missing data handling to a greater degree than was previously emphasized (Henson & Roberts, 2006). As such, the information provided in early analyses is sparse in comparison to current practices. One confirmatory factor analysis (CFA) of the PALS personal goal orientations scale, which reported both factor pattern loadings and factor correlations, found items to load on the appropriate factors with all pattern loadings exceeding .49 and a factor correlation of -.017 between the performance and mastery scales (Jagacinski, & Duda, 2001, p. 1024); however, neither the estimation method, structure coefficients, or approach to handling missing data was mentioned, leaving future researchers unable to weigh in on the accuracy of or replicate the CFA results. Prior analyses using principal components analysis (PCA) to assess the scale properties have been performed with both orthogonal and oblique rotations. In both of these cases, only factor correlations, when pertinent, and estimated reliability coefficients (αs) were reported. Roeser, Midgley, and Urdan (1996, p. 412) used an oblique rotation and found a correlation of -.45 between mastery and performance personal goal orientation subscales. Estimated reliability coefficients calculated in these PCA analyses for the personal mastery subscale have been found to range from .71 to .81 whereas personal performance subscales have been found to range from .65 to .84 with α values generally increasing with the age of the participants (Anderman & Midgley, 1997, p. 278; Roeser, Midgley, & Urdan, 1996, p. 412-413). A study including personal goal orientations and classroom goal structures performed by Anderman, Griesinger, and Westerfield (1998, p. 86) mentioned that a factor analysis had been done, but they did not provide actual output in the paper. Overall, the primary focus of analyses seemed to be on the classic ideas of reliability and validity with minimal concern for the issues entailed in dimensionality assessment and dealing with ordinal data. Similar issues will be seen when we examine research involving the trichotomous formulation.

Trichotomous formulation. Despite the general consensus that a mastery goal orientation offered better outcomes when compared to a performance goal orientation (see, Ames, 1992; Urdan, 1997), a small subset of researchers were troubled by the seemingly inconsistent results found for performance goal orientations across studies (e.g., Elliot, 1999). Upon further analysis, Elliot and Church (1997), resurrecting early theoretical conceptualizations of motivation espoused by achievement goal orientation founders Dweck and Leggett (1988) as well as Nicholls (1984), determined that inconsistency in goal orientation correlates across research studies partially derived from a confusion in the operationalization of the performance construct across different goal orientation instruments. Prior to 1997, operationalization of the performance construct could be grouped into two main categories: 1) instruments using only positively-valenced items, and 2) instruments using both positively- and negatively-valenced items. Mastery items exhibited no such variability in valence across instruments with a consistent approach orientation being pervasive (Elliot, 1994; Elliot & Church, 1997). When prior research was parsed to divide studies into groups based upon whether the performance scale reflected solely approach motivation versus a hybrid of approach and avoidance motivation, the confusion in correlates to performance orientation was greatly alleviated (Elliot & Church, 1997). Once parsed, performance approach motivation appeared to demonstrate a relatively consistent positive relationship to academic outcomes, contrasting with the inconsistent results displayed across studies using the hybrid conception of performance motivation (Elliot & Church, 1997). Subsequently, a new theoretical model was created for goal orientations in which performance motivation was bifurcated into performance approach and performance avoid motivation while mastery orientation was retained as solely an approach construct. As such, the trichotomous conception of goal orientations became widely adopted across the predominant goal orientation instruments of the time (e.g., Middleton & Midgley, 1997; Vandewalle, 1997).

The most prominent studies evaluating the factor structure of the newly formulated trichotomous goal orientation scales used either exploratory factor analysis (EFA) with undisclosed factor extraction techniques or CFAs using maximum likelihood (ML) estimation. The earliest investigation of the factor structure for the revised scale was conducted by Middleton and Midgley (1997) and presented at the American Educational Research Association. Notably, this analysis used a scale constructed at the personal goal orientation level and resulting dimensionality results were extended by proxy to classroom and school goal structures. The initial EFA conducted on the first half of the sample was performed with an oblique rotation, resulting in support for the three-factor conceptualization. Supporting evidence consisted of adequate pattern loadings and acceptable explained variance (61.3%); however, factor correlations were not reported. A CFA using ML estimation was subsequently performed on the second half of the sample resulting in support for the three-factor model according to the goodness-of-fit index, the adjusted goodness-of-fit index (GFI), the chi-square value, and pattern loadings: factor correlations were not reported for this analysis either. When the entire sample was combined, the correlation between mastery and performance approach was .04, between mastery and performance-avoid the correlation was .01, and between performance approach and performance avoid the correlation was .56 (Middleton & Midgley, 1997, p. 10). The second large-scale study used to support the three factor structure of the personal goal orientations scale was performed by Midgley et al. (1998, p. 123) using more current diagnostic indices to assess fit (specifically: GFI, Tucker-Lewis Index [TLI], Comparative Fit Index [CFI], and the Root Mean Square of Error Approximation [RMSEA]); however, ML estimation with list wise deletion of missing data was still implemented. A correlated three-factor model showed good fit although the factor correlations were not reported. It was stated in the discussion section that “there is some overlap [between the performance-approach and performance-avoid scales] as indicated by the correlation between the two scales” although the exact correlation was not reported in the paper (Midgley et al., 1998, p. 127). These studies considered in tandem appear to have led other authors to feel that dimensionality information is well-established: it is not uncommon to find research in which an exploratory method of dimensionality assessment was performed on the PALS personal goal orientation scales with no reported results attributed to the fact that “factor structure, reliability, and validity have been previously examined” (Linnenbrink, 2005, p. 201). Overall, with the exception of one study applying a Rasch analysis to each of the three PALS scales related to personal goal orientation, much of our dimensionality information derives from the studies cited above which were performed prior to the proliferation and easy implementation of estimation methods designed to account for the ordinal nature of Likert-type data (Muis, Winne, & Edwards, 2009).

Issues in prior assessment. With the history of goal theory spanning over four decades, techniques for examining dimensionality as well as handling ordinal measures have advanced in both form and accessibility. Such progress renders prior techniques outdated despite representing best practice at the time of publication. In reviewing the PALS literature, composed of both studies with the stated purpose of establishing scale properties and studies including dimensionality assessment as a precursor to testing broader relationships, several common methods of assessing dimensionality emerge. Almost all of the studies are purportedly couched in the factor analytic framework with some studies using an exploratory approach while other studies, typically more recent, used a confirmatory approach. The word purportedly is key to this prior statement as it is not uncommon to see a study claim to be performing an EFA while implementing principal components extraction (e.g., Duda & Nicholls, 1992; Elliot & Church, 1997). The use of principal components extraction changes the nature of the technique to a dimension reduction process as embodied in PCA rather than a dimensionality assessment process as embodied in EFA (Osborne, Costello, & Kellow, 2008). One reason that this distinction arises is that PCA uses all of the variance embodied in the dataset, a practice directly contradictory to the premises embodied in Classical Test Theory (CTT), which details the composition of variance as consisting of both true and error variance. In contrast to PCA, EFA uses only shared (common) variance for extraction, presuming that unique variance represents the error component and should be parceled out. Current recommendations frown upon the use of PCA in dimensionality assessment (Osborne, Costello, & Kellow, 2008; Reise, Waller, & Comrey, 2000).

As discussed in the section concerning the trichotomous framework for personal goal orientations, CFA has gradually replaced EFA and PCA as the default dimensionality assessment technique for the PALS scales as our understanding of the structure of goal orientations has progressed. The movement to more sophisticated methods of factor analysis from PCA does represent improvement in dimensionality assessment; however the implementation of CFA still does not address many issues present when analyzing Likert-type data. Both EFA and CFA originate from the factor analytic framework which traditionally uses summary statistics such as a correlation matrix as an intermediary between the individual item level responses and the actual factor extraction procedure (Osborne, Costello, & Kellow, 2008). A factor analysis can generally proceed so long as the sample correlation matrix is provided, even without the inclusion of individual item-level responses. The implication of this process is twofold in that: 1) a loss of item-level information occurs, and 2) the factor analysis is susceptible to any shortcomings entailed in the correlation calculation process (Flora, LaBrish, & Chalmers, 2012).

In response to the first implication, full-information factor analysis, termed item-factor analysis (IFA), has become an area of burgeoning research; however, the limited-information approach to factor analysis is the most prominent method used by applied researchers (Flora, LaBrish, & Chalmers, 2012; Forero & Maydeu-Olivares, 2009). The full-information approach to factor analysis derives from MIRT; however, IFA still differs from MIRT in estimation methods and overall focus due to different overarching frameworks driving the analysis, with IFA deriving from the structural equation modeling (SEM) perspective and MIRT deriving from the IRT perspective. IFA and item response theory (IRT) using non-continuous variables has been repeatedly demonstrated as formally similar over the past three decades (e.g., Takane & De Leeuw, 1987). Despite the prevalence of IFA in the SEM literature as a source of burgeoning research directions, traditional limited-information factor analysis has remained the prominent dimensionality assessment technique for applied researchers (DeVellis, 2006; Flora, LaBrish, & Chalmers, 2012). Considering that the most prominent studies investigating the goal orientation construct have used the limited-information factor analysis (FA) approach, the remaining focus of this method review will solely consider this approach when discussing FA.

The second limitation of the FA approach is the common use of the Pearson correlation coefficient in the factor extraction process. The original development of FA occurred to aid understanding of cognitive abilities through the analysis of overall cognitive test scores, scores which were developed to be continuous. Consequently, FA emerged through the analysis of the Pearson product-moment correlation matrix, rendering the procedure optimized for analysis of linear continuous variables as well as sensitive to many of the same limitations of the Pearson product-moment correlation presented in any introductory statistics textbook. Psychology has since expanded the application of FA to Likert-type scales designed to assess attitudes and opinions. Despite applied psychological researchers' common utilization of ordinal scales, the traditional reliance on FA with ML estimation continues relatively unabated (Flora, LaBrish, & Chalmers, 2012).

FA of polytomous items using estimation procedures designed for continuous variables has known shortcomings, which are accentuated the more coarsely categorized (i.e., five or fewer observed response categories) the variable (Flora & Curran, 2004; Flora, LaBrish, & Chalmers, 2012). With the most common Likert-type scale format for the PALS goal orientation measures consisting of five response categories, these shortcomings apply to most scenarios in which FA is applied to the PALS (Midgley et al., 2000). Specific known shortcomings of using an approach designed for continuous variables with discrete variables include: attenuation of correlations leading to an underestimation of parameters, overestimation of the accuracy of these parameters in the form of deflated standard error estimates, and inflation of the chi-squared statistic (as discussed in Flora & Curran, 2004). One alternative to Pearson-based FA while remaining within the FA framework involves using the polychoric instead of the product moment correlation matrix; however, the straight substitution of the polychoric matrix has been found to produce erroneous statistics and standard error calculations partially remedied through corrected estimation methods, such as robust WLS, developed to counter these errors (Flora & Curran, 2004). A second alternative is to shift to a method designed to accommodate categorical variables such as IRT or MIRT. Shifting to this framework has the additional advantage of using full item information in calculating the model parameters and fit statistics.

Although EFA and CFA are traditionally construed as being differentiated by the intention of the analysis, as epitomized in the first terms of their names, the true difference is predicated on the assumptions ingrained within the analysis, with both EFA and CFA possessing the potential to be used in an exploratory or a confirmatory fashion[1]. Using an EFA model, the factor loading matrix is traditionally freely estimated such that each item can load on all factors; in contrast, using a CFA model generally entails constraining the estimation of factor loadings such that each item can load on only one factor (creating simple structure). Although there are further differences and various implementation possibilities of each procedure, this summarizes the primary difference in the most common implementation of these two procedures which both fall within the realm of the common factor model. Further discussion will thus discuss FA as a whole rather than particularly referencing either of these two procedures specifically.

Some historically common methods of assessing dimensionality for the PALS scales include: use of PCA to assess dimensionality under the guise of EFA (e.g., Duda & Nicholls, 1992; Elliot, A. J. & Church, 1997; Harackiewicz, Durik, Barron, Linnenbrink, & Garcia, 2008); use of PCA, labelled as a PCA, to determine dimensionality (e.g., Deemer, 2004; Elliot, A. J. & McGregor, 2001; Kaplan, Gheen, & Midgley, 2002); use of CFA with ML estimation (e.g., Middleton & Midgley, 1997; Midgley et al., 1998); and use of CFA with no named method of estimation, suggesting the default most common estimation method, ML, was used (e.g., Lau & Nie, 2008; Urdan, 2004). Although these methods represented accepted practice for assessing the psychometric qualities of a new instrument at the time, technological progress has facilitated access to and the easy implementation of more advanced psychometric techniques. Reporting

[1] Target rotation can be used to conduct an EFA in a confirmatory fashion whereas a comparison of a series of models can be used to conduct a CFA in an exploratory fashion (Flora, LaBrish, & Chalmers, 2012).

practices have also become more rigorous as the importance has been broadly recognized of reporting the estimation method, structure coefficients, specific rotation method chosen, and the data considerations that determined the number of factors to extract (Henson & Roberts, 2006). Unfortunately, many of these early articles failed to provide information concerning not only the actual rotation method used, but also whether the chosen rotation was even fundamentally oblique or orthogonal in nature. Further compounding this problem is that the dimensionality determined for personal goal orientations appears to have been presumed as applicable to goal structures at the classroom level.

Extension to classroom measures. As was mentioned in the historical review of the goal orientation construct, the development of goal structures was a response to perceived shortcomings in previous theories, such as achievement motive and attributional motivational theories. One driving issue that the goal orientation framework intended to redress was the perceived overemphasis in prior motivational theories of dispositional traits to the detriment of situational influences, limiting the efficacy of interventions (Elliot, 2005). Intriguingly enough, goal orientation research gradually shifted from a focus on experimental manipulations designed to generate particular goal structures towards an emphasis on assessing already extant personal goal orientations (e.g., Anderman, Griesinger, & Westerfield, 1998; Bong, 2004; Church, Elliot, & Gable, 2001; Duda & Nicholls, 1992; Greene, Miller, Crowson, Duke, & Akay, 2004). With personal goal orientations dominating the achievement goal literature, much of the current knowledge concerning the dimensionality and psychometric properties of achievement goals is based on assuming that the information derived from research with personal goal orientations remains true when perceived classroom goal structures are assessed. Just as intra-individual dimensionality of a construct may differ from inter-individual dimensionality, the structure of perceived personal goal orientations may differ from the structure of perceived classroom goal structures (Borsboom, 2005); thus, investigation into the dimensionality of classroom goal structures is a vital step to progressing the theory of achievement goals.

Current state of the PALS. The PALS have a long history of development and revision across the years, modifying the scale structure to reflect the most prominent theoretical framework of the time. The PALS currently includes scales designed to assess personal goal orientations and classroom goal structures, as well as a bevy of additional scales. Both the personal goal orientation and classroom goal structure scales were revised to conform to the trichotomous framework proposed in the late 1990s and rely predominantly on assessments of goal structure using the factor analytic model as optimized for continuous data (Anderman & Midgley, 2002). As discussed previously, prior studies assessing the properties of the PALS goal orientation scales may represent best practice at the time, but now appear outdated in light of current progress in psychometric methods. Only one study has been performed using item-level information to assess scale properties, a study using Rasch analysis to compare the scale functioning of three different instruments (Muis, Winne, & Edwards, 2009). Although representing progress, this study was conducted on the personal goal orientation level and analyzed each proposed dimension (or subscale) separately, potentially obfuscating any issues of shared variance across dimensions that may affect measurement in studies implementing multiple subscales. In order to understand the potential issues inherent when relying on dated methods for assessing dimensionality, the theoretical and practical frameworks of these two predominant theories, IRT and CTT, necessitates review.

Statistical Foundations

The shift in psychometric recommendations from methods relying on summary statistics as a necessary intermediary for model estimation to methods deriving from the analysis of full item information has been a relatively recent phenomena. Paradoxically, the foundational exposition of the tenants underlying these two different approaches are located in the same text published mid-twentieth century, with computational power limitations preventing the wide scale dissemination of full item information approaches until the beginning of the twenty-first century. Prior to the computational advancements which expanded the accessibility of item-based models, factor analytic methods provided a useful lens for analyzing an instrument's characteristics by providing an intermediary method that was implementable with the computational capacities available to the applied researcher of the time. Many shortcomings of the factor analytic framework as commonly applied are alleviated when full item information is used to estimate instrument properties, rendering this a useful avenue to pursue even in an instrument that is already being widely applied. One could argue that a re-examination of instrument properties with more modern psychometric techniques is especially prudent when an instrument has been widely used in the literature. A fundamental overview of these two methods will be provided to contextualize the discussion of the benefits of IRT, and MIRT, when analyzing ordinal data.

Classical Test Theory Basics

CTT can be considered the liberating theory of mental test measurement. Models had been proposed to link physical and mental stimuli, such as psychophysical models in the early nineteenth century; however, the computational difficulty involved in modeling item and person characteristics prevented the common application of these models to applied research. Instead, a simplified model, reliant on definitions to sustain the precarious framework, was formed for wide applicability (Lord & Novick, 1968/2008). This model considers the constraints of computational power while using an ideal but unattainable hypothetical scenario to model the true, observed, and error variance scores. As a result, this model offered both great advancements at the time and hindered future research with unnecessary constraints.

Basic model of true score theory. The advancement followed from a pre-dominating theory, rooted in the theory of errors, that described the observed score as a function of the "true" and error scores such that:

$$X_{ij} = T_{ij} + E_{ij}, \tag{1}$$

where X_{ij} = observed score for person i on test j; T_{ij} = trait or "true" score for person i on test j; and E_{ij} = error score for person i on test j.

This decomposition can be manipulated to provide further definitions integral to CTT. For example, error is now defined as the difference between any individual's true and observed scores:

$$E_{ij} \;=\; X_{ij} - T_{ij}\ . \tag{2}$$

When the above formula is combined with the theory of errors, the true score is defined as the expected observed score for any given individual.

Several additional assumptions were necessary to ensure that the above relationship would hold. In order for the error scores to be expected to disappear upon repeated replications, the error scores must be random and unbiased. The following assumptions were implemented to define error as truly random:

$$\mu_E = 0, \tag{3}$$

where μ_E is the average error score for a population of participants. The average error score across participants is assumed to be zero, a result we would expect if the error scores are truly random. An additional assumption to ensure unbiased error is that true and error scores are uncorrelated within a population:

$$\rho_{TE} = 0, \tag{4}$$

where ρ_{TE} is the correlation between true and error scores.

The testing instrument is also of importance and needs to have unbiased errors for the overall CTT model to hold. Therefore, the final assumption for the basic CTT model is that when two administrations of parallel tests or the same test are implemented, errors from the two test administrations are not correlated:

$$\rho_{E1E2} = 0, \tag{5}$$

where $E1$ is error from administration one and $E2$ is error from administration two.

Implicit assumptions of CTT and Psychometric Shortcomings. Traditional practice in creating summary scores for CTT measures entails a set of implicit assumptions that are rarely discussed (Streiner, 2010). The issues inherent in these implicit assumptions can be compressed into two pertinent potentially error-inducing areas: the practice of assigning numbers and the treatment of these numbers. Both of these implicit assumptions were once explicit and appear to derive from a seminal piece by Stevens (1946), popularly perceived to mean that the assignment of a number according to a rule is enough to qualify a number as approximately interval. In 1946, this practice both enabled otherwise difficult data an avenue for analysis and prevented applied researchers from being burdened with overly complicated methods of determining appropriate values for the data. Unfortunately, this practice has persisted past utility as we now have the tools for improved quantification of this data. A common instantiation of this practice is to create a standard scoring rule across all items such that each ascending answer on a Likert-type scale is worth an additional digit: strongly disagree = 0, disagree = 1, agree = 2, strongly agree = 3. Use of this scoring rule has additional assumptive implications: each threshold is presumed to be of an equal size, meaning that the amount of difference between “strongly disagree” and “disagree” is assumed to be the same as the amount of difference in opinion that it takes to decide to mark “disagree” rather than “agree” for any given item. The veracity of this assumption goes unchecked in CTT; however, polytomous IRT models are available that can assess the degree of truth in this assumption (Streiner, 2010, p.181).

If several such items exist within a measurement instrument, as is likely since more items tend to equate to increased reliability in the CTT framework, two additional assumptions are invoked. First, all items are assumed to have equal thresholds whenever the same scoring rule is applied across items indiscriminately: the movement from “disagree” to “agree” is assumed as equally difficult to make for all items (Streiner, 2010). Second, whenever items are unweighted, all items have been effectively constrained to have equal weight. In the context of a depression measure, this is equivalent to saying that a yes response to both of the following dichotomous items means the same: “my appetite is not like it used to be” and “I think I might be better off dead"[2]. In contrast, IRT includes item difficulty parameters and uses this information when formulating person parameter estimates, thus negating these assumptions.

The second issue with this assumption is the treatment of these assigned numbers as continuous. One of the most common methods of assessing construct validity, FA, derives from an early extension of the initial CTT framework. FA has traditionally depended on Pearson correlations as the driving engine to determine the associations between items and the underlying latent trait(s) (Helgado-Tello, Chacon-Moscoso, Barbero-Garcia, & Vila-Abad, 2010). Pearson correlations assume an interval scale and are known to be reduced in cases of increased sample homogeneity on the variables in question. By using a reduced scale such as the common Likert-type response scale, values are inherently restricted to a few predetermined numbers rather than spanning a truly continuous range of possibilities, leading such increased homogeneity to be much more likely. One demonstrated consequence of this issue is a systematic underestimation of the strength of inter-item relationships when using Pearson's correlation with observed ordinal items. This underestimation can potentially lead to extraction errors when using FA to determine dimensionality and underestimation factor loadings (Olsson, 1979). Alternative estimation procedures that account for the ordinal nature of the variables and use full item information are now recommended as best practice to avoid well-known estimation errors consequent to the mistreatment of ordinal variables (Muthén & Kaplan, 1985; Wirth & Edwards, 2007).

Item Response Theory Basics

Item response theory is a modeling method that imposes strong, yet testable, assumptions on the data to which it is applied (Embretson & Reise, 2000). IRT offers a platform that can assess dimensionality and obviate many of the concerns raised about mistreatment of ordinal variables by: 1) using all item response information when estimating parameters, 2) estimating

---

[2] This example is actually from an online inventory at http://www.depression-test.net/depression-questionnaire.html; furthermore, it is actually a dichotomous item and presumed to be of equal weight as used in the example.

the difficulty in moving between item response categories through use of item-level information, and 3) allow for the testing of assumptions to ascertain their tenability (Birnbaum, 1968/2008; De Ayala, 2009; Embretson & Reise, 2000). Technological innovations have allowed a method first introduced mid-twentieth century to become accessible to applied researchers in the early twenty-first (Lord, 1952; Lord & Novick, 1968/2008, Zickar & Broadfoot, 2007).

Limitations in scope. Many variants of IRT exist, such as: nonparametric IRT, the normal ogive model, and models using spline functions for the item response function (IRF) (Embretson & Reise, 2000). For the purposes of this paper, the dependent variable, ability level or theta, is assumed to be continuous; furthermore, the following assumptions are assumed to be true in describing the premises of unidimensional IRT: the instrument measures a unidimensional trait, data demonstrates good fit to the model in terms of IRF, and conditional independence holds. These premises will then be expanded upon and modified to describe the MIRT framework. IRT will serve as an explanatory and simplified platform for the understanding of the MIRT framework and assumptions.

Assumptions of IRT. Any model has assumptions; however, IRT models differ from CTT models in that the assumptions are readily testable. The following can be considered the big three assumptions of IRT.

Dimensionality. Dimensionality broadly refers to the idea that the appropriate number of dimensions has been specified to account for the latent variables that exist. The converse of this statement also deserves consideration, such that, the item responses result from only the latent dimensions specified. For the traditional IRT model, this dimensionality assumption is concretized in a unidimensionality assumption. The theoretical concept of unidimensionality is exactly as it sounds: the measurement instrument captures a particular single dimensioned construct. Theoretically, unidimensional measures offer an advantage; however, in practice, unidimensionality is an assumption never fully achieved, only approached with the necessary level of specificity being driven by the research questions(s) or hypotheses in tandem with the intended population for sampling.

With the difficulties of ever achieving perfect unidimensionality, or even defining what this may look like, the concept, when applied to intangible variables, can be understood in terms of factor structure. To attain unidimensionality, one factor must be predominant in explaining the covariance of the items. This factor may consist of either one underlying construct, or a contribution of two or more underlying constructs if they contribute equally to responses to each item on the measurement instrument (Cronbach, 1951; Rupp & Zumbo, 2004). Constructs that are theoretically multidimensional in nature should not be forced to conform to a unidimensional model; rather, a multidimensional IRT model should be chosen to model the items and responses in such a scenario. In MIRT, the assumption of unidimensionality is altered to reflect an intent to model multiple dimensions; thus, the researcher assumes that the number of dimensions specified in the MIRT model accurately reflect the true number of dimensions underlying the construct.

Conditional independence. Conditional independence, or local independence, is the second assumption of IRT models. In terms of unidimensional IRT, it is assumed that some unidimensional factor underlies the items of the measurement instrument; however, past this dimension, all item responses are assumed to be independent conditional on the latent theta, or ability, level of each individual participant. In colloquial language, if an individual takes a math test, we assume that the probability for a correct response on each item is only dependent on the test-taker's math ability and the difficulty of the item. Violations in this assumption could occur in tandem with a violation of the dimensionality assumption: if a model is incorrectly specified in terms of dimensionality, the presence of an unidentified additional latent variable could be influencing person responses. This assumption may also be violated due to question inter-relationships that are not representative of erroneous dimensionality specification.

In MIRT, the conditional independence assumption remains essentially unchanged except that the responses to an item are presumed independent of all other responses conditional on all of the individual's latent theta, or ability, levels. Thus, if one were to have a group of individuals with identical values on all presumed factors underlying item responses, the conditional item response distribution would be independent between people (de Ayala, 2009). Another way to imagine this is that after accounting for individual theta levels, the remaining variance is independent between people.

Item response function shape. Early psychometric research in the field of applied psychology focused on methods of scaling, with many competing theories emerging in the early twentieth century (e.g., Coombs, 1944; Guttman, 1944; Thurstone, 1925; Thurstone, 1928). A monumental contribution to this arena arrived in the form of early item response theory; however, the early form of this used the cumulative normal distribution, or normal ogive model, to specify the relationship between theta, or examinee proficiency level, and the probability of correct response to an item at a particular difficulty level (Lord, 1952; Lord & Novick, 1968/2008).

This innovation allowed the item to become the fungible unit of analysis, lending item properties readily examinable parameters. The normal ogive model was linked to a classical test theory basis with some additional assumptions, yet the computational difficulty prohibited an easy transition to this more complex model (Lord & Novick, 1968/2008; Tucker, 1946). Part of this difficulty was resolved through a subtle shift in the shape of the IRF from being rooted in the normal ogive model to the logistic model (Birnbaum, 1968/2008). The logistic model has proven more tractable, obviating usage of the original model, and adopting a similar IRF shape.

Both the normal ogive model and the logistic model have the same essential characteristics. Across both models, the item is most discriminating when well-targeted to the individual. Less information is acquired about individual ability level when the item is poorly targeted, as the individual is almost certainly likely to get the item correct if it is far too easy or to miss the item if it is too difficult. This is both intuitive and reflected in the sigmoidal shape of the IRF. The logistic and normal ogive models are almost identical with the most pronounced differences being at the extremes of where the probability of correct response is approaching zero or one (Embretson & Reise, 2000). With the increased tractability of the logistic function, this has been selected as the preferred method of modeling IRFs over time.

With the main difference between CTT and IRT being the direct estimation of item parameters in IRT, the IRF plays a pivotal role in this theory and deserves further discussion. As can be seen from Figure 2.2, the proposed IRF curve relates probability of correct response to theta level (for associated item parameters, see Table 2.1). Less visible is the relationship shown between item difficulty and theta level: for the 1-parameter logistic and 2-parameter logistic (PL) models, item difficulty is defined as the point of slope inflection which is also where an individual at that ability level has an even chance of success or failure on the item (see, Figure 2.3 for visual representation and Table 2.2 for associated item parameters).

The shape of the IRF is a testable assumption of IRT. Item misfit is able to be assessed, and either the items or the model can be deemed inadequate and replaced with different items or an alternative scaling method. The decision as to which route to go depends on the purpose of the instrument being developed and the philosophical perspective of the researcher: Rasch adherents are likely to keep the model and drop or alter the item to achieve better data-model fit (Andrich, 2004; Henning, 1989). So long as the data and model provide adequate fit for the purposes of the study, the increased complexity of IRT may be a worthwhile trade for the greater information garnered about instrument items (e.g., Sharkness & DeAngelo, 2011). The shape of the IRF remains a testable assumption of MIRT since the IRF is still assessable through investigating goodness-of-fit indices and residuals.

Dichotomous IRT Models

The simplest form of the logistic item response model predicts only one of two outcomes for any given achievement item: correct or incorrect. This dichotomous model represents the most parsimonious outcome structure, omitting possibilities such as partial-credit in representing item response. If the test to be considered were to be an attitude measure, the outcomes would be represented as agree or disagree, and the overall measure would theoretically assess level of endorsement overall for each examinee.

Although there are logistic models for polytomous response structures such as a 4-point Likert-type response scales ranging from ‘strongly disagree’ to ‘strongly agree,’ the following models will be restricted to dichotomous to serve as an introduction to the typical item-parameters considered in applied IRT. The models discussed will be limited to the 1-PL to provide a basic framework and the 2-PL which frees the discrimination parameter, allowing discrimination to vary across items.

1-Parameter logistic model. The 1-PL model is the simplest logistic model out of all the IRT models, estimating only person ability, item difficulty, and a common item discrimination. As in all IRT models, the IRF specifies the relation between person ability, item difficulty, and probability of correct response. The basic 1-PL model can be expressed with the following equation:

$$P\left(x = 1 \middle| \theta_j, \beta_i, \alpha\right) = \frac{e^{\alpha(\theta_j - \beta_i)}}{1 + e^{\alpha(\theta_j - \beta_i)}}, \tag{6}$$

where θ_j is the ability level for person j, β_i is the difficulty for item i, and α represents the item discrimination. It may be noted that alpha involves no subscripts: this is representative of a particular characteristic of the 1-PL model, specifically that one item discrimination is estimated across all items and constrained to constancy.

It can be seen from this model that the probability of success exhibits a non-linear relationship to person ability (also see the IRFs in Figure 2.2 or 2.3). This non-linear relationship reflects that the most informative items for any given person are those targeted to their ability level (i.e., where the item is most discriminating based on the slope of the IRF). Targeted items are also preferred in CTT, but the relationship between the individual and the item is subverted to the relationship between the individual and the average item difficulty on the test (Tucker, 1946).

Because differentiated item discrimination and guessing parameters are not included in this model, certain additional assumptions apply. The discrimination is constrained to an estimated value held constant across items. Looking at the IRFs in Figure 2.2, one can see that the slope of the IRFs remains constant across all items: this will differentiate the 1-PL from the 2-PL model. The 2-PL model relaxes the assumption of equivalent item discriminations while maintaining the assumption that guessing is not a pertinent parameter.

2-Parameter logistic model. The 2-PL model frees item discriminations to vary, but in doing so adds additional parameters to be estimated. It is important to always assess whether the sample size is large enough to produce stability in estimates as model complexity increases. The 2-PL model can be represented with this equation:

$$P\left(x = 1 \middle| \theta_j, \beta_i, \alpha_i\right) = \frac{e^{\alpha_i(\theta_j - \beta_i)}}{1 + e^{\alpha_i(\theta_j - \beta_i)}}\,, \tag{7}$$

where θ_j is the ability level for person *j*, β_i is the difficulty for item i, and α_i represents the discrimination for item *i*. The addition of the subscript to item discrimination reflects the separate estimation of discrimination for each item. Using this model, the item discrimination relates to the item-total biserial correlation used to represent item discrimination in traditional CTT analysis (Embretson & Reise, 2000; Lord & Novick, 1968/2008).

This model allows all parameters to be estimated except a guessing parameter which is constrained to equal zero. By allowing item discrimination to differ, the slopes of the IRFs vary as can be seen in Figure 2.3. If enough variability exists, IRFs can cross; furthermore, with this model, the total sum score is not enough to determine the person ability estimate. Instead a weighted sum score is used with more discriminating items given greater weight when estimating theta (Embretson & Reise, 2000).

Although an examination of dichotomous IRT models can be instructive in understanding the general nature of the IRT framework, these dichotomous models are not able to remedy many of the perceived shortcomings of treating Likert-type data as continuous. To remedy the issues of assigning numbers according to a rule and treating the resultant data as approximately interval, polytomous IRT models need to be implemented.

Polytomous IRT Models

Polytomous IRT models represent an extension of the simple dichotomous IRT models described above. These models were designed to handle situations, such as in attitudinal measures with Likert-type rating scales, where the responses cannot be easily divided into two categories. Unlike in the traditional treatment of ordered responses in CTT (i.e., each category given a subsequent integer value), IRT polytomous models have the capacity to estimate the amount of the latent trait needed to cross a threshold between response categories: an example would be the threshold between strongly disagree and disagree. Figure 2.4 shows an example of a five-point Likert-type scale with four item thresholds. If the number of response option is "*m*", the number of thresholds will always be "*m*-1".

Although there has been a proliferation of polytomous IRT models since the first extension of the dichotomous model into the world of polytomous responses by Samejima in 1969, only the model selected as optimal for the purpose of analyzing the PALS scales will be considered at this time (Bock, 1997). The graded response (GR) model allows item discriminations and threshold values to vary across items, but does require the thresholds to be ordered within an item, which will be expanded upon further after an introduction of the GR model equation (Embretson & Reise, 2000).

Graded Response Model. The GR model can be seen as an extension of the traditional 2PL model:

$$P_{it}(\theta) = \frac{e^{\alpha_i(\theta_j - \beta_{it})}}{1+e^{\alpha_i(\theta_j - \beta_{it})}}, \tag{8}$$

$t_i = 1,\ldots,k_i - 1$; k_i = the number of response categories; and $k_i - 1$ = the number of thresholds, or m_i.

Equation 8 represents the probability that a person will respond in a particular category, such as strongly disagree rather than all higher categories (i.e., disagree, agree, strongly agree), for a particular item as conditioned on theta, or ability level. As can be seen by examining the subscript of the discrimination parameter, item slope is allowed to vary across items, thus being an extension of the 2-PL dichotomous model; however, threshold slopes are restricted to constancy within an item. This constancy constraint is a result of this model being what Thissen and Steinberg (1986) term a "difference model." Item response thresholds are also constrained to be ordinal in nature, meaning that endorsing disagree would be presumed to reflect less of the latent trait than would be reflected by endorsing agree on the same item. One benefit of this model is that items with different response categories pose no problem: question one can ask participants to choose an answer on a 4-point scale (i.e., Strongly Disagree, Disagree, Agree, Strongly Agree), whereas question two can ask participants to select an answer from a six-point scale (i.e., Strongly Disagree, Disagree, Slightly Disagree, Slightly Agree, Agree, Strongly Agree) with no model-based problems.

This model estimates response probabilities in a two-stage process. The first stage involves Equation 9 and consists of comparing two groups of item responses to determine where the thresholds fall on the latent continuum (i.e., SD vs. D, A, SA; SD, D vs. A, SA; SD, D, A, vs. SA). The second stage involves estimating the actual probability of response in each category. This process is the basis of the term "difference" model as it involves taking the full distribution and subtracting off portions until all probabilities are estimated. Mathematically, probabilities are calculated as follows assuming a Likert-type agreement scale:

$$P_{i0} = 1.0 - P^{*}_{i1}(\theta) \qquad \text{For strongly disagree}$$

$$P_{i1}(\theta) = P^{*}_{i1}(\theta) - P^{*}_{i2}(\theta) \qquad \text{For disagree}$$

$$P_{i2}(\theta) = P^{*}_{i2}(\theta) - P^{*}_{i3}(\theta) \qquad \text{For agree}$$

$$P_{i3}(\theta) = P^{*}_{i3}(\theta) - 0 \qquad \text{For strongly agree.} \qquad (9)$$

Bifactor Model

The bifactor model is a unique model in its ability to estimate both a general factor which parcels out variance common across all items and multiple orthogonal specific factors representing common variance among subgroups of specified items simultaneously, allowing for comparison of the respective importance of these factors (Reise, Moore, & Haviland, 2010; Toland, et al., 2017). This allows all factors, general and specific, to exist on the same conceptual level rather than imposing a hierarchy among them as is common in higher-order factor models. The result of this is the ability to easily compare variance explained by the general factor to the overall variance explained by the model, allowing a quantitative understanding of the relative contribution of each specified factor in explaining the full estimated variance (Reise et al., 2010; Toland et al., 2017).

Following the mathematical notation set out in Toland et al. (2017), I will present three equations to elucidate the bifactor model using MIRT and the GR model. The first step necessary for easily transforming the previously presented unidimensional GR model to the multidimensional GR model is rewriting the unidimensional GR model in slope-intercept form as follows (Toland et al., 2017):

$$\mathrm{P}(Y_{ji} \geq k|\theta_j, c_{ik}, a_i = \frac{e^{c_{ik} + a_i\theta_j}}{1+e^{c_{ik} + a_i\theta_j}}, \tag{10}$$

where Y_{ji} = person j's response to item i, c_{ik} = -$\alpha_i\beta_{ik}$, and a_i = α_i. In this representation, the IRT item discrimination parameter and category-threshold parameter are converted to the more traditional CFA slope and category-intercept form. The main benefit of this representation in Equation 10 is that it allows a direct extension to the multidimensional GR model as follows (Toland et al., 2017):

$$\mathrm{P}(Y_{ji} \geq k|\boldsymbol{\theta}, c_{ik}, a_i = \frac{e^{c_{ik} + a_i{}^1\, \theta_j{}^1 + \cdots + a_i{}^M \theta_i{}^M}}{1+e^{c_{ik} + a_i{}^1\, \theta_j{}^1 + \cdots + a_i{}^M \theta_i{}^M}}. \tag{11}$$

The alterations to the multidimensional GR model in Equation 11 reflects the shift from the probability of a response in a given response category or higher depending on a single dimension, dimension 1, to the probability of an item response in a given response category or higher depending on dimensions 1 through M, with a common estimated category-intercept across all dimensions. Each item has a distinct estimated slope for any given dimension which reflects the relationship between the latent trait level on a given dimension and the probability of an individual responding in a given category for that item; thus, it is logical that each person has an estimated latent trait value for each given dimension from 1 to M as well. When using the confirmatory bifactor GRM, a response to any individual item will be influenced by item slopes, intercept, and person latent trait estimates from one general dimension and one specific dimension, orthogonally estimated, with all other dimensions being constrained to loadings of zero. To conceptualize this, the multidimensional GRM will be rewritten in terms of a single specific trait (*S*) and a single general trait (*G*).

$$\mathrm{P}(Y_{ji} \geq k|\boldsymbol{\theta}, c_{ik}, a_i = \frac{e^{c_{ik}+\, a_i{}^S\, \theta_j{}^S + \cdots + a_i{}^G \theta_i{}^G}}{1+e^{c_{ik}+\, a_i{}^S\, \theta_j{}^S + \cdots + a_i{}^G \theta_i{}^G}}\,, \tag{12}$$

where P is the probability of person *j* providing a response equal or greater than *k* on item *i* given a person's estimated location ($\boldsymbol{\theta}$) on both the general and one specific trait, category *k*'s item-intercept (c_{ik}), and the conditional item slope parameters on both the general (a_i^G) and one specific trait (a_i^S). One difference between the MIRT and the bifactor models is that the MIRT model allows an item to contribute to the measurement of only one latent trait whereas the bifactor model allows for each item to contribute to both the general and one orthogonal specific trait. This difference allows these models to be used in a complementary fashion to understand the functioning of a scale.

Plausible Models

Based on the extant literature, seven different models as depicted in Figures 2.5 to 2.11 could be used to understand the internal structure of the PALS Classroom Goal Structures scales.

These seven models fall into three different model categories: unidimensional IRT, MIRT, and bifactor.

A unidimensional IRT model will model all item responses as indicative of a single latent trait. In this particular case, a unidimensional model is theoretically untenable in light of current research; however, two purposes are served by including this model: 1) this model is the simplest representation of the data and thus deserves to be ruled out as a potential model to satisfy Occam's razor, and 2) this model will serve as a baseline comparison for model fit when examining the more complex models detailed below. A single unidimensional IRT model will be considered:

- in the standard unidimensional model, all item responses are estimated to reflect a single latent trait representing overall perceived classroom motivation (see Figure 2.5 – Model A)

MIRT models allow for multiple different correlated dimensions to be modeled simultaneously, offering a chance to compare different representations of classroom goal structures using full item information and treating the data as ordinal. Three MIRT models will be considered:

- A correlated traits model with all performance items, both approach and avoid, constrained to load on a performance dimension while all mastery items are constrained to load on a mastery dimension. The mastery and performance dimensions are allowed to covary. This represents a historical perspective on goal structures (see Figure 2.6 – Model B);
- A correlated traits model with three estimated dimensions, performance approach, performance avoid, and mastery, which are allowed to covary. This model represents the most theoretically plausible model and the intended structure of the scale (see Figure 2.7 – Model C);

- A correlated traits model with all approach items, both performance approach and mastery, constrained to load on an approach dimension while performance avoid items are constrained to load on an avoid dimension. The approach and avoid dimensions are allowed to covary. This model represents an extension of A. J. Elliot's (1999) work emphasizing the historical utility of parsing constructs into approach and avoid components (see Figure 2.8 – Model D)..

The bifactor model offers a unique opportunity to parse variance at the item level rather than between constructs, allowing for supplemental information about the extent to which items are representing the general versus the specific dimensions (all orthogonal). When considering a scale designed to assess three dimensions, the desired result would be minimal variance being attributed to the general dimension and the preponderance of the variance being attributed to the specific dimensions. Three bifactor models will be considered:

- a bifactor model which complements Model B, having one general factor and two specific factors, performance (including both approach and avoid) and mastery (See Figure 2.9 – Model E);
- a bifactor model which complements Model C, having one general factor and three specific factors (See Figure 2.10 – Model F);
- a bifactor model which complements Model D, having one general factor and two specific factors, approach (including both mastery and performance approach) and avoid (see Figure 2.11 – Model G).

Study Purpose

The intent of this study is to provide an updated understanding of one scale, the PALS perception of classroom goals scale, from a commonly implemented motivational instrument. Since the initial inception of the goal orientation construct, the theoretical basis has been altered, leading to further modification of the initial scale. Despite these changes and subsequent revisions to the scale, all formal analysis has been performed from a CTT perspective (Midgley et al., 2000). The current accessibility of MIRT and bifactor analysis allow for the unique opportunity to examine the dimensionality and item properties of the PALS goal orientation in greater depth and with tools that appropriately treat ordinal responses as polytomous items to minimize distortion of scale structure during the estimation process. As such, the historically proposed potential construct structures will be examined and compared in order to enhance knowledge of the properties of this scale. The bifactor model will be utilized to better parse any sources of multidimensionality, potentially elucidating the latent trait dimensionality further.

In the current study, the following research questions will be addressed:

(1) What are the psychometric properties of the PALS classroom goal structure scale?

    a. Which multidimensional representation of the PALS classroom goal structure scale provides the best fit and most interpretable solution?

    b. Have items been appropriately parsed onto scales such that items are assessing the intended construct?

(2) Based on the model determined above, what are the psychometric properties of each latent dimension considered as a unitary scale?

    a. Which items are most and least informative for each scale?

    b. How does construct coverage compare to the theta range for each scale?

    c. What is the level of measurement precision for each scale?

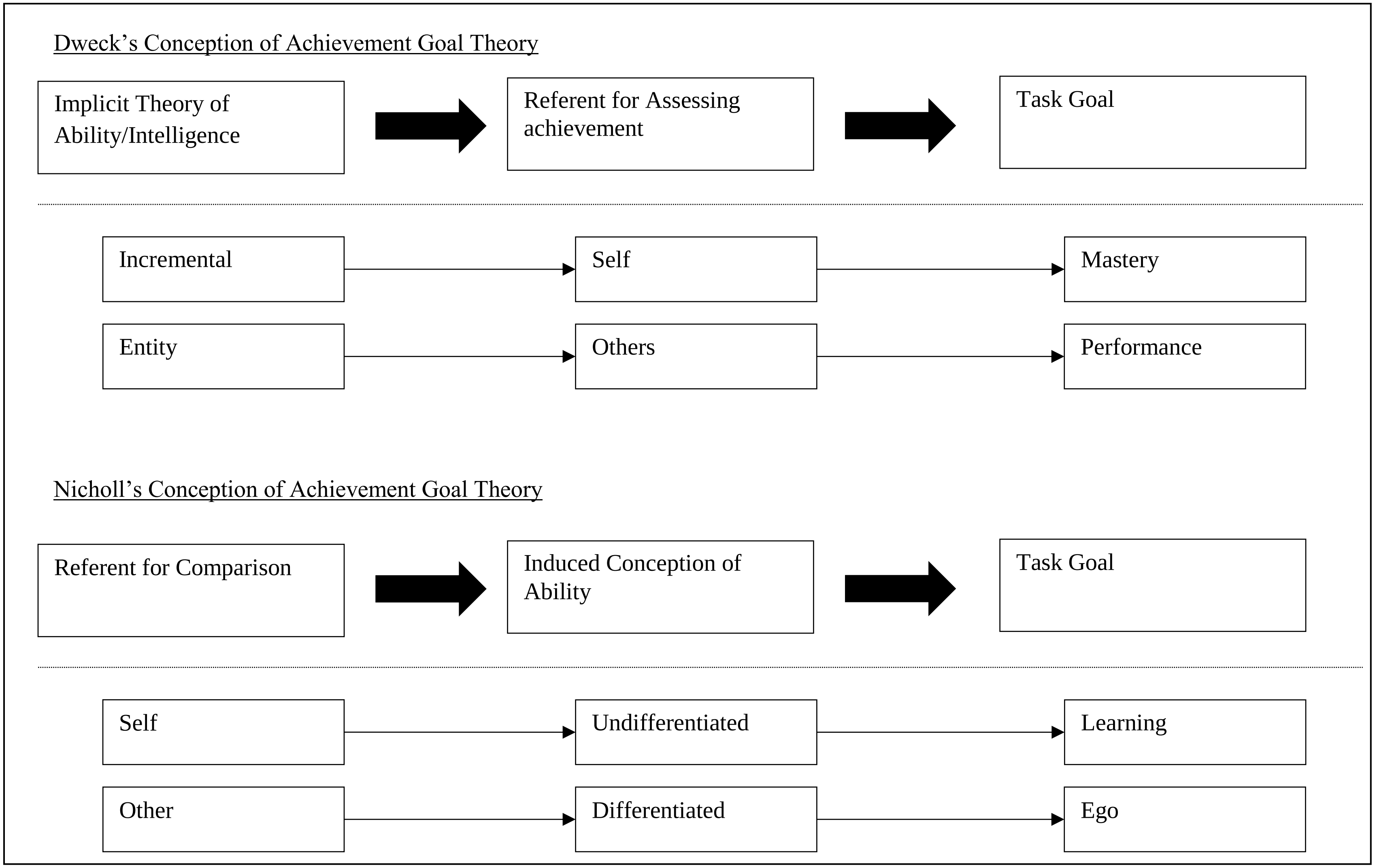

Figure 2.1. Conceptions of goal orientation precursors contrasted between models proposed by Dweck and Elliot.

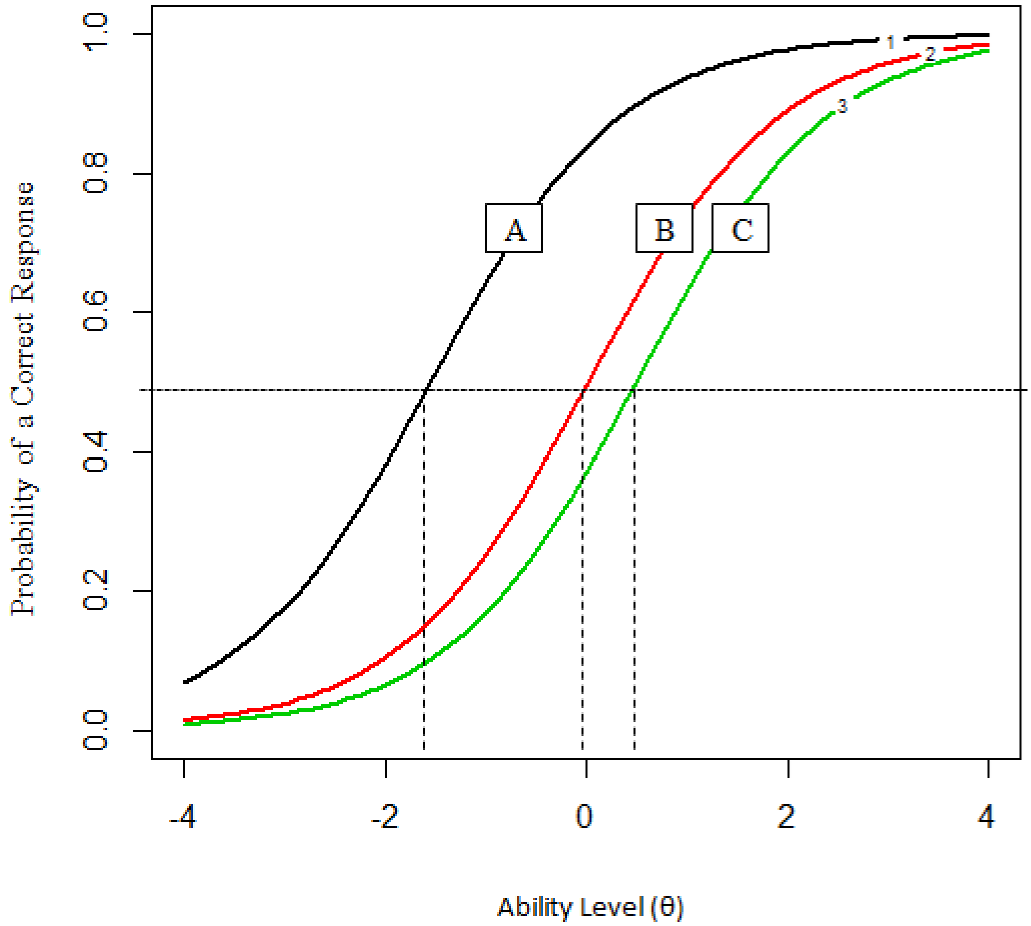

Figure 2.2. 1-PL item response functions showing the relation of person ability and item difficulty to probability of correct response.

Table 2.1 Item parameters associated with Figure 2.2

Item (i)	Discrimination (α)	Difficulty (β)
A	1.06	-1.90
B	1.06	-.025
C	1.06	0.20

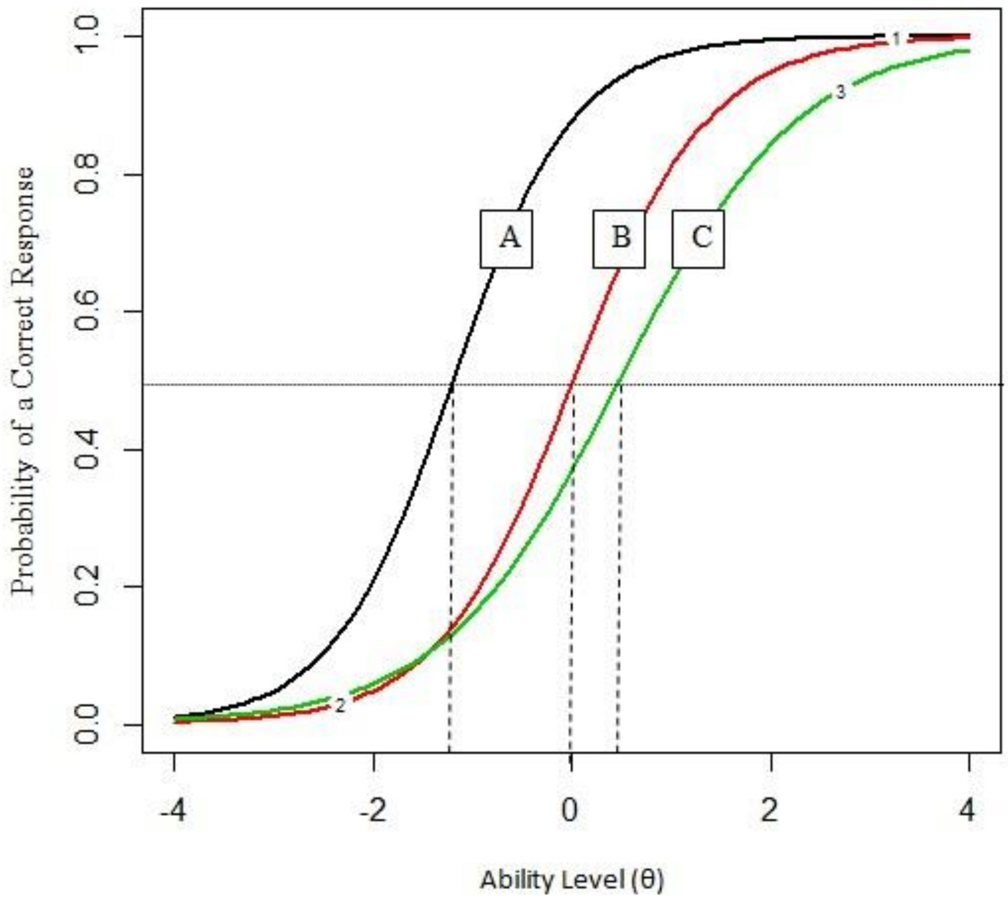

Figure 2.3. 2-PL item response functions showing the relation of person ability and item difficulty to probability of correct response.

Table 2.2 Item parameters associated with Figure 2.3

Item (i)	Discrimination (α)	Difficulty (β)
A	1.64	-1.19
B	1.47	0.02
C	1.65	0.49

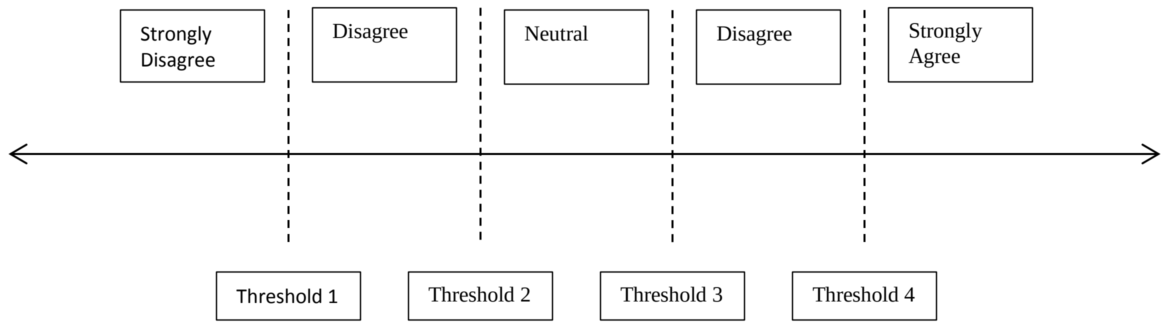

Figure 2.4. Example of a likert-type response scale with five response choices and the 4 item thresholds between them.

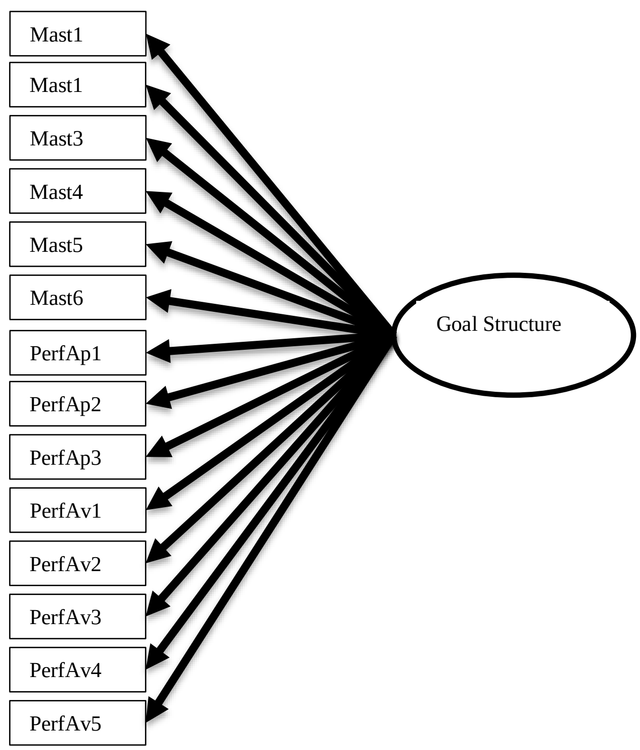

Figure 2.5. Model A, a unidimensional representation of the perceived classroom goal structures PALS scale.

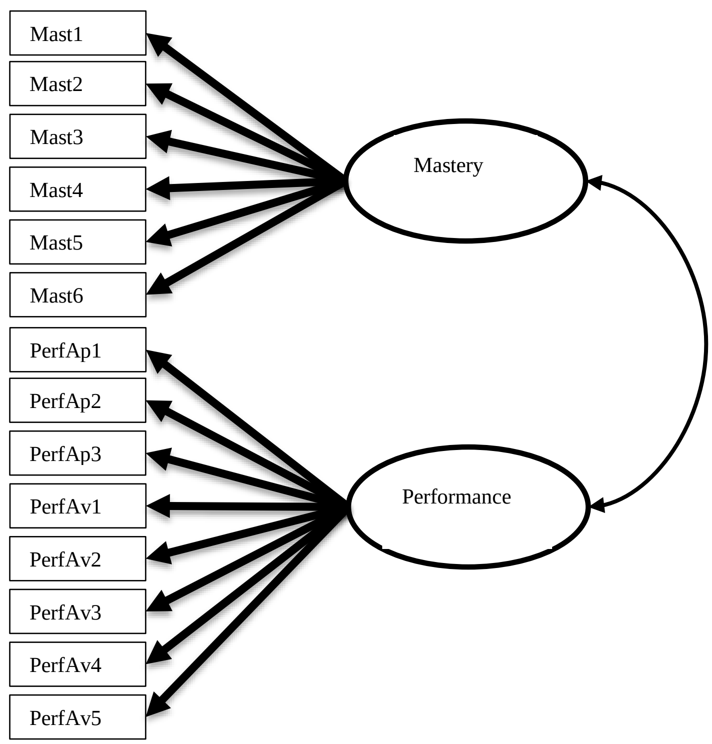

Figure 2.6. Model B, a correlated traits model with two estimated latent traits (mastery, performance).

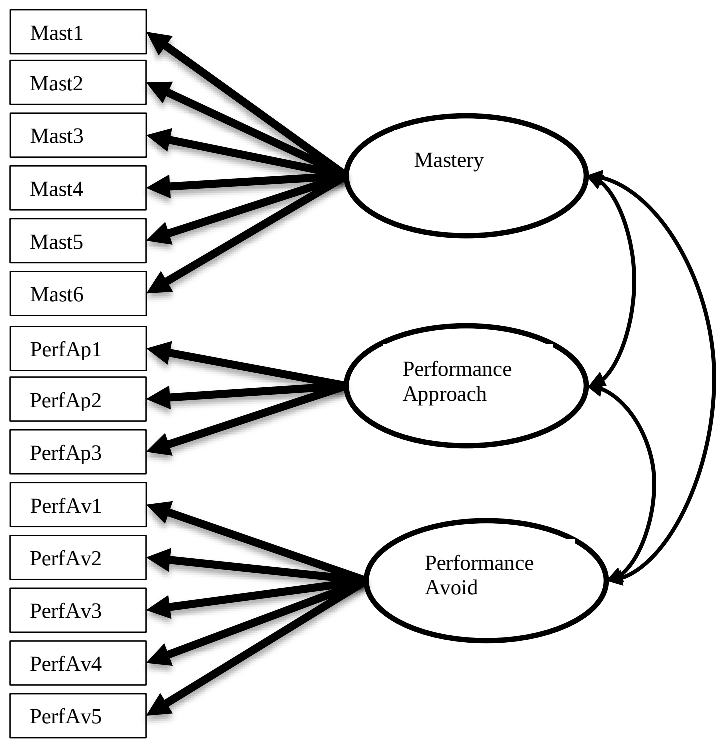

Figure 2.7. Model C, a correlated traits model with three estimated latent traits (mastery, performance approach, performance avoid).

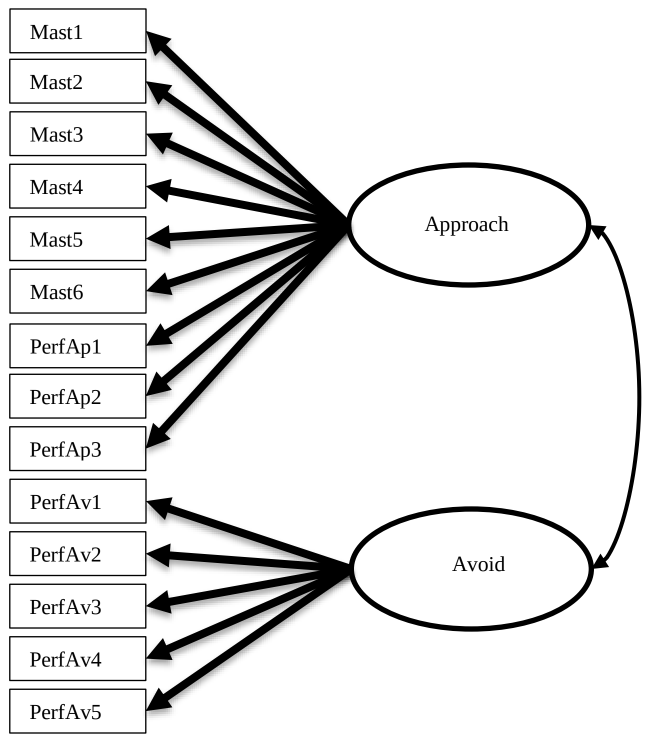

Figure 2.8. Model D, a correlated traits model with two estimated latent traits (approach, avoid).

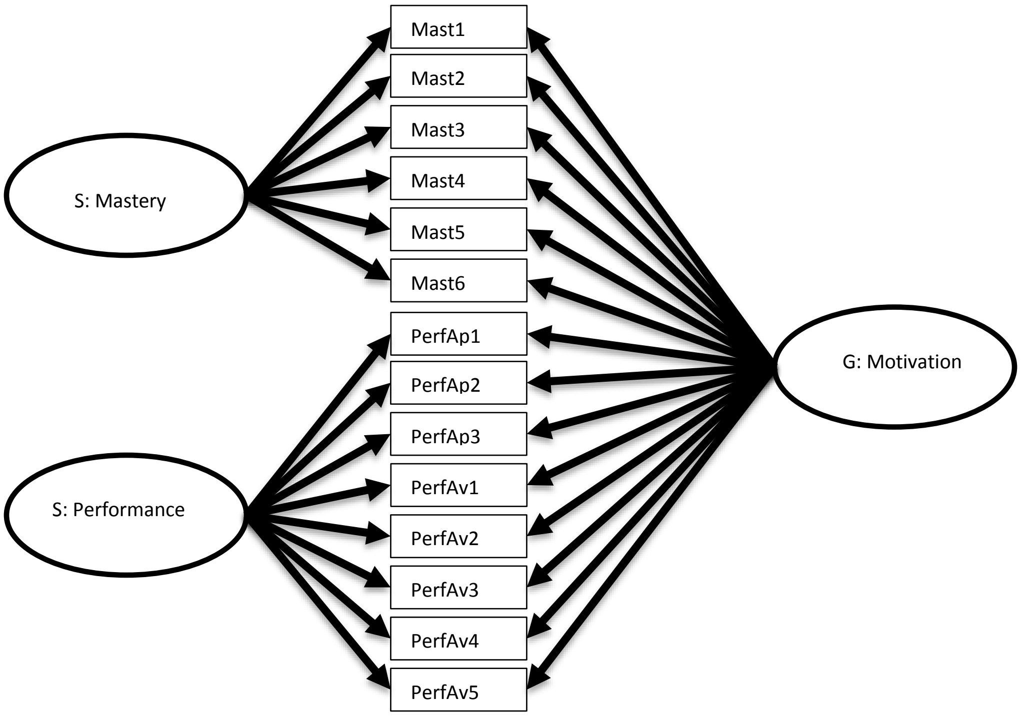

Figure 2.9. Model E, a bifactor model with two specific factors (mastery, performance).

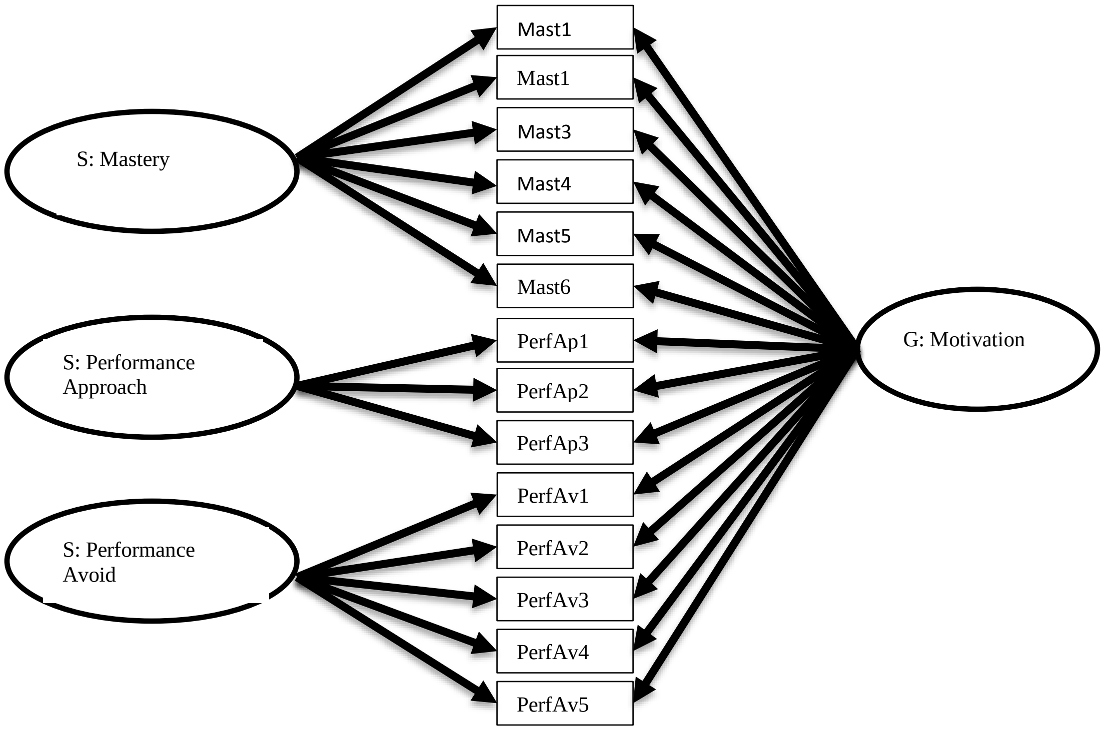

Figure 2.10. Model F, a bifactor model with three specific dimensions (mastery, performance approach, performance avoid).

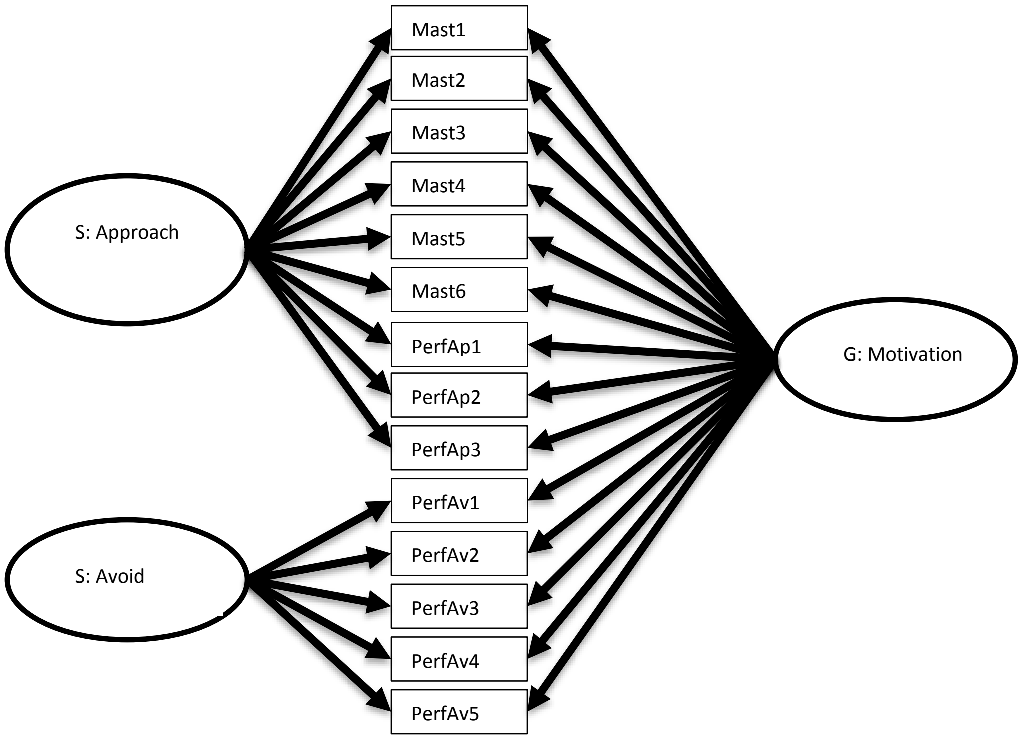

Figure 2.11. Model G, a bifactor model with two specific dimensions (approach, avoid).

Chapter 3: Method

Sample

Students, enrolled in grades 9-12, were recruited from three different southeastern public high schools in the United States. Samples will be divided by school membership with one sample, Sample A (the initial sample), used to compare models representing different representations of dimensionality while the second sample, Sample B (the validation sample), will be used to confirm the validity of the results found using Sample A. Sample A will only include one school while Sample B will include two smaller schools.

Students in both samples were predominantly white with almost 60% of each sample self-identifying as being of white ethnicity. The students in Sample A were 58.5% white, 12.7% African-American, 3.4% Hispanic, 7% Asian, 6.3% identifying as some other ethnicity, and 12.2% choosing not to respond. Students in Sample B were 59.9% white, 23.5% African-American, 5.5% Hispanic, 2.4% Asian, 5.9% identifying as some other ethnicity, and 2.8% choosing not to respond. Gender was relatively evenly split with Sample A consisting of 43.2% female students and 41.7% male students with 15.1% choosing not to identify gender: sample B consisted of 50.1% female students and 44.8% male students with 5.1% choosing not to identify gender. Both samples had more respondents deriving from the lower high school grades. Sample A included 28.6% ninth grade students, 28% tenth grade students, 22% eleventh grade students, 10.9% twelfth grade students and 10.5% of students who choose not to indicate grade while Sample B included 35.1% ninth grade students, 26.5% tenth grade students, 22.3% eleventh grade students, 15.9% twelfth grade students and <1% of students who choose not to indicate grade (see Table 3.1).

Instrumentation and Procedure

Perceived classroom goal structures were assessed using a pencil-and-paper survey with the PALS Perception of Classroom Goal Structures scales (Midgley et al., 2000). The subsection querying perceived classroom goal structures included: 6 items designed to assess perception of mastery classroom goal orientation, 3 items to assess performance approach classroom goal orientation, and 5 items to assess performance avoid classroom goal orientation. This subsection was preceded with instructions requesting the student to reflect upon either their English or Social Studies class and to answer the following questions to the extent that the items reflected their perceptions in this class (for specific item stems, see Table 3.2. All questions used the same 5-point Likert-type response format, where 1 = *not at all true*, 2 = *a little bit true*, 3 = *somewhat true*, 4 = *quite a bit true*, 5 = *very true*.

Data Analysis Plan

Given the proposed structure of the PALS goal structures scales, the unidimensionality assumption is untenable; thus, the following analyses will be couched in MIRT using the GR model to account for the categorical nature of the data[3].

Due to the assumption that the multidimensional structure of the data conform to the multidimensional structure specified in the MIRT model as well as the broader familiarity of factor analysis for the general reader, many researchers choose to assess dimensionality from a factor analytic framework prior to beginning analysis from the IRT framework (e.g., Edelen & Reeve, 2007). Although this approach may provide an accessible entrance to the topic for those unfamiliar with IRT, the IRT and in particular MIRT approach was selected for two main reasons: 1) IRT provides more information by estimating not only an item's discrimination parameter, similar to the factor loading assessed in the traditional framework, but also an intercept for each item in the form of a difficulty estimate (or step thresholds in the polytomous models); 2) although the factor analytic approach and the IRT approach will often lead to similar conclusions, factor analysis has been shown to obfuscate the true nature of the data in certain

[3] The choice between polytomous models has been largely found to produce similar results, rendering the selection of a model a relatively trivial decision so long as it fits the data structure (Edelen & Reeve, 2007)

cases (De Ayala, 2009). Considering these twofold reasons to select the IRT framework over the FA approach, the assumption of appropriate dimensionality will be examined through the traditional item indices provided in a MIRT analysis.

The following analysis plan is most easily understood as a multiple stage process. Stage 1 consists of a comparison Models A through D, the correlated trait models, using a MIRT framework and a common polytomous model, the GR model, to determine the best representation of dimensionality. Stage 2 consists of assessing additional sources of evidence concerning the nature of any multidimensionality of the data by reviewing Models E through G, the bifactor models. This stage is particularly important as the PALS scales have frequently been parsed in research and treated as unique subscale scores (e.g., Deemer, 2004; Greene et al., 2004; Gutman, 2006). The results from stages 1 and 2 will lead to the selection of the most appropriate dimensional representation of the data. Stage 3 will involve parsing the selected optimal MIRT model into unidimensional IRT models by identified constructs to provide a more in-depth examination of item and subscale properties. Although the bifactor model is intended to be a complementary model to the MIRT models, if the bifactor model is deemed the most successful model, stage 3 will consist of a further breakdown of the item parameters. Stage 4 is essentially a repetition of the above steps using only the optimally identified model from stage 1 and 2 on a sample B for validation of the results.

The following analysis plan for stage 1 borrows heavily from the framework proposed by De Ayala (2009), Embretson and Reise (2000), and Toland (2014), with the focus being on assessing both conformance of the data to the model assumptions and item-level fit before proceeding to a review of model-fit indicators for each of the seven proposed models. The first component of stage one involves checking the empirical data in regards to the assumptions of functional form and local, or conditional, independence (LI). Due to the complexity of examining IRFs plotted in a multidimensional space, LI will be used as an indirect indicator of functional form when reviewing MIRT models as a violation of LI would indicate violations of the functional form assumption. Once the final model is selected and examined using unidimensional IRT for each dimension identified in the MIRT model, a graphical examination of predicted IRFs will be used to further assess the tenability of the functional form assumption.

LI is the assumption that responses to items are only reflective of the latent trait variables included within a given model. With this particular scale, we are assuming that responses solely indicate each individual's latent traits concerning classroom goal structures; thus, the scale has been written in such a way that neither possible latent traits such as reading ability nor other items included within the classroom goal structure scale are influencing an individual's response to any given item. If LI is an issue, other item parameters may be distorted. Particular distortions include: inflated slopes (suggesting better discrimination of items than is actually present), reduced standard error estimates at both the item and the latent trait level (suggesting greater accuracy than is actually present), and more homogeneous item thresholds. Considered in tandem, these distortions tend to artificially inflate the apparent information provided by the scale (De Ayala, 2009; Edelen & Reeve, 2007; Toland, 2014). LI will be assessed using the standardized local dependency (LD) χ^2 statistic with absolute values greater than 10 deemed large and absolute values that range between 5 and 10 considered worthy of further examination through item content analysis (Cai, du Toit, & Thissen, 2011, p. 77). Particular attention will be paid to any item which shows recurrent LD issues across multiple item pairs. If an item is flagged as potentially problematic when considering both context and statistical indicators, the appropriateness of item removal will be assessed through a comparison of models estimated both with and without the potentially problematic item. Model comparison will be focused on potential changes in slope and threshold estimates in terms of either magnitude or pattern between the full and partial models. If LD is found to be an issue for an item when considering all information in combination, the item will be removed from the final calibration for that model.

After satisfactory item-level fit has been established, model-level fit will be examined. The primary indicator used to examine item-level fit will be the generalized S-χ^2 item-fit statistic, which is based upon a comparison of the empirical response data to the expected response frequencies if the data were to conform to the defined model. This polytomous generalization depends on empirical response frequencies and is subsequently sensitive to data sparseness across the response category matrix. Significant p-values indicate departure of an item from appropriately fitting the specified model. Due to the large sample size and usage of multiple significance tests, items will be examined for adequate fit through evaluation at the 1% significance level as suggested by Toland (2014). Information garnered from the generalized S-χ^2 item-fit statistic will be considered in combination with previously mentioned item information to determine the appropriate course of action. The previously described actions for evaluating both item fit and model fit will be performed for each of Models A through D prior to moving on to cross-model comparison of model-fit.

Subsequent to completing all item-fit and model-level fit decisions for each of the four correlated trait models, model-fit will be examined using several complementary model-fit statistics, specifically the Akaike Information Criteria (AIC), the Bayesian Information Criteria (BIC), and the C_2 limited information goodness-of-fit statistic with its associated RMSEA index (Cai & Monroe, 2014). The AIC, BIC, and RMSEA[4] are commonly used statistics from the FA framework and maintain the same meaning in the context of assessing fit in MIRT models. The AIC and BIC are both measures of relative model fit which function as indicators of misfit for a data to a model; thus, lower values indicate better fit. These two statistics do differ in penalizing for increased model complexity with the BIC introducing greater cost for increased complexity. A less familiar statistic when approaching from a FA framework is the C_2 statistic which assesses goodness of fit and can be considered similar to the traditional χ^2 statistic (Toland et al., 2017). The C_2 statistic has an associated RMSEA index which can hold any value between 0 and 1 with

[4] From the traditional FA framework, the RMSEA would be associated with the χ^2 statistic. It is of note that the RMSEA mentioned here is associated with the C_2 statistic.

lower values indicating greater fit. For models which can be considered nested, the likelihood ratio test (LRT) based on the change in the -2 log likelihood value between two nested models will also be used as a source of evidence. The LRT indicates whether the additional complexity of the full model provides substantially better fit when compared to the reduced model. The end result of the first stage of the data analysis will be a decision as to which MIRT model best represents the empirical data in terms of dimensionality.

The next stage, Stage 2, of the data analysis involves accruing additional evidence concerning the dimensionality of the PALS classroom goal orientation scales by modelling the data with Models E through G, the complementary bifactor models. The initial stages of assessing model data fit follow the same pattern as analysis outlined in stages 1 and 2 for the correlated trait MIRT models. One important caveat is that the bifactor model should not be considered nested; thus, the LRT will not be used in model comparison. After assessing item-level fit and model-level fit using the previously delineated procedure, marginal bifactor slopes will be calculated and compared following the procedure and formulas detailed by Stucky and Edelen (2015).

An additional source of information, crucial to understanding the dimensional structure in a bifactor model is explained common variance (ECV), a calculation that determines the proportion of common variance explained by either a general or specific factor: this can be converted to a percentage by multiplying the value by 100 (Reise et al., 2010; ten Berge & Socan, 2004; Toland et al., 2017). The ECV can be calculated for the general dimension as an index of unidimensionality or for the specific dimensions to determine the uniqueness of each dimension (Stucky & Edelen, 2015). An item-based ECV (IECV) can be calculated for both the general and specific dimensions to parse out common variance at the item level: this process allows determination of how representative an item is of the general and specific factors (see, Stucky & Edelen, 2015, Equation 9.17). The extension of the IECV to assess representation of the specific trait was introduced by Toland et al. (2017, Equation 16).

Stage 3 of the data analysis involves taking the decided upon polytomous MIRT model and breaking it down into separate unidimensional IRT models to allow for the examination of item properties by dimension. As prior research has traditionally used each defined subscale to construct separate construct scores for inclusion in analysis, an understanding of the items properties by subscale may prove fruitful to future researchers. The two part process outlined below will be repeated for each identified subscale from the prior MIRT analysis.

Using the IRT model, item properties such as slope and threshold values will be examined. Item slope parameters will first be checked for validity by assessing whether the range of item slopes includes extreme values that may indicate artificial inflation. Slope parameters will also be used to determine which items are the most and least informative in relation to the construct of interest. The range of item thresholds as well as the size of the gap between item thresholds will be used to determine both whether the scale is functioning as desired and if the thresholds appear to represent a valuable shift along the theta continuum.

After item properties have been reviewed, information functions at both the item and the test level will be used to further examine the properties of the individual items and the subscale as a whole in its current iteration. Item information functions (IIFs) will be used to examine the range for which each item provides information across the scale and the location at which the item is most appropriate at discriminating person theta level. The ability to overlay these IIFs in a single chart allows for the discernment of potentially redundant items as well as the identification of the most and least informative items on the subscale. The test information function (TIF) is a summation of all of the IIFs for the scale and allows for the assessment of scale precision at different points along the latent continuum. The TIF will be used for identification of possible gaps in information across the theta continuum and the range of ability levels which for which the theta scores are most precisely estimated.

The final stage of the data analysis involves using an independent sample to validate the prior data conclusions. As such the polytomous MIRT model determined as the optimal model in stages 1 and 2 of the data analysis will be applied to a different sample and assessed for fit using the same procedures outlined in stages 1 and 3. To assess whether the dimensionality is appropriate, the empirical data will be examined with the following process: 1) the data will be evaluated for conformance to MIRT assumptions with the standardized LD χ^2 statistic and a review of the item thresholds; 2) item-model fit will be assessed using the generalized S-χ^2 item-fit statistic; and 3) model-fit statistics will be reviewed for plausibility. The MIRT model will then be broken down into distinct subscales based on dimensionality and reanalyzed by separate unidimensional IRT models to assess item properties. In assessing item properties, slope estimates, item thresholds, IIFs, and the TIF will be used in tandem to suggest future directions for each subscale's possible revision as well as to provide a description of each subscale's current state.

Table 3.1.
Sample Characteristics Split by Sample.

Item	Category	Sample A		Sample B	
		Student 'N'	Percentage	Student 'N'	Percentage
Ethnicity:					
	African American	149	12.7	414	23.5
	Asian	82	7.0	43	2.4
	Hispanic	40	3.4	97	5.5
	White	688	58.5	1057	59.9
	Other	74	6.3	105	5.9
	Missing	144	12.2	49	2.8
	Total:	**1177**	**100.0**	**1765**	**100.0**
Gender:					
	Female	508	43.2	885	50.1
	Male	491	41.7	790	44.8
	Missing	178	15.1	90	5.1
	Total:	**1177**	**100.0**	**1765**	**100.0**
School Grade:					
	Nine	337	28.6	619	35.1
	Ten	329	28.0	467	26.5
	Eleven	259	22.0	393	22.3
	Twelve	128	10.9	281	15.9
	Missing	124	10.5	5	0.3
	Total:	**1177**	**100.0**	**1765**	**100.0**

Table 3.2.
Response Counts and Percentages for the 14-Item Classroom Goal Structures PALS Scale Split by School

			Response Label					
Item	Sample	Statement	Not at all true	A little bit true	Somewhat true	Quite a bit true	Very true	Missing
Mastery1	ALL	Trying hard is important.	135 (4.6%)	177 (6.0%)	465 (15.8%)	804 (27.3%)	1269 (43.1%)	92 (3.1%)
	A		53 (4.5%)	64 (5.4%)	189 (16.1%)	326 (27.7%)	523 (44.4%)	22 (1.9%)
	B		82 (4.6%)	113 (6.4%)	276 (15.6%)	478 (27.1%)	746 (42.3%)	70 (4.0%)
Mastery2	ALL	How much you improve is really important.	109 (3.7%)	260 (8.8%)	622 (21.1%)	899 (30.6%)	959 (32.6%)	93 (3.2%)
	A		42 (3.6%)	107 (9.1%)	261 (22.2%)	364 (30.9%)	382 (32.5%)	21 (1.8%)
	B		67 (3.8%)	153 (8.7%)	361 (20.5%)	535 (30.3%)	577 (32.7%)	72 (4.1%)
Mastery3	ALL	Really understanding the material is the main goal.	127 (4.3%)	235 (8.0%)	594 (20.2%)	874 (29.7%)	1004 (34.1%)	108 (3.7%)
	A		47 (4.0%)	103 (8.8%)	243 (20.6%)	359 (30.5%)	400 (34.0%)	25 (2.1%)
	B		80 (4.5%)	132 (7.5%)	351 (19.9%)	515 (29.2%)	604 (34.2%)	83 (4.7%)
Mastery4	ALL	It’s important to understand the work, not just memorize it.	142 (4.8%)	249 (8.5%)	612 (20.8%)	870 (29.6%)	949 (32.3%)	120 (4.1%)
	A		59 (5.0%)	100 (8.5%)	265 (22.5%)	366 (31.1%)	360 (30.6%)	27 (2.3%)
	B		83 (4.7%)	149 (8.4%)	347 (19.7%)	504 (28.6%)	589 (33.4%)	93 (5.3%)
Mastery5	ALL	Learning new ideas and concepts is very important.	123 (4.2%)	262 (8.9%)	719 (24.4%)	912 (31.0%)	804 (27.3%)	122 (4.1%)
	A		47 (4.0%)	97 (8.2%)	272 (23.1%)	389 (33.1%)	343 (29.1%)	29 (2.5%)
	B		76 (4.3%)	165 (9.3%)	447 (25.3%)	523 (29.6%)	461 (26.1%)	93 (5.3%)
Mastery6	ALL	It’s OK to make mistakes so long as you are learning.	180 (6.1%)	231 (7.9%)	639 (21.7%)	837 (28.5%)	921 (31.3%)	134 (4.6%)
	A		73 (6.2%)	85 (7.2%)	269 (22.9%)	338 (28.7%)	379 (32.2%)	33 (2.8%)
	B		107 (6.1%)	146 (8.3%)	370 (21.0%)	499 (28.3%)	542 (30.7%)	101 (5.7%)

Table 3.2 continued

Item	Sample	Statement	Response Label: Not at all true	A little bit true	Somewhat true	Quite a bit true	Very true	Missing
Papp1	ALL	Getting good grades is the main goal.	73 (2.5%)	188 (6.4%)	439 (14.9%)	780 (26.5%)	1372 (46.6%)	90 (3.1%)
	A		23 (2.0%)	74 (6.3%)	177 (15.0%)	317 (26.9%)	566 (48.1%)	20 (1.7%)
	B		50 (2.8%)	114 (6.5%)	262 (14.8%)	463 (26.2%)	806 (45.7%)	70 (4.0%)
Papp2	ALL	Getting right answers is very important.	101 (3.4%)	222 (7.5%)	634 (21.5%)	956 (32.5%)	929 (31.6%)	100 (3.4%)
	A		38 (3.2%)	87 (7.4%)	242 (20.6%)	391 (33.2%)	395 (33.6%)	24 (2.0%)
	B		63 (3.6%)	135 (7.6%)	392 (22.2%)	565 (32.0%)	534 (30.3%)	76 (4.3%)
Papp3	ALL	It's important to get high scores on tests.	107 (3.6%)	185 (6.3%)	502 (17.1%)	871 (29.6%)	1170 (39.8%)	107 (3.6%)
	A		41 (3.5%)	69 (5.9%)	199 (16.9%)	377 (32.0%)	467 (39.7%)	24 (2.0%)
	B		66 (3.7%)	116 (6.6%)	303 (17.2%)	494 (28.0%)	703 (39.8%)	83 (4.7%)
Pavoid1	ALL	Showing others that you are not bad at class work is really important.	381 (13.0%)	497 (16.9%)	859 (29.2%)	628 (21.3%)	432 (14.7%)	145 (4.9%)
	A		145 (12.3%)	187 (15.9%)	361 (30.7%)	273 (23.2%)	175 (14.9%)	36 (3.1%)
	B		236 (13.4%)	310 (17.6%)	498 (28.2%)	355 (20.1%)	257 (14.6%)	109 (6.2%)
Pavoid2	ALL	It's important that you don't make mistakes in front of everyone.	512 (17.4%)	577 (19.6%)	787 (26.8%)	522 (17.7%)	419 (14.2%)	125 (4.2%)
	A		183 (15.5%)	227 (19.3%)	326 (27.7%)	232 (19.7%)	178 (15.1%)	31 (2.6%)
	B		329 (18.6%)	350 (19.8%)	461 (26.1%)	290 (16.4%)	241 (13.7%)	94 (5.3%)
Pavoid3	ALL	It's important not to do worse than other students.	353 (12.0%)	506 (17.2%)	868 (29.5%)	665 (22.6%)	416 (14.1%)	134 (4.6%)
	A		129 (11.0%)	204 (17.3%)	355 (30.2%)	273 (23.2%)	180 (15.3%)	36 (3.1%)
	B		224 (12.7%)	302 (17.1%)	513 (29.1%)	392 (22.2%)	236 (13.4%)	98 (5.6%)
Pavoid4	ALL	It's very important not to look dumb.	434 (14.8%)	525 (17.8%)	730 (24.8%)	565 (19.2%)	561 (19.1%)	127 (4.3%)
	A		167 (14.2%)	211 (17.9%)	317 (26.9%)	235 (20.0%)	216 (18.4%)	31 (2.6%)
	B		267 (15.1%)	314 (17.8%)	413 (23.4%)	330 (18.7%)	345 (19.5%)	96 (5.4%)
Pavoid5	ALL	One of the main goals is to avoid looking like you can't do the work.	533 (18.1%)	536 (18.2%)	794 (27.0%)	535 (18.2%)	413 (14.0%)	131 (4.5%)
	A		199 (16.9%)	214 (18.2%)	345 (29.3%)	224 (19.0%)	163 (13.8%)	32 (2.7%)
	B		334 (18.9%)	322 (18.2%)	449 (25.4%)	311 (17.6%)	250 (14.2%)	99 (5.6%)

Chapter 4: Results

The overall intent of this analysis is to expand our understanding of the PALS classroom goal structure questionnaire using MIRT models to explore the relationship between items and dimensionality. This analysis will provide information about: 1) the best correlated trait model for representing this data with these samples, 2) sources of multidimensionality for this questionnaire, 3) the level of information and precision provided by each subscales, and 4) potential shortcomings of this scale.

Initial Analyses: Sample A

The correlated traits models, when compared to the bifactor models, provide optimal representation of the classroom goal structure construct as per prior research and development of the questionnaire; thus, the correlated traits models are the main focus of this analysis with the bifactor models providing additional insight by estimating the general and specific factors simultaneously. This simultaneous estimation process provides insight into item function by allowing for within item multidimensionality rather than forcing multidimensionality to exist between items. For a refresher on the models being examined, see Figures 2.5 through 2.11.

Correlated Traits Models (Models A through D)

Analysis begins by fitting sample A to a unidimensional IRT model, Model A, to provide a baseline. Although the unidimensional model is not anticipated to provide adequate representation of the data, it has utility as a comparison model. Models B and D provide alternative two dimensional correlated trait models with Model B providing a historical representation of the data, where both performance avoid and performance approach items are grouped onto a single performance dimension, while Model D offers an alternative representation, where mastery and performance approach are grouped onto a single approach dimension. Model C is the currently accepted three dimensional correlated traits model with separate dimensions for mastery, performance approach, and performance avoid.

Assumptions: Evaluation of conditional independence. Conditional independence, a critical assumption in MIRT, is based on the notion that all of the covariance between item responses is explained by the underlying modelled dimensionality; thus, an individual's response to an item is due solely to his or her ability on the underlying dimension being measured. This assumption was assessed for veracity using Chen and Thissen's (1997) standardized LD χ^2 statistic. Absolute values of this statistic greater than 10 are considered large and potentially problematic (see, Steinberg & Thissen, 2013, p. 342). As is common on self-report questionnaires, negative LD is a prevalent issue across the unidimensional and correlated trait models (see Table 4.1 for a count of positive and negative LD pairs flagged as large by each model). None of these four models had fewer than 50% of the item pairs flagged for negative LD issues. Model C had the fewest negative LD pairs flagged with a count of 48 (57%) problematic negative LD pairs whereas Model B had the largest number of negative LD pairs with 55 (65%) of the item pairs being flagged for problematic negative LD.

Positive LD presents a greater threat in terms of parameter estimate inflation (slope values), inflation in information (reliability) estimates, and deflated estimates of measurement precision around scores, which can lead a researcher to believe that more information is present in a score than actually exists. Such inflation can artificially suggest improved accuracy and lead to the discovery of relationships that are not truly present (see Toland, 2014). Because LD can reflect un-modeled dimensions, inappropriately specified dimensionality, item order effects or item wording effects, or item content overlap, examining the pattern of positive LD across models can provide insight into model misspecification (see Table 4.2). Although models A, C, and D seem to demonstrate item-based issues when looking at the pattern of positive LD, model B presents a different story. Model B, which combines performance approach and performance avoid items into one dimension, evidences a positive LD problem across the entire block of performance approach items: this pattern lends support to the theoretical decision to parse performance into the two dimensions of performance approach and performance avoid. The disappearance of this block pattern in models C and D suggests that model B is misspecified (see Table 4.3).

Although, items could be dropped at this time in an attempt to eliminate positive LD issues, the intent of this initial analysis is to gain insight on the appropriate dimensionality of the instrument as a whole. Retaining all items allows for both a better understanding of the instrument as traditionally used in applied research as well as a comparison between nested models using the -2LL (Deviance) using a likelihood ration test (LRT) that is chi-square distributed with df being the difference in number of parameters between two nested models. When unidimensional subscales are being analyzed, reduced item models to eliminate positive LD issues will be considered.

Assumptions: Evaluation of functional form. The functional form assumption of MIRT models follows the same logic as the functional form assumption in unidimensional IRT models – the data is assumed to match the specified form. Due to the complexity of the multidimensional surface of item response functions, the veracity of this assumption is assessed at a unidimensional level and assumed to hold in the multidimensional format when dimensions are appropriately specified. To assess this assumption the initial and validation samples are used in tandem, with the initial sample being used for parameter estimation (calibration) and the validation sample being used to provide empirical item response data. In a way, the second sample is acting as a cross-validation sample using sample 1's item parameters. The empirical data is then able to be compared to the estimated parameters, using MODFIT, to heuristically assess whether the graded response logistic model accurately models the empirical response pattern (Stark, 2001). An examination of each of the three subscales individually suggested that the functional form assumption is tenable.

Comparing models: Assessing item-level model-data fit. Item-level model-data fit is an index of accurate response prediction at the item level by comparing observed item responses to modelled predictions (Orlando & Thissen, 2003). This was assessed using the Orlando-

Thissen-Bjorner item-fit S-χ^2 statistic as implemented for polytomous item response data in flexMIRT® version 3.0.3 (Cai, 2015). This statistic was evaluated at the .01 level to correct for potential Type I error inflation. As can be seen in Table 4.1 none of the MIRT models performed well with the unidimensional Model A showing only 1 item as demonstrating acceptable fit whereas models C and D performed the best with 3 of the 14 items demonstrating acceptable fit. Model B performed only slightly better than model A with 2 of the 14 items showing adequate fit to the model. Because the accuracy of the item fit statistic depends upon correct specification of the latent model in terms of dimensionality, more emphasis should be ascribed to models C and D where the pattern of positive LD did not suggest model misspecification (Toland et al., 2017).

Comparing models: Global model-data fit. Despite the issues in item level fit, global measures of model-data fit can be cautiously used as an additional source of information when comparing alternative models (see Table 4.1 for fit indices across models). As expected, Model A demonstrates the worst global fit across all indices with the largest AIC, BIC, and -2LL values. Model A also shows poor fit when using the RMSEA associated with the C_2 statistic (RMSEA = .18). Although demonstrating poor global fit, Model A is still useful as a baseline model for comparison.

When assessing Models B through D, it is important to note that Model A is nested within both Models B and D and Models B and D are both nested within Model C; however, Models B and D cannot be placed in any sort of nested structure to each other. Thus, the deviance statistic (-2LL) will only be pertinent for model comparison in some cases (comparing models A to B to C or models A to D to C). Comparisons between Models B and D will be rooted in the AIC, BIC and the RMSEA associated with the C_2 statistic. When using these comparative indices, smaller values demonstrate improved global model-data fit.

The additional dimension specified for Model B shows significant improvement in global fit over the unidimensional Model A based on a difference in each models deviance statistic or what is known as a LRT, $\chi^2(1) = 503.58$, $p < .001$. Model B continues to show evidence of improved fit when comparing the AIC, BIC, and RMSEA to Model A; however, Model B still demonstrated less than adequate global fit with an RMSEA value of .17.

Model D presents an alternative two dimensional correlated traits model in which mastery and performance approach were combined into a single approach dimension while performance avoid was specified as a separate dimension. Similar to Model B, Model D shows improvement in global fit over Model A based on both the LRT ($\chi^2[1] = 1394.58$, $p < .001$) and all other fit indices (AIC, BIC, RMSEA). Although Model D and Model B cannot be compared using a LRT, a comparison of the global fit indices suggests that Model D demonstrates improved fit over Model B; furthermore, Model D has an associated RMSEA value of .09, indicating marginal global fit of this model.

Model C, the three factor correlated traits model that represents the current theoretical perspective on the PALS classroom goal orientation scale, shows the best fit out of all tested correlated traits models across all comparative indicators. The addition of a separate dimension for the performance approach items significantly improves global fit when compared to both Model B ($\chi^2[2] = 1102.54$, $p < .001$) and Model D ($\chi^2[2] = 211.54$, $p < .001$). Model C also demonstrates the smallest values for the AIC, BIC, and RMSEA (.07). Despite relatively poor item-model fit across all three correlated trait models, Model C demonstrates the best overall fit to the data out of the estimated correlated trait models.

Bifactor Models (Models E through G)

The bifactor models provide a complement to the correlated traits models: by estimating orthogonal general and specific dimensions situated at equivalent levels (non-hierarchical), the bifactor model allows multidimensional relationships to be parsed at the item level. Thus the multidimensionality is able to be examined on an item by item basis rather than as an overall relationship between the latent dimensions. Models E through G provide insight into the driving impetus behind the latent trait correlations shown in models B through D.

As expected, the bifactor models demonstrate improved fit over the correlated traits models as evidenced by fewer positive LD issues flagged, reduced deviance statistic, reduced AIC and BIC, and improved model fit based on the RMSEA value. The bifactor is designed to account for LD by allowing for the modeling of within-item multidimensionality, so improved fit is expected (Toland et al., 2017). When ECV is computed and discussed within the context of the bifactor models, it will become apparent that despite the bifactor models evidencing improved fit, the existence of an underlying general dimension for these subscales is not tenable.

In-depth glance at our comparison model: Model A. When discussing the bifactor model, a strong focus is on the way that within-item multidimensionality is being parsed and the effect of this multidimensionality on slope parameters. A comparison of the general dimension to the specific dimensions within the bifactor model is augmented by a comparison to a model where one single underlying dimension is being fit to the entire set of items with no option of parsing the multidimensionality, Model A. An in-depth discussion of Model A slopes will facilitate further comparisons with the subsequent bifactor models.

As discussed previously, Model A exhibited issues with positive LD and poor overall fit; however, the pattern in the positive LD provides insight into the reason for the poor model-data fit. Positive LD existed for every item pair within the performance avoidance subset of items, which means that more covariance exists within this subset of items than is predicted by the unidimensional model (see Table 4.4). Next, slope parameters are examined to determine possible dominant item subsets. Using Table 4.4, it is apparent that both mastery (average slope = 2.00) and performance approach (average slope = 1.98) items dominate the factor structure when a unidimensional solution is applied. The sharp drop in slopes for the performance avoidance items (average slope = 0.77) when combined with the positive LD issues suggest that a distinct construct is being tapped with the avoidance items. This information will be used as a comparative base when examining the bifactor models.

Model E: Bifactor with specific mastery and specific performance. Model E is the complement to Model B, representing the historic proposed dimensionality of the goal orientation scales where mastery is contrasted with performance (inclusive of both an approach and avoid dimension). Model E is able to provide insight into the workings of Model B through comparison with the unidimensional model (see Table 4.5 for Model B parameter estimates, Table 4.6 for Model B factor intercorrelations, and Table 4.11 for Model E parameter estimates). For a comparison of these models to be valid, marginal slopes are calculated which remove the conditional relationship between the general and specific factors, placing the slopes of the bifactor onto a unidimensional metric (Stucky & Edelen, 2015).

The unidimensional model (Model A) suggested that if there were a single dimension representing all items, that dimension would be driven by mastery items closely followed by performance approach items; in contrast, the bifactor Model E flips this relationship, estimating that the general dimension is driven primarily by performance approach items, with performance approach item 1 being the primary representative (‘Getting good grades is the main goal’). The domination of the general dimension by the performance approach items left minimal unique variance available for the performance approach items to load onto the specific dimension (or for the specific factor loadings).

Despite minimal variance being attributed to the specific dimension (*a^{S2}, Table 4.11) for the performance approach items, the specific performance dimension explains enough of the item response variance to be considered a distinct dimension (due to the contribution of the performance avoid items). The ECV for the performance dimension has a value of .28 suggesting that this is a unique construct for this sample and these items. Further look at the IECVs for the performance dimension suggest that this dimension is primarily representing the performance avoid items with IECV values ranging from .66 to .92. In contrast, the performance approach items have IECV values that range from .00 to .02 on this same dimension. A further comparison of the IECVs on the general and specific dimension shows that IECVs are unilaterally higher on the specific dimension for the avoidance items, adding credence to the idea that the avoidance items represent a unique dimension.

Using Model E to supplement Model B, stronger conclusions can be made about the misspecification of Model B. Examining the pattern of large positive LD values evidenced in Model B, the two-factor correlated traits model appears to be showing an excess of unmodeled covariation between the performance approach and mastery items as well as within the performance approach block (see Table 4.2). The performance dimension in Model B was being driven by the performance avoid items. This pattern led to decreased slopes for the performance approach items when these items were forced onto the same dimension as the performance avoid items as well as an inflation of the correlation between the two dimensions (.60) due to the unaccounted for relationship between mastery and performance approach items. From these two models (Models B and E), it can be concluded that performance approach should not be grouped with performance avoid onto a single dimension.

Model F: Bifactor with specific mastery, performance approach, and performance avoid. Model F is the complement to Model C, representing the current proposed dimensionality of the goal orientation scales where mastery, performance approach, and performance avoid are each considered unique dimensions (for Model C parameters, see Table 4.7). When examining the output of Model C, the estimated correlations between the latent variables show a story which will be emphasized by the results of the bifactor analysis in Model F. Model C shows a strong correlation between Mastery and Performance Approach with an estimated value of .82 (see Table 4.8). Performance Avoid shows a weaker relationship to the other subscales with a correlation of .34 with Mastery and .49 with Performance Approach. Positive LD in Model C did not show a strong pattern, suggesting no obvious dimensional misspecification.

Examining marginal slopes on the general factor for complementary bifactor Model F, performance approach items still appear as the driving force (for Model F parameters, see Table 4.12). However, this model allows each subscale a distinct specific factor, consequently altering the patterns of marginal slopes in the model. Separating the performance approach and avoid items into two distinct specific factors allows the performance approach items to parse within item variance in a more theoretically appropriate manner. Consequently, we see some of the mastery items joining the performance approach items in driving the general factor. Most notable amongst the mastery items is Mastery1, "Trying hard is important." Nevertheless, the majority of the mastery items show relatively large marginal slopes (loadings) on the specific trait suggesting these items form a unique construct ($ECV_{mastery}$ = .12).

As mentioned, the performance approach items tend to dominate the general trait, leaving little unique (or residual) variance for estimating the specific performance approach dimension. Steinberg and Thissen (2013, p. 362) caution that encountering a specific factor with reduced loadings should not prevent the conclusion that a scale is best represented by a multidimensional model. They also state that the loading pattern should not be overinterpreted substantively. Other samples may show a different set of items driving the general trait.

The performance avoid items again show evidence of functioning as a distinct construct. This is expected based on the low estimated latent trait correlation between performance avoid and the other latent dimensions in Model C (see Table 4.8). Unlike the mastery and performance approach items, the performance avoid items consistently show larger marginal slopes (loadings) on the specific dimension ($*a^{S3}$ or $*S_3$) compared to the general dimension ($*a^{G}$ or *G, see Table 4.12). With an overall ECV of .26, Performance Avoid can be considered a unique construct, distinct from Mastery and Performance Approach. Overall, bifactor Model F, estimating three specific dimensions, supports a three-factor correlated traits model.

Model G: Bifactor with specific approach and avoid. The two-factor correlated traits Model D is complemented by the bifactor Model G, which groups performance approach items with mastery items onto a single specific approach dimension and performance avoid items onto a separate specific dimension (for Model D parameters, see Table 4.9; for Model D inter-factor correlations, see Table 4.10; for Model G parameters, see Table 4.13).

The pattern of this bifactor model is remarkably similar to the pattern seen when looking at bifactor Model E. Marginal slopes (loadings) on the general dimension ($*a^G$ or $*G$, respectively) indicate the driving impetus of the general dimension is the performance approach items. Similar to the other bifactor models, the minimal available unique variance for the performance approach items after estimation of the general dimension renders small marginal slopes on the specific approach factor ($*a^{S1}$ or $*S_1$). However, the specific approach dimension remains representative of a unique construct with an ECV value of .17. Examination of IECVs suggests that this specific dimension is largely representative of the mastery items (IECVs range from .12 to .58) rather than the performance approach items (IECVs range from .00 to .01). The performance avoid items again show larger marginal slopes on the specific avoid dimension when compared to the general dimension, suggesting that these items form a unique construct (ECV_{pavoid} = .25). Despite the apparent within item multidimensionality present for the mastery items, this model also suggests that a three-factor correlated traits model is the best representation of this data.

Unidimensional Subscales

With both the bifactor and correlated traits models suggesting the supremacy of the three-factor correlated traits model for this data, each subscale will be broken down separately using unidimensional IRT models. This empirical result, combined with theoretical support, also serves to satisfy the first critical assumption of unidimensional IRT, appropriate dimensionality. In order to evaluate the subscales using IRT, the assumptions of functional form and local independence will also need to be satisfied.

The functional form assumption was also assessed prior to the confirmatory MIRT models using MODFIT, a program designed to compare empirical response patterns from a validation sample to the parameters estimated from an initial sample heuristically (Stark, 2001). As mentioned previously, this heuristic approach suggested that item response patterns were well represented by the graded response logistic representation used for each of the three subscales. A third assumption, local independence will be assessed for each scale separately.

The following subscale breakdowns will start by assessing the remaining assumption of the IRT framework, namely local independence, continue with analyzing model-data fit through both item- and model-level fit, and finish by providing a more in-depth look at the utility of the scale for producing informative scores across the continuum. Examining the utility of the scale will involve a discussion of factor loadings and slopes as well as an overview of the most informative range of the overall subscale. This breakdown will allow for a better understanding of how each individual subscale functions as well as suggestions for revision in the discussion section.

Mastery subscale of the PALS classroom goal structures scale. The mastery subscale consists of 6 items with response choices on a 5-point likert-type scale ranging from ‘not at all true of me’ to ‘very true of me.’ When initially fitting these 6 items to a unidimensional IRT model, mastery1 and mastery2 show evidence of positive local dependence using the Chen and Thissen’s (1997) standardized localized dependency (LD) χ^2 statistic with a cutoff value of greater than $|10|$ as suggested by Cai, du Toit, and Thissen (2011, p.77). Results need careful consideration as this statistic has been shown to be overly sensitive to LD violations when the number of items is small, which has been previously defined as 10 items (Chen & Thissen, 1997, p. 288). To assess the magnitude of this issue with this small number of items, a sensitivity calibration was conducted of the subscale. The process for a sensitivity calibration is as follows: 1) drop mastery1 and calibrate the subscale, 2) drop mastery2 and calibrate the subscale, 3) compare the effect of this calibration on item parameters from the full set of items. This sensitivity calibration process supported the idea that the positive LD was distorting item parameters as evidenced by a shift in parameter magnitude and order between the reduced models. A closer examination of the slope parameters suggest that the presence of mastery1 is the source of the distortion[5]. Considering that mastery1 was shown to be strongly related to the performance approach items across all bifactor analyses and is also estimated as the least discriminating item in this unidimensional calibration, mastery1 was selected for removal. All following analyses will use the reduced 5-item mastery subscale with mastery1 removed (for parameters of the reduced mastery subscale, see Table 4.14).

Item level fit used the same S-χ^2 statistic as was used when analyzing the MIRT models (Orlando & Thissen, 2000, 2003). An important caveat to this statistic is that simulation studies have been performed under dichotomous response conditions with individual item misfit (Orlando & Thissen, 2003) and under polytomous conditions where the entire dataset is generated under a different IRT model (Kang & Chen, 2008), but analysis of individual item misfit under polytomous data conditions has not yet been studied. Although the S-χ^2 statistic indicated that none of the items fit, the heuristic approach of comparing empirical response data to item parameters as implemented in MODFIT showed minimal aberrations. With the understudied performance of the statistical approach in detecting individual item fit under polytomous data conditions, the heuristic approach was relied upon in this instance and item level fit was deemed acceptable. Model level fit was assessed using the C_2 statistic with more emphasis placed on the associated RMSEA value due to the assumption in the C_2 statistic of perfect model data fit (Cai & Monroe, 2014). For the mastery subscale, model level fit was indicated as acceptable, $C_2(5) = 24.46$, $p < .001$, RMSEA = .06.

Table 4.14 presents the estimated parameters for the mastery subscale. The subscale showed acceptable parameter estimation with item slopes ranging from 1.44 (mastery6) to 2.59 (mastery4) and threshold estimates ranging from -2.41 (mastery6 and mastery2, b_1) to .66 (mastery6, b_4). Mastery4, 'it's important to understand the work, not just memorize it,' is the

[5] Slope parameter order remained the same in the full model and the reduced model with mastery2 dropped; however, removing mastery1 (and adding back in mastery2) caused a shift in slope parameter order. This suggests that the presence of mastery1 was unduly influencing the parameter estimation.

most discriminating item, showing the strongest relationship to the underlying latent trait of perceived classroom mastery (λ=.84) while mastery6, 'it's OK to make mistakes so long as you are learning,' is the least discriminating item, showing the weakest relationship to the underlying latent trait of perceived classroom mastery (λ=.65). The threshold estimates convey additional information about the range of optimal scale efficacy. Despite a marginal reliability, interpreted much as traditional reliability in the CTT framework, of .83 for the entire scale, the thresholds suggest that this reliability is only applicable to a subset of this range. Outside of the range of these thresholds, limited information is provided to discriminate between individuals, reducing the accuracy of latent trait estimates for those who have a score greater than half a standard deviation above average. In CTT terms, the conditional reliability would be reduced for those much above the average latent trait estimate for the group.

To better convey this, IIFs and a TIF can be created. The IIF aids visualization of some key components of IRT, such as the notion that precision of an estimated latent trait score depends on location along the continuum. Rather than estimate one reliability score for an entire scale, information estimates based on theta location can translate to standard errors of the estimate (standard error of the estimate $\cong 1/\sqrt{\text{Information}}$). This relationship allows for the construction of a TIF, consisting of a summation of each individual item information function, which displays the range of the continuum for which score accuracy is greatest. Figure 4.1 displays the individual IIFs for mastery items 2 through 6. Figure 4.1 graphically displays the relationship of slopes and thresholds to item information: mastery6, found to be the least discriminating can be seen to provide the least information across the continuum while mastery3 and mastery4 provide the most information, reflecting higher estimated slopes (and factor loadings). Small peaks in this information curve represent the threshold values which show when an individual with an equivalent latent trait estimate has an equal chance of choosing either category.

When these IIFs are summed, a TIF for the scale is created. Figure 4.2 shows the TIF for the perceived classroom mastery scale where the maximum precision, a combination of large information and reduced standard error of the estimate, is found for latent traits estimated between a theta of -2 to +.5. Comparing this to a histogram of estimated thetas for this sample, approximately 70.35% (n=828) of people are well targeted by this scale, with 1.7% (n=20) scoring lower than the optimal information area and 27.95% (n=329) scoring better than the optimal information area (see Figure 4.3). Overall, this scale appears to be slightly easier than needed to effectively differentiate between individuals with higher latent trait estimates.

Performance approach subscale of the PALS classroom goal structures scale. The performance approach subscale consists of 3 items with response choices on a 5-point likert-type scale ranging from 'not at all true of me' to 'very true of me.' Initially fitting these 3 items to a unidimensional IRT model led to no identified issues of positive local dependency with all (LD) χ^2 values being negative. Item level fit using the S-χ^2 statistic indicated lack of adequate fit for all items; however, keeping in mind the caveats to this statistic mentioned during analysis of the mastery subscale, the heuristic approach as implemented in MODFIT was re-examined. Once again, visual examination of the empirical data plotted against estimated item parameters suggested adequate fit. Model level fit was assessed using a full information fit statistic, G_2, due to issues in estimating the traditional limited information fit statistic, C_2 , when less than 4 items compose a scale (Cai &Monroe, 2014). For the performance approach subscale, model level fit was indicated as acceptable through the RMSEA value, $G_2(80) = 300.84$, $p < .001$, RMSEA = .05.

Table 4.15 presents the estimated parameters for the performance approach subscale. This subscale showed acceptable parameter estimation with item slopes ranging from 2.04 (papp1) to 2.96 (papp3) and threshold estimates ranging from -2.75 (papp1, b_1) to .51 (papp2, b_4). Papp3, 'it's important to get high scores on tests,' is the most discriminating item, showing the strongest relationship to the underlying latent trait of perceived classroom performance approach (λ=.87) while papp1, 'getting good grades is the main goal,' is the least discriminating item, showing the weakest relationship to the underlying latent trait of perceived classroom performance approach (λ=.77). Marginal reliability of the perceived classroom performance approach scale is .76, but this value would only be reflective of the true reliability if the distribution of information were uniform across the scale.

Visual examination of the IIFs show that papp3, 'it's important to get high scores on tests', is providing the majority of the information for this subscale with perfapp1 and perfapp2 providing approximately equivalent information (see Figure 4.4). All three items show similar patterns of information in terms of the range of the latent trait distribution covered. The TIF provides a simpler representation of overall available information about the estimated latent trait (see Figure 4.5). The TIF for performance approach mimics the pattern found in the mastery TIF with greatest precision being found towards the lower estimated latent trait values, specifically between a theta of -2 to +.25. Comparing this to a histogram of estimated theta values, only 60.15% (n=708) of respondents in this sample have estimated thetas that fall within the optimal range of precision (see Figure 4.6). Respondents who are poorly targeted by this scale are mostly those scoring higher on the latent trait of classroom performance approach with 38.57% of this sample (n=454) exceeding the optimal precision range of the items; in contrast, only 1.27% (n=15) of respondents score below the optimal target range of these items. Once again, this scale would need items that are more difficult to agree with to increase the precision of the scale for almost half of the respondents.

Performance Avoid subscale of the PALS classroom goal structures scale. The performance avoid subscale consists of 5 items with response choices on a 5-point likert-type scale ranging from 'not at all true of me' to 'very true of me.' When initially fitting these 5 items to a unidimensional IRT model, pavoid1 and pavoid5 show evidence of positive local dependence using the Chen and Thissen's (1997) standardized localized dependency (LD) χ^2 statistic with a value of + 20.2. To assess the magnitude of this issue, sensitivity calibration was conducted for this subscale. Calibration using all items and the two abbreviated calibrations showed very similar results. Order of magnitude of the slope parameters remained the same no matter which item was dropped, but the actual magnitude of the slope parameters differed. By dropping either pavoid1 or pavoid5, the estimated slope values for pavoid4 and pavoid2 increased, suggesting that the local dependence between pavoid1 and pavoid5 was essentially creating a super-item. This dependency could distort the meaning of the latent dimension towards the cause of the local dependency rather than the intended construct. In turn, this appears to be non-trivial dependency that necessitates the dropping of one of the two problematic items. Since these items have comparable slope magnitudes and thresholds, no strong empirical rationale exists for determining which item to drop. Instead, an examination of the item stems suggests that pavoid5 may be more difficult to understand due to the inclusion of two negative terms in the item stem ('one of the main goals is to avoid looking like you can't do the work'). All subsequent analyses will be performed on the limited 4-item perceived performance avoid classroom goal orientation scale. Using the modified performance avoid scale, no positive LD issues were identified.

Item level fit used the same S-χ^2 statistic again suggested poor item fit; however the heuristic graphical MODFIT approach supports the notion that adequate fit is an acceptable assumption. Model level fit was again analyzed using the C_2 statistic with emphasis placed on the associated RMSEA value (Cai & Monroe, 2014). For the performance avoid subscale, model level fit was indicated as acceptable based on the RMSEA value, $C_2(2) = 7.22$, $p < .001$, RMSEA = .05.

Table 4.16 presents the estimated parameters for the performance avoid subscale. The subscale showed acceptable parameter estimation with item slopes ranging from 1.44 (pavoid1) to 2.77 (pavoid4) and threshold estimates ranging from -1.77 (pavoid1, b_1) to 1.60 (pavoid1, b_4). Pavoid4, 'it's very important not to look dumb,' is the most discriminating item, showing the strongest relationship to the underlying latent trait of perceived classroom performance avoid (λ=.85) while pavoid1, 'showing others that you are not bad at classwork is really important,' is the least discriminating item, showing the weakest relationship to the underlying latent trait of perceived classroom performance avoid (λ=.64).

Unlike the previous subscales, the performance avoid subscale demonstrates a more balanced distribution of item information in relation to the latent ability distribution, rendering the marginal reliability of .81 more accurate across the spectrum of the latent trait distribution. This relationship is reflected graphically in the IIFs (see Figure 4.7) and in the summative display of the TIF for the scale (see Figure 4.8). The TIF shows that the maximum precision is between -1.2 and +1 for this subscale. When compared to the estimated latent trait scores for this sample, 77.66% (n = 914) of respondents are well-targeted by this scale, with 13.42% (n = 158) of respondents scoring above the area of maximum precision and 8.92% (n = 105) of respondents scoring below the area of maximum precision (see Figure 4.9). Although this scale covers the central area of respondents relatively well, other items which better target the ends of the spectrum would improve precision.

Cross-Validation Sample

Although the initial intent was to restrict the validation analyses to the optimal correlated traits model and a breakdown of the unidimensional subscales, analysis on the initial sample revealed some potential problems that need further investigation in the confirmatory sample: namely, problematic LD and difficulties with item fit. To address this, the full series of analyses will be performed on the confirmatory sample to assess the stability of the results.

Correlated Traits Models (Models A through G)

This correlated traits models will be analyzed comparably to the initial sample with a focus on result consistency across the two samples. Data anomalies that consistently occur across this second validation sample may be indicative of information worthy of interpretation. Particular attention will be given to: the potential persistence of negative LD issues as a possible indicator of item redundancy, overall item fit as a potential indicator of scale stability, and the fit of Models C and D.

Assumptions: Evaluation of conditional independence. In the initial sample, negative LD was highly prevalent across all models, positive LD helped reveal the issues with Model B specification, and positive LD problems persisted at the individual item level across all correlated trait models. For the validation sample, the assumption of conditional independence was again assessed for veracity using Chen and Thissen's (1997) standardized LD χ^2 statistic with absolute values in excess of 10 being flagged as potentially problematic. Once again, positive LD is particularly concerning due to the possibility of inflating item parameters, thus suggesting more precision in the scale than is actually present (e.g., Toland, 2014).

In the validation sample, negative LD continues to be prevalent. Specifically, negative LD occurs in slightly less than 50% for models A, C, and D while model B demonstrates negative LD in approximately 57% of item pairs (see Table 4.17). The pattern of positive LD, which provides information about model misspecification, follows a pattern consistent with the initial sample but with slightly clearer implications (see Table 4.18). Model A, our unidimensional model, shows a more cohesive pattern of positive LD on the diagonals suggesting that item interrelationships within each of the three subscales is greater than predicted when estimated with a unidimensional model. Model B, which estimates two correlated constructs of Mastery and Performance (both approach and avoid), displays a pattern of positive LD within both the performance approach and performance avoid block suggesting that these constructs should be split. Model D, which estimates two correlated constructs of approach (both mastery and performance) and avoid, shows a pattern of positive LD in the performance approach block suggesting that this construct should be separated from mastery as a distinct construct. Model C, the three construct correlated traits model, demonstrates item issues rather than distinct patterns suggestive of dimension misspecification. All items will be kept in for model comparison in order to maintain consistency with the prior analysis.

Comparing models: Assessing fit. Item-level model-data fit showed marked improvement for the validation sample with Models A and B each estimating 4 items as fitting and Models C and D each estimating 7 items as fitting (see Table 4.17). This difference may be attributable to sample size differences (n=1765 in the validation sample compared to n=1177 in the initial sample).

Global model-data fit was analyzed comparably to procedures described in the initial sample with comparable results. Models A and B showed poor global fit and markedly poorer performance when compared to Model C (see Table 4.17). Both Models C and D demonstrated acceptable global fit when using the RMSEA associated with the C_2 statistic (RMSEA=.06 for the former and .08 for the latter); however, Model C showed significant improvement in global fit based on the deviance statistic ($\chi^2(2) = 276.15$, $p < .001$). Model C best represents the data in the validation sample.

Bifactor Models (Models E through G)

The bifactor models show the same patterns and relationships described in the initial analysis (to view the results and compare to the correlated traits models, see Tables 4.19 through 4.28). Overall, parsing the models by within-item multidimensionality rather than between subscale multidimensionality has a few consistent components worthy of note. First, the performance avoid items appear to be a distinct construct with minimal overlap with mastery and performance approach items. Specific ECV for the performance avoid items ranged from .26 to .27 across all bifactor models suggesting that approximately 26% of the common variance is being explained by this specific trait. A second result consistent across the two samples is that there is a great deal of overlap in the approach constructs (mastery and performance approach) despite the mastery items representing a distinct construct ($ECV_{mastery} \approx .15$). Of particular note is Mastery1, ‘trying hard is important,’ which tends to load strongly on the general factor driven by performance approach items and weakly on the mastery specific factor across all models. A final conclusion is that the three-factor correlated traits model is suggested as optimal across all bifactor models, with the general trait only ever explaining between 53% and 58% of the common variance. Stucky and Edelen (2015) suggest values of .85 or higher are necessary for a measure to be considered unidimensional in nature.

Unidimensional Subscales

A breakdown of the three subscales separately allows for an assessment of whether items identified as problematic in the initial analysis continue to show LD issues using the validation sample. Focus will also be placed on whether negative LD issues persist and the theta range identified as providing maximum precision for each subscale.

Mastery subscale of the PALS classroom goal structures scale. Fitting the mastery items with a unidimensional IRT model presented some concerns with positive LD between item pairs. As was found in the initial sample, mastery1 and mastery2 showed evidence of positive LD (χ^2 = +15.9). An additional item pair, mastery4 and mastery6 also showed evidence of concerning positive LD (χ^2 = +16.6). Because the local dependency issue between mastery1 and mastery2 was identified in both the initial and validation samples, these two items were selected for sensitivity analysis.

Dropping mastery1 eliminated all positive LD issues between items and led to parameter shifts in slopes, which suggest that mastery1 may have been distorting parameter estimates. Removal of mastery2 did not solve the positive LD issues between item pairs: all subsequent analysis uses the same reduced item mastery subscale as was analyzed with the initial sample.

Using the reduced mastery subscale, slope parameters appear consistent with the initial sample. Both magnitude and order of slopes was consistent with mastery4 indicated as the most discriminating item and mastery6 indicated as the least discriminating item (see Table 4.29). Threshold estimates indicate that the validation sample shows similar coverage of the latent mastery continuum, with the subscale still demonstrating a lack of precision for those who are estimated as high in perceived classroom mastery.

Performance approach subscale of the PALS classroom goal structures scale. Similar to the initial calibration, the performance approach subscale calibrated on the validation sample showed no positive LD issues, but exhibited some negative LD issues that suggest potential item redundancy (even with only three items). Although overall model fit, optimal precision range, and item fit remained consistent from the initial to the validation sample, the slope parameter estimates shifted slightly (see Table 4.30). Papp3 remained the most discriminating item but the order of Papp1 and Papp2 shifted. Papp2, 'getting right answers is most important,' was estimated as the least discriminating item for the performance approach subscale in the validation sample. The performance approach scale may be suffering from too few items, causing a lack of stability in item estimates.

Performance avoid subscale of the PALS classroom goal structures scale. Fitting the unidimensional performance avoid scale on the validation sample resulted in positive LD issues between items that needed to be handled. Pavoid1 and pavoid5 showed positive LD issues as were found in the initial sample, but the validation sample also exhibited positive LD between pavoid2 and pavoid5. Because the former pair of items showed consistent LD problems across both samples, these were selected for a sensitivity analysis. Dropping either pavoid1 or pavoid5 eliminated all positive LD item pair issues. For consistency with the initial sample, pavoid5 was dropped for the remaining analysis.

Further analysis on the modified 4-item performance avoid subscale agreed with the results found with the initial sample (see Table 4.31). Slope magnitude and order remained the same suggesting that results found in this sample may be applicable in other implementations of this subscale. The scale continued to perform the best of the three subscales in terms of optimally matching estimating theta ranges of respondents, with the scale encompassing the same optimal range of precision based on the TIF.

Table 4.1

Results from the Uni-GR, Multi-GR, and Bifac-GR Models fit to the 14-Item PALS Classroom Goal Structures Scale

	Uni-GR	Multi-GR			Bifac-GR		
Index	Model A	Model B	Model C	Model D	Model E	Model F	Model G
# positive LD pairs flagged	18	27	14	17	11	10	9
# negative LD pairs flagged	51	55	48	51	52	54	55
# of items fit by model	1	2	3	3	3	3	3
# of parameters	70	71	73	71	84	84	84
-2LL	41706.91	41203.33	40100.79	40312.33	39940.78	39940.13	39952.84
AIC	41846.91	41345.33	40246.79	40454.33	40108.78	40108.13	40120.84
BIC	42201.86	41705.35	40616.95	40814.35	40534.72	40534.07	40546.78
$C_2(df)$	2975.97(77)	2769.42(76)	496.82(74)	753.63(76)	339.21(63)	335.19(63)	354.60(63)
RMSEA based on C_2	.18	.17	.07	.09	.06	.06	.06

Note. -2LL = -2 log likelihood or deviance statistic; AIC = Akaike information criterion; BIC = Bayesian information criterion; C_2 = Cai and Monroe's (2014) limited-information goodness-of-fit statistic for ordinal response data; RMSEA based on C_2 = root mean square error of approximation based on C_2. 91 total LD pairs possible.

Table 4.2

Positive LD values for MIRT models A through D

		Mastery						Performance Approach			Performance Avoid			
	Item	1	2	3	4	5	6	7	8	9	10	11	12	13
Mastery	1													
	2													
	3													
	4			A,D										
	5			A	A,D									
	6			A	A,B,C,D									
Performance Approach	7	A,B,C,D	B	B	B	B	B							
	8	B,C	B,C	B	B	B	B	B						
	9	B	B	B	B	B	B	B	A,B,D					
Performance Avoid	10		C,D	C,D	C,D	C,D	C		D					
	11								A,D		A			
	12								C,D	D	A	A,B		
	13					C			D	D	A	A,B,C,D	A,B	
	14		C								A,B,C,D	A,B	A	A,B

Note. As is common in survey instruments, there was an excessive amount of negative LD (Toland, 2014); thus, for simplicity, only positive LD values of 10 or greater are displayed in this table; A = Model A [unidimensional]; B = Model B [mastery versus performance (approach and avoid)]; C = Model C [mastery, performance approach, performance avoid]; D = Model D [approach (mastery and performance approach) versus avoid (performance avoid)].

Table 4.3

Positive LD values for MIRT models C-D

		Mastery						Performance Approach			Performance Avoid			
	Item	1	2	3	4	5	6	7	8	9	10	11	12	13
Mastery	1													
	2													
	3													
	4			D										
	5				D									
	6				C,D									
Performance Approach	7	C,D												
	8	C	C											
	9								D					
Performance Avoid	10		C,D	C,D	C,D	C,D	C		D					
	11								D					
	12								C,D	D				
	13					C			D	D		C,D		
	14		C								C,D			

Note. As is common in survey instruments, there was an excessive amount of negative LD (Toland, 2014); thus, for simplicity, only positive LD values of 10 or greater are displayed in this table; C = Model C [mastery, performance approach, performance avoid]; D = Model D [approach (mastery and performance approach) versus avoid (performance avoid)].

Table 4.4

Unidimensional Model Item Parameter Estimates for the 14-item PALS Classroom Goal Structures Scale

	Factor Loading	Slope	Intercept			
Item	λ	a	c_1	c_2	c_3	c_4
Mastery1	0.79	2.21	4.70	3.52	1.70	-0.31
Mastery2	0.79	2.19	4.99	3.11	1.05	-1.15
Mastery3	0.79	2.20	4.91	3.09	1.11	-1.04
Mastery4	0.75	1.95	4.25	2.82	0.90	-1.19
Mastery5	0.79	2.21	4.83	3.12	0.98	-1.40
Mastery6	0.58	1.23	3.27	2.27	0.66	-0.88
Papp1	0.76	2.00	5.52	3.56	1.81	-0.06
Papp2	0.75	1.96	4.84	3.12	1.19	-0.99
Papp3	0.76	1.97	4.75	3.32	1.53	-0.59
PAvoid1	0.50	0.98	2.25	1.10	-0.45	-1.98
PAvoid2	0.32	0.57	1.78	0.67	-0.57	-1.77
PAvoid3	0.44	0.84	2.30	1.04	-0.42	-1.86
PAvoid4	0.42	0.78	1.98	0.85	-0.43	-1.60
PAvoid5	0.36	0.66	1.70	0.67	-0.68	-1.92

Note. λ = full information factor loadings; a = slope; c_1-c_4 = intercepts.

Table 4.5

2-Dimensional Model Item Parameter Estimates for the 14-item PALS Classroom Goal Structures Scale (Mastery, Performance)

	Factor loading		Slope		Intercept			
Item	λ_1	λ_2	a_1	a_2	c_1	c_2	c_3	c_4
Mastery1	0.75		1.93		4.40	3.28	1.58	-0.31
Mastery2	0.79		2.22		5.09	3.18	1.05	-1.21
Mastery3	0.83		2.53		5.37	3.41	1.23	-1.17
Mastery4	0.81		2.39		4.83	3.23	1.03	-1.39
Mastery5	0.83		2.54		5.31	3.46	1.08	-1.58
Mastery6	0.63		1.40		3.43	2.40	0.70	-0.94
Papp1		0.60		1.26	4.59	2.91	1.46	-0.06
Papp2		0.65		1.45	4.22	2.71	1.03	-0.86
Papp3		0.63		1.39	4.07	2.83	1.32	-0.49
PAvoid1		0.71		1.71	2.73	1.33	-0.60	-2.47
PAvoid2		0.65		1.44	2.22	0.87	-0.70	-2.23
PAvoid3		0.71		1.72	2.89	1.34	-0.55	-2.41
PAvoid4		0.73		1.82	2.6	1.14	-0.58	-2.16
PAvoid5		0.63		1.38	2.03	0.80	-0.85	-2.35

Note. λ_1- λ_2 = full information factor loadings; a_1-a_2 = MIRT slopes; c_1-c_4 = intercepts.

Table 4.6

Inter-factor Correlations for the 2-Dimensional Model (Mastery, Performance)

	Factor 1	Factor 2
Factor 1	1.00	
Factor 2	0.60	1.00

Note. Factor 1 = Mastery; factor 2 = Performance.

Table 4.7

3-Dimensional Model Item Parameter Estimates for the 14-item PALS Classroom Goal Structures Scale (Mastery, Performance Approach, Performance Avoid)

	Factor loading			Slope			Intercept			
Item	λ_1	λ_2	λ_3	a_1	a_2	a_3	c_1	c_2	c_3	c_4
Mastery1	0.77			2.07			4.57	3.42	1.65	-0.32
Mastery2	0.79			2.22			5.08	3.19	1.06	-1.21
Mastery3	0.83			2.52			5.35	3.41	1.23	-1.18
Mastery4	0.80			2.29			4.69	3.14	1.00	-1.35
Mastery5	0.82			2.42			5.14	3.36	1.04	-1.55
Mastery6	0.64			1.41			3.44	2.42	0.71	-0.95
Papp1		0.80			2.25		5.94	3.87	1.98	-0.06
Papp2		0.81			2.33		5.36	3.51	1.36	-1.12
Papp3		0.83			2.57		5.60	3.98	1.86	-0.70
PAvoid1			0.71			1.73	2.76	1.29	-0.66	-2.49
PAvoid2			0.77			2.02	2.64	1.01	-0.87	-2.65
PAvoid3			0.75			1.92	3.09	1.39	-0.64	-2.57
PAvoid4			0.83			2.54	3.2	1.34	-0.78	-2.66
PAvoid5			0.70			1.68	2.22	0.84	-0.97	-2.57

Note. λ_1- λ_3 = full information factor loadings; a_1-a_3 = MIRT slopes; c_1-c_4 = intercepts.

Table 4.8

Inter-factor Correlations for the 3-Dimensional Model (Mastery, Performance Approach, Performance Avoid)

	Factor 1	Factor 2	Factor 3
Factor 1	1.00		
Factor 2	0.82	1.00	
Factor 3	0.34	0.49	1.00

Note. Factor 1 = Mastery; factor 2 = Performance Approach; factor 3 = Performance Avoid.

Table 4.9

2-Dimensional Model Item Parameter Estimates for the 14-item PALS Classroom Goal Structures Scale (Approach, Avoid)

	Factor loading		Slope		Intercept			
Item	λ_1	λ_2	a_1	a_2	c_1	c_2	c_3	c_4
Mastery1	0.79		2.22		4.76	3.57	1.72	-0.33
Mastery2	0.79		2.19		5.05	3.16	1.05	-1.19
Mastery3	0.81		2.33		5.10	3.23	1.15	-1.11
Mastery4	0.77		2.05		4.41	2.94	0.93	-1.26
Mastery5	0.80		2.24		4.91	3.19	0.99	-1.46
Mastery6	0.62		1.33		3.37	2.35	0.68	-0.92
Papp1	0.75		1.92		5.45	3.51	1.78	-0.08
Papp2	0.73		1.82		4.70	3.04	1.14	-0.99
Papp3	0.74		1.88		4.67	3.26	1.48	-0.61
PAvoid1		0.72		1.74	2.77	1.29	-0.67	-2.50
PAvoid2		0.76		2.02	2.64	1.01	-0.87	-2.65
PAvoid3		0.74		1.90	3.08	1.38	-0.63	-2.56
PAvoid4		0.83		2.52	3.19	1.34	-0.78	-2.65
PAvoid5		0.71		1.70	2.24	0.85	-0.98	-2.59

Note. λ_1- λ_2 = full information factor loadings; a_1-a_2 = MIRT slopes; c_1-c_4 = intercepts.

Table 4.10

Inter-factor Correlations for the 2-Dimensional Model (Approach, Avoid)

	Factor 1	Factor 2
Factor 1	1.00	
Factor 2	0.41	1.00

Note. Factor 1 = Approach; Factor 2 = Avoid.

Table 4.11

Bifactor Model Item Parameter Estimates for the 14-item PALS Classroom Goal Structures Scale (Mastery, Performance)

	Conditional slope			Intercept				Marginal factor loading						Marginal slope		
Item	a^{G}	a^{S1}	a^{S2}	c_1	c_2	c_3	c_4	$^*G$	$^*S_1$	$^*S_2$	$IECV_G$	$IECV_{S1}$	$IECV_{S2}$	$^*a^{G}$	$^*a^{S1}$	$^*a^{S2}$
Mastery1	2.19	0.55		4.82	3.61	1.75	-0.33	0.77	0.19		0.94	0.06		2.08	0.34	
Mastery2	1.93	0.97		5.00	3.13	1.04	-1.18	0.70	0.35		0.80	0.20		1.68	0.64	
Mastery3	2.08	1.52		5.43	3.47	1.25	-1.19	0.67	0.49		0.65	0.35		1.55	0.96	
Mastery4	1.88	1.82		5.12	3.44	1.10	-1.48	0.60	0.58		0.52	0.48		1.28	1.22	
Mastery5	1.98	1.60		5.31	3.47	1.08	-1.60	0.65	0.52		0.60	0.40		1.44	1.04	
Mastery6	1.13	0.90		3.47	2.43	0.71	-0.96	0.51	0.40		0.61	0.39		1.00	0.75	
Papp1	2.51		0.04	6.36	4.18	2.13	-0.07	0.83		0.01	1.00		0.00	2.51		0.02
Papp2	2.22		0.32	5.24	3.43	1.33	-1.10	0.79		0.11	0.98		0.02	2.18		0.19
Papp3	2.32		0.21	5.26	3.73	1.74	-0.66	0.80		0.07	0.99		0.01	2.30		0.12
PAvoid1	1.01		1.41	2.75	1.31	-0.64	-2.49	0.42		0.58	0.34		0.66	0.78		1.21
PAvoid2	0.63		2.12	2.79	1.05	-0.95	-2.80	0.23		0.76	0.08		0.92	0.39		1.99
PAvoid3	0.96		1.65	3.07	1.39	-0.62	-2.56	0.38		0.65	0.25		0.75	0.69		1.44
PAvoid4	1.00		2.38	3.24	1.35	-0.80	-2.70	0.32		0.77	0.15		0.85	0.58		2.05
PAvoid5	0.60		1.54	2.22	0.83	-0.98	-2.56	0.25		0.65	0.13		0.87	0.44		1.45
ECV								0.59	0.14	0.28						

Note. G = general PALS trait; S_1 = specific trait 1 – Mastery; S_2 = specific trait 2 – Performance (approach and avoid); a^{G} = conditional slopes for the general trait; a^{S1} – a^{S2} = conditional slopes for specific traits 1-2; c_1-c_4 = intercepts; ECV = explained common variance by factor; $IECV_G$ = item explained common variance for the general trait; $IECV_{S1}$ –$IECV_{S2}$= item explained common variance for specific traits 1-2; ECV values do not sum to 1 due to rounding error; $^*a^{G}$ = marginal slope for general trait; $^*a^{S1}$ - $^*a^{S2}$ = marginal slopes for specific traits 1-2.

Table 4.12

Bifactor Model Item Parameter Estimates for the 14-item PALS Classroom Goal Structures Scale (Mastery, Performance Approach, Performance Avoid)

	Conditional slope				Intercept				Marginal factor loading								Marginal slope			
Item	a^G	a^{S1}	a^{S2}	a^{S3}	c_1	c_2	c_3	c_4	$*G$	$*S_1$	$*S_2$	$*S_3$	$IECV_G$	$IECV_{S1}$	$IECV_{S2}$	$IECV_{S3}$	$*a^G$	$*a^{S1}$	$*a^{S2}$	$*a^{S3}$
Mastery1	2.34	0.44			4.96	3.73	1.81	-0.33	0.80	0.15			0.97	0.03			2.27	0.26		
Mastery2	2.01	0.89			5.04	3.16	1.05	-1.19	0.72	0.32			0.84	0.16			1.78	0.57		
Mastery3	2.10	1.49			5.43	3.47	1.25	-1.19	0.68	0.48			0.67	0.33			1.58	0.94		
Mastery4	1.92	1.80			5.15	3.46	1.11	-1.49	0.61	0.57			0.53	0.47			1.32	1.19		
Mastery5	2.02	1.53			5.30	3.47	1.08	-1.59	0.66	0.50			0.64	0.36			1.50	0.99		
Mastery6	1.14	0.88			3.47	2.43	0.71	-0.95	0.51	0.39			0.63	0.37			1.01	0.73		
Papp1	2.46		0.25		6.29	4.12	2.10	-0.07	0.82		0.08		0.99		0.01		2.43		0.14	
Papp2	2.13		0.61		5.18	3.40	1.32	-1.10	0.76		0.22		0.92		0.08		2.00		0.38	
Papp3	3.18		2.22		7.64	5.50	2.60	-0.97	0.75		0.52		0.67		0.33		1.93		1.05	
PAvoid1	1.09			1.37	2.76	1.31	-0.64	-2.50	0.45			0.56	0.39			0.61	0.85			1.15
PAvoid2	0.72			2.08	2.78	1.04	-0.95	-2.79	0.26			0.75	0.11			0.89	0.46			1.92
PAvoid3	1.03			1.60	3.07	1.39	-0.62	-2.56	0.40			0.63	0.29			0.71	0.75			1.37
PAvoid4	1.11			2.33	3.24	1.35	-0.80	-2.70	0.36			0.75	0.18			0.82	0.65			1.95
PAvoid5	0.69			1.52	2.22	0.83	-0.98	-2.57	0.29			0.64	0.17			0.83	0.51			1.41
ECV									0.58	0.12	0.04	0.26								

Note. G = general PALS trait; S_1 = specific trait 1 – Mastery; S_2 = specific trait 2 – Performance Approach; S_3 = specific trait 3 – Performance Avoid; a^G = conditional slopes for the general trait; a^{S1} – a^{S3} = conditional slopes for specific traits 1-3; c_1-c_4 = intercepts; ECV = explained common variance by factor; $IECV_G$ = item explained common variance for the general trait; $IECV_{S1}$ –$IECV_{S3}$= item explained common variance for specific traits 1-3; ECV values do not sum to 1 due to rounding error; $*a^G$ = marginal slope for general trait; $*a^{S1}$ - $*a^{S3}$ = marginal slopes for specific traits 1-3.

Table 4.13

Bifactor Model Item Parameter Estimates for the 14-item PALS Classroom Goal Structures Scale (Approach, Avoid)

	Conditional slope			Intercept				Marginal factor loading						Marginal slope		
Item	a^G	a^{S1}	a^{S2}	c_1	c_2	c_3	c_4	$*G$	$*S_1$	$*S_2$	$IECV_G$	$IECV_{S1}$	$IECV_{S2}$	$*a^G$	$*a^{S1}$	$*a^{S2}$
Mastery1	2.10	0.76		4.77	3.58	1.73	-0.32	0.75	0.27		0.88	0.12		1.92	0.48	
Mastery2	1.84	1.14		5.01	3.13	1.04	-1.18	0.67	0.41		0.72	0.28		1.53	0.77	
Mastery3	1.93	1.70		5.42	3.46	1.25	-1.19	0.63	0.55		0.56	0.44		1.36	1.12	
Mastery4	1.71	1.99		5.13	3.44	1.11	-1.48	0.55	0.64		0.42	0.58		1.11	1.4	
Mastery5	1.84	1.73		5.29	3.46	1.08	-1.59	0.60	0.57		0.53	0.47		1.29	1.17	
Mastery6	1.02	1.03		3.48	2.44	0.72	-0.96	0.46	0.46		0.50	0.50		0.87	0.88	
Papp1	2.48	0.16		6.31	4.14	2.11	-0.07	0.82	0.05		1.00	0.00		2.47	0.09	
Papp2	2.25	0.19		5.26	3.45	1.33	-1.11	0.80	0.07		0.99	0.01		2.24	0.11	
Papp3	2.34	0.21		5.29	3.76	1.75	-0.67	0.81	0.07		0.99	0.01		2.32	0.12	
PAvoid1	1.10		1.35	2.75	1.31	-0.64	-2.49	0.45		0.55	0.40		0.60	0.86		1.13
PAvoid2	0.78		2.04	2.77	1.04	-0.94	-2.78	0.28		0.74	0.13		0.87	0.50		1.85
PAvoid3	1.08		1.57	3.07	1.39	-0.62	-2.56	0.42		0.61	0.32		0.68	0.79		1.33
PAvoid4	1.17		2.30	3.24	1.35	-0.80	-2.69	0.38		0.74	0.21		0.79	0.70		1.89
PAvoid5	0.70		1.52	2.23	0.83	-0.99	-2.58	0.29		0.64	0.17		0.83	0.52		1.41
ECV								0.57	0.17	0.25						

Note. G = general PALS trait; S_1 = specific trait 1 – Approach (Mastery and Performance Approach); S_2 = specific trait 2 – Performance Avoid; a^G = conditional slopes for the general trait; a^{S1} – a^{S2} = conditional slopes for specific traits 1-2; c_1-c_4 = intercepts; ECV = explained common variance by factor; $IECV_G$ = item explained common variance for the general trait; $IECV_{S1}$ –$IECV_{S2}$= item explained common variance for specific traits 1-2; ECV values do not sum to 1 due to rounding error; $*a^G$ = marginal slope for general trait; $*a^{S1}$ - $*a^{S2}$ = marginal slopes for specific traits 1-2.

Table 4.14

Unidimensional Model Item Parameter Estimates for the Reduced 5-item Mastery Subscale of the PALS Classroom Goal Structures Scale (Mastery1 removed)

	Factor Loading	Slope	Threshold				Item-fit Statistics	
Item	Λ	a	b_1	b_2	b_3	b_4	S-χ^2	*p*
Mastery2	0.76 (.03)	2.00 (.11)	-2.41 (.13)	-1.50 (.08)	-0.50 (.05)	0.56 (.05)	70.30	.0307
Mastery3	0.83 (.02)	2.58 (.15)	-2.12 (.10)	-1.36 (.07)	-0.49 (.05)	0.46 (.05)	99.20	<.0001
Mastery4	0.84 (.02)	2.59 (.14)	-1.97 (.10)	-1.32 (.07)	-0.43 (.05)	0.56 (.05)	98.50	<.0001
Mastery5	0.83 (.02)	2.53 (.14)	-2.11 (.10)	-1.38 (.07)	-0.43 (.05)	0.63 (.05)	113.2	<.0001
Mastery6	0.65 (.04)	1.44 (.08)	-2.41 (.14)	-1.69 (.10)	-0.50 (.06)	0.66 (.07)	111.4	<.0001

Note. λ = full information factor loadings; a = item slope (discrimination); b_1-b_4 = thresholds, S-χ^2 = item-fit statistic, *p* = p value associated with the item-fit statistic. Values in parentheses are standard error estimates. Mastery1 was removed due to positive LD issues with Matery2.

Table 4.15

Unidimensional Model Item Parameter Estimates for the 3-item Performance Approach Subscale of the PALS Classroom Goal Structures Scale

	Factor Loading	Slope	Threshold				Item-fit Statistics	
Item	Λ	a	b_1	b_2	b_3	b_4	S-χ^2	*p*
Papp1	0.77 (.03)	2.04 (.13)	-2.75 (.16)	-1.81 (.09)	-0.93 (.06)	0.04 (.05)	69.6	<.0001
Papp2	0.79 (.03)	2.17 (.13)	-2.38 (.13)	-1.58 (.08)	-0.62 (.05)	0.51 (.05)	92.8	<.0001
Papp3	0.87 (.03)	2.96 (.25)	-2.10 (.11)	-1.51 (.07)	-0.72 (.05)	0.27 (.04)	58.7	<.0001

Note. λ = full information factor loadings; a = item slope (discrimination); b_1-b_4 = thresholds, S-χ^2 = item-fit statistic, *p* = p value associated with the item-fit statistic. Values in parentheses are standard error estimates.

Table 4.16

Unidimensional Model Item Parameter Estimates for the Reduced 4-item Performance Avoid Subscale of the PALS Classroom Goal Structures Scale (Pavoid5 is dropped due to LD issues)

	Factor Loading	Slope	Threshold				Item-fit Statistics	
Item	Λ	a	b_1	b_2	b_3	b_4	S-χ^2	*p*
Pavoid1	0.64 (.04)	1.44 (.08)	-1.77 (.10)	-0.82 (.07)	0.43 (.06)	1.6 (.10)	138.9	<.0001
Pavoid2	0.79 (.03)	2.21 (.12)	-1.27 (.07)	-0.48 (.05)	0.43 (.05)	1.28 (.07)	162.9	<.0001
Pavoid3	0.74 (.03)	1.85 (.10)	-1.66 (.09)	-0.74 (.06)	0.35 (.05)	1.37 (.08)	151.1	<.0001
Pavoid4	0.85 (.02)	2.77 (.17)	-1.24 (.06)	-0.51 (.05)	0.31 (.04)	1.03 (.06)	127.2	<.0001

Note. λ = full information factor loadings; a = item slope (discrimination); b_1-b_4 = thresholds, S-χ^2 = item-fit statistic, *p* = p value associated with the item-fit statistic. Values in parentheses are standard error estimates. Pavoid5 was dropped due to positive LD issues with Pavoid2.

Table 4.17

Cross-Validation Results from the Uni-GR, Multi-GR, and Bifac-GR Models fit to the 14-Item PALS Classroom Goal Structures Scale

	Uni-GR	Multi-GR			Bifac-GR		
Index	Model A	Model B	Model C	Model D	Model E	Model F	Model G
# positive LD pairs flagged	21	25	18	24	15	13	18
# negative LD pairs flagged	42	52	42	40	45	48	44
# of items fit by model	4	4	7	7	7	7	7
# of parameters	70	71	73	71	84	84	84
-2LL	61854.08	61251.90	59677.12	59953.27	59456.19	59468.85	59460.62
AIC	61994.08	61393.90	59823.12	60095.27	59624.19	59636.85	59628.62
BIC	62377.40	61782.69	60222.86	60484.06	60084.16	60096.82	60088.60
$C_2(df)$	3926.75(77)	4055.56(76)	606.70(74)	855.24(76)	397.08(63)	395.16(63)	400.23(63)
RMSEA based on C_2	0.17	0.17	0.06	0.08	0.05	0.05	0.06

Note. -2LL = -2 log likelihood or deviance statistic; AIC = Akaike information criterion; BIC = Bayesian information criterion; C_2 = Cai and Monroe's (2014) limited-information goodness-of-fit statistic for ordinal response data; RMSEA based on C_2 = root mean square error of approximation based on C_2. 91 total LD pairs possible. All cross-validation analyses are run on Sample B.

Table 4.18

Positive LD values for MIRT models A-D

		Mastery						Performance Approach			Performance Avoid			
	Item	1	2	3	4	5	6	7	8	9	10	11	12	13
Mastery	1													
	2													
	3													
	4		A	A,C,D										
	5			A,D	A,D									
	6				A,B,C,D	A,D								
Performance Approach	7	A,B,C,D	B	B										
	8	B	B,C	B		B		A,B,D						
	9	B	B		B	B		A,B,D	A,B,D					
Performance Avoid	10		C,D	C	C,D	C,D			D	D				
	11								A,D		A,B			
	12		C			C,D			C,D	C,D	A,B	A,B		
	13		C,D						D	D	A,B	A,B,C,D	A,B	
	14	C	C			C,D					A,B,C,D	A,B	A,B	A,B

Note. As is common in survey instruments, there was an excessive amount of negative LD (Toland, 2014); thus, for simplicity, only positive LD values greater than 10 are displayed in this table; A = Model A [unidimensional]; B = Model B [mastery versus performance (approach and avoid)]; C = Model C [mastery, performance approach, performance avoid]; D = Model D [approach (mastery and performance approach) versus avoid (performance avoid)].

Table 4.19

Cross-Validation Unidimensional Model Item Parameter Estimates for the 14-item PALS Classroom Goal Structures Scale

	Factor Loading	Slope	Intercept			
Item	Λ	a	c_1	c_2	c_3	c_4
Mastery1	0.79	2.17	4.62	3.30	1.59	-0.40
Mastery2	0.79	2.16	4.85	3.06	1.12	-1.08
Mastery3	0.78	2.11	4.57	3.07	1.11	-0.97
Mastery4	0.74	1.88	4.23	2.75	1.02	-0.91
Mastery5	0.74	1.85	4.32	2.64	0.58	-1.44
Mastery6	0.60	1.28	3.32	2.18	0.67	-0.94
Papp1	0.76	1.98	5.03	3.36	1.69	-0.16
Papp2	0.74	1.85	4.56	2.94	0.96	-1.15
Papp3	0.75	1.95	4.64	3.16	1.39	-0.52
PAvoid1	0.46	0.87	2.06	0.87	-0.55	-1.90
PAvoid2	0.27	0.48	1.49	0.42	-0.77	-1.84
PAvoid3	0.41	0.78	2.08	0.91	-0.52	-1.98
PAvoid4	0.39	0.72	1.84	0.74	-0.38	-1.46
PAvoid5	0.36	0.65	1.52	0.52	-0.69	-1.85

Note. λ = full information factor loadings; a = slope; c_1-c_4 = intercepts.

Table 4.20

Cross-Validation 2-Dimensional Model Item Parameter Estimates for the 14-item PALS Classroom Goal Structures Scale (Mastery, Performance)

	Factor loading		Slope		Intercept			
Item	λ_1	λ_2	a_1	a_2	c_1	c_2	c_3	c_4
Mastery1	0.76		1.98		4.39	3.14	1.52	-0.39
Mastery2	0.79		2.20		4.91	3.11	1.14	-1.11
Mastery3	0.82		2.44		4.98	3.39	1.25	-1.07
Mastery4	0.80		2.27		4.70	3.09	1.17	-1.02
Mastery5	0.78		2.09		4.60	2.85	0.64	-1.56
Mastery6	0.66		1.49		3.51	2.33	0.72	-1.01
Papp1		0.71		1.73	4.68	3.11	1.57	-0.14
Papp2		0.73		1.84	4.53	2.93	0.97	-1.14
Papp3		0.74		1.86	4.50	3.07	1.38	-0.48
PAvoid1		0.62		1.36	2.32	0.99	-0.66	-2.21
PAvoid2		0.49		0.95	1.66	0.49	-0.85	-2.04
PAvoid3		0.62		1.33	2.39	1.06	-0.62	-2.33
PAvoid4		0.59		1.23	2.09	0.86	-0.44	-1.69
PAvoid5		0.54		1.09	1.69	0.58	-0.79	-2.09

Note. λ_1- λ_2 = full information factor loadings; a_1-a_2 = MIRT slopes; c_1-c_4 = intercepts.

Table 4.21

Inter-factor Correlations for the 2-Dimensional Model (Mastery, Performance)

	Mastery	Performance
Mastery	1.00	
Performance	0.70	1.00

Note. Mastery = Mastery items; Performance = Performance Approach and Performance Avoid items.

Table 4.22

Cross-Validation Bifactor Model Item Parameter Estimates for the 14-item PALS Classroom Goal Structures Scale (Mastery, Performance)

	Conditional slope			Intercept				Marginal factor loading						Marginal slope		
Item	a^G	a^{S1}	a^{S2}	c_1	c_2	c_3	c_4	$*G$	$*S_1$	$*S_2$	$IECV_G$	$IECV_{S1}$	$IECV_{S2}$	$*a^G$	$*a^{S1}$	$*a^{S2}$
Mastery1	2.19	0.55		4.76	3.39	1.62	-0.43	0.77	0.19		0.94	0.06		2.08	0.33	
Mastery2	1.93	0.97		4.83	3.05	1.12	-1.1	0.7	0.35		0.8	0.2		1.65	0.64	
Mastery3	2.08	1.52		4.97	3.38	1.24	-1.08	0.66	0.48		0.65	0.35		1.5	0.94	
Mastery4	1.88	1.82		5.12	3.39	1.29	-1.13	0.58	0.6		0.49	0.51		1.22	1.28	
Mastery5	1.98	1.6		4.61	2.86	0.64	-1.58	0.6	0.5		0.59	0.41		1.26	0.98	
Mastery6	1.13	0.9		3.65	2.44	0.76	-1.05	0.48	0.49		0.49	0.51		0.94	0.96	
Papp1	2.51		0.04	6.04	4.08	2.06	-0.21	0.84		-0.01	1		0	2.64		-0.02
Papp2	2.22		0.32	4.9	3.17	1.04	-1.25	0.77		0.1	0.98		0.02	2.04		0.18
Papp3	2.32		0.21	5.06	3.46	1.53	-0.58	0.8		0.05	1		0	2.23		0.09
PAvoid1	1.01		1.41	2.65	1.08	-0.81	-2.54	0.37		0.63	0.25		0.75	0.67		1.39
PAvoid2	0.63		2.12	2.17	0.63	-1.14	-2.65	0.19		0.71	0.07		0.93	0.34		1.72
PAvoid3	0.96		1.65	2.75	1.19	-0.75	-2.69	0.35		0.64	0.23		0.77	0.64		1.42
PAvoid4	1		2.38	2.74	1.08	-0.64	-2.23	0.3		0.72	0.15		0.85	0.54		1.78
PAvoid5	0.6		1.54	2.05	0.66	-1.01	-2.53	0.25		0.67	0.13		0.87	0.45		1.54
ECV								0.58	0.15	0.27						

Note. G = general PALS trait; S_1 = specific trait 1 – Mastery; S_2 = specific trait 2 – Performance (approach and avoid); a^G = conditional slopes for the general trait; a^{S1} – a^{S2} = conditional slopes for specific traits 1-2; c_1-c_4 = intercepts; ECV = explained common variance by factor; $IECV_G$ = item explained common variance for the general trait; $IECV_{S1}$ –$IECV_{S2}$= item explained common variance for specific traits 1-2; ECV values may not sum to 1 due to rounding error; $*a^G$ = marginal slope for general trait; $*a^{S1}$ - $*a^{S2}$ = marginal slopes for specific traits 1-2.

Table 4.23

Cross-Validation 3-Dimensional Model Item Parameter Estimates for the 14-item PALS Classroom Goal Structures Scale (Mastery, Performance Approach, Performance Avoid)

	Factor loading			Slope			Intercept			
Item	λ_1	λ_2	λ_3	a_1	a_2	a_3	c_1	c_2	c_3	c_4
Mastery1	0.78			2.09			4.53	3.24	1.56	-0.41
Mastery2	0.79			2.19			4.90	3.10	1.14	-1.12
Mastery3	0.82			2.43			4.98	3.37	1.24	-1.08
Mastery4	0.79			2.19			4.61	3.03	1.14	-1.01
Mastery5	0.77			2.02			4.53	2.80	0.62	-1.54
Mastery6	0.66			1.50			3.52	2.34	0.72	-1.01
Papp1		0.81			2.32		5.54	3.73	1.88	-0.19
Papp2		0.79			2.19		5.05	3.27	1.07	-1.29
Papp3		0.82			2.41		5.29	3.62	1.60	-0.60
PAvoid1			0.74			1.87	2.69	1.08	-0.84	-2.57
PAvoid2			0.71			1.72	2.09	0.61	-1.09	-2.55
PAvoid3			0.73			1.84	2.76	1.19	-0.76	-2.70
PAvoid4			0.78			2.11	2.71	1.07	-0.63	-2.20
PAvoid5			0.72			1.76	2.06	0.67	-1.01	-2.54

Note. λ_1- λ_3 = full information factor loadings; a_1-a_3 = MIRT slopes; c_1-c_4 = intercepts.

Table 4.24

Inter-factor Correlations for the 3-Dimensional Model (Mastery, Performance Approach, Performance Avoid)

	Mastery	Papp	PAvoid
Mastery	1.00		
Papp	0.83	1.00	
PAvoid	0.32	0.46	1.00

Note. Mastery = Mastery items; Papp = Performance Approach items; PAvoid = Performance Avoid items.

Table 4.25

Cross-Validation Bifactor Model Item Parameter Estimates for the 14-item PALS Classroom Goal Structures Scale (Mastery, Performance Approach, Performance Avoid)

	Conditional slope				Intercept				Marginal factor loading								Marginal slope			
Item	a^G	a^{S1}	a^{S2}	a^{S3}	c_1	c_2	c_3	c_4	*G	$*S_1$	$*S_2$	$*S_3$	$IECV_G$	$IECV_{S1}$	$IECV_{S2}$	$IECV_{S3}$	$*a^G$	$*a^{S1}$	$*a^{S2}$	$*a^{S3}$
Mastery1	2.18	0.56			4.76	3.4	1.63	-0.43	0.77	0.2			0.94	0.06			2.07	0.34		
Mastery2	1.90	0.98			4.84	3.06	1.12	-1.10	0.70	0.36			0.79	0.21			1.65	0.65		
Mastery3	1.95	1.46			4.97	3.37	1.24	-1.08	0.66	0.49			0.64	0.36			1.48	0.96		
Mastery4	1.8	1.89			5.12	3.39	1.29	-1.13	0.58	0.61			0.48	0.52			1.20	1.30		
Mastery5	1.60	1.36			4.61	2.86	0.64	-1.58	0.59	0.50			0.58	0.42			1.25	0.99		
Mastery6	1.12	1.19			3.65	2.44	0.76	-1.05	0.47	0.50			0.47	0.53			0.92	0.99		
Papp1	2.73		-0.18		6.18	4.18	2.11	-0.21	0.85		-0.06		1.00		0.00		2.71		-0.10	
Papp2	2.13		0.53		5.06	3.28	1.07	-1.30	0.77		0.19		0.94		0.06		2.03		0.33	
Papp3	2.33		0.60		5.28	3.61	1.6	-0.60	0.79		0.20		0.94		0.06		2.20		0.35	
PAvoid1	0.98			1.55	2.66	1.08	-0.82	-2.55	0.39			0.62	0.29			0.71	0.72			1.34
PAvoid2	0.55			1.76	2.16	0.62	-1.14	-2.64	0.22			0.70	0.09			0.91	0.38			1.67
PAvoid3	0.93			1.55	2.73	1.19	-0.75	-2.68	0.37			0.62	0.26			0.74	0.69			1.36
PAvoid4	0.91			1.95	2.74	1.08	-0.64	-2.23	0.33			0.71	0.18			0.82	0.60			1.72
PAvoid5	0.68			1.63	2.06	0.66	-1.02	-2.55	0.28			0.66	0.15			0.85	0.49			1.51
ECV									0.58	0.15	0.01	0.26								

Note. G = general PALS trait; S_1 = specific trait 1 – Mastery; S_2 = specific trait 2 – Performance Approach; S_3 = specific trait 3 – Performance Avoid; a^G = conditional slopes for the general trait; a^{S1} – a^{S3} = conditional slopes for specific traits 1-3; c_1-c_4 = intercepts; ECV = explained common variance by factor; $IECV_G$ = item explained common variance for the general trait; $IECV_{S1}$ –$IECV_{S3}$= item explained common variance for specific traits 1-3; ECV values may not sum to 1 due to rounding error; $*a^G$ = marginal slope for general trait; $*a^{S1}$ - $*a^{S3}$ = marginal slopes for specific traits 1-3.

Table 4.26

Cross-Validation 2-Dimensional Model Item Parameter Estimates for the 14-item PALS Classroom Goal Structures Scale (Approach, Avoid)

	Factor loading		Slope		Intercept			
Item	λ_1	λ_2	a_1	a_2	c_1	c_2	c_3	c_4
Mastery1	0.79		2.20		4.68	3.34	1.60	-0.43
Mastery2	0.79		2.16		4.88	3.08	1.12	-1.11
Mastery3	0.80		2.24		4.76	3.20	1.15	-1.03
Mastery4	0.76		1.97		4.36	2.84	1.05	-0.95
Mastery5	0.74		1.90		4.39	2.69	0.58	-1.48
Mastery6	0.63		1.38		3.41	2.26	0.69	-0.97
Papp1	0.75		1.92		4.99	3.33	1.67	-0.17
Papp2	0.72		1.75		4.48	2.88	0.93	-1.13
Papp3	0.74		1.86		4.56	3.10	1.35	-0.53
PAvoid1		0.74		1.88	2.70	1.08	-0.84	-2.58
PAvoid2		0.71		1.72	2.08	0.61	-1.09	-2.54
PAvoid3		0.73		1.81	2.73	1.18	-0.75	-2.67
PAvoid4		0.78		2.12	2.71	1.07	-0.63	-2.21
PAvoid5		0.72		1.78	2.07	0.67	-1.02	-2.56

Note. λ_1- λ_2 = full information factor loadings; a_1-a_2 = MIRT slopes; c_1-c_4 = intercepts.

Table 4.27

Inter-factor Correlations for the 2-Dimensional Model (Approach, Avoid)

	Approach	Avoid
Approach	1.00	
Avoid	0.38	1.00

Note. Approach = Mastery and Performance Approach items; Avoid = Performance Avoid items.

Table 4.28

Cross-Validation of Bifactor Model Item Parameter Estimates for the 14-item PALS Classroom Goal Structures Scale (Approach, Avoid)

	Conditional slope			Intercept				Marginal factor loading						Marginal slope		
Item	a^G	a^{S1}	a^{S2}	c_1	c_2	c_3	c_4	$*G$	$*S_1$	$*S_2$	$IECV_G$	$IECV_{S1}$	$IECV_{S2}$	$*a^G$	$*a^{S1}$	$*a^{S2}$
Mastery1	2.11	0.8		4.76	3.40	1.63	-0.43	0.75	0.28		0.87	0.13		1.91	0.50	
Mastery2	1.76	1.21		4.84	3.06	1.12	-1.10	0.64	0.44		0.68	0.32		1.43	0.84	
Mastery3	1.75	1.68		4.96	3.36	1.24	-1.07	0.59	0.57		0.52	0.48		1.24	1.17	
Mastery4	1.55	2.13		5.16	3.42	1.30	-1.14	0.49	0.68		0.35	0.65		0.97	1.57	
Mastery5	1.41	1.56		4.61	2.86	0.64	-1.58	0.52	0.58		0.45	0.55		1.04	1.20	
Mastery6	0.96	1.33		3.66	2.44	0.76	-1.05	0.41	0.56		0.34	0.66		0.76	1.16	
Papp1	2.82	0.16		6.33	4.28	2.16	-0.21	0.86	0.05		1.00	0		2.81	0.08	
Papp2	1.99	0.37		4.83	3.12	1.02	-1.23	0.75	0.14		0.97	0.03		1.94	0.24	
Papp3	2.13	0.43		4.96	3.38	1.49	-0.57	0.77	0.16		0.96	0.04		2.06	0.27	
PAvoid1	1.00		1.54	2.66	1.08	-0.82	-2.55	0.4		0.62	0.30		0.7	0.74		1.33
PAvoid2	0.59		1.74	2.16	0.62	-1.14	-2.64	0.24		0.70	0.10		0.9	0.41		1.64
PAvoid3	0.97		1.53	2.73	1.19	-0.75	-2.68	0.39		0.62	0.29		0.71	0.72		1.33
PAvoid4	0.94		1.94	2.74	1.08	-0.64	-2.23	0.34		0.71	0.19		0.81	0.62		1.70
PAvoid5	0.7		1.62	2.06	0.66	-1.02	-2.55	0.29		0.66	0.16		0.84	0.51		1.50
ECV								0.53	0.21	0.26						

Note. G = general PALS trait; S_1 = specific trait 1 – Approach (Mastery and Performance Approach); S_2 = specific trait 2 – Performance Avoid; a^G = conditional slopes for the general trait; a^{S1} – a^{S2} = conditional slopes for specific traits 1-2; c_1-c_4 = intercepts; ECV = explained common variance by factor; $IECV_G$ = item explained common variance for the general trait; $IECV_{S1}$ –$IECV_{S2}$= item explained common variance for specific traits 1-2; ECV values may not sum to 1 due to rounding error; $*a^G$ = marginal slope for general trait; $*a^{S1}$ - $*a^{S2}$ = marginal slopes for specific traits 1-2.

Table 4.29

Cross-Validation of Unidimensional Model Item Parameter Estimates for the 5-item Mastery Subscale of the PALS Classroom Goal Structures Scale (Mastery1 removed)

	Factor loading	Slope	Threshold				Item-fit statistic	
Item	Λ	A	b_1	b_2	b_3	b_4	S-χ^2	*p*
Mastery2	0.75 (.03)	1.95 (.09)	-2.35 (.11)	-1.49 (.07)	-0.54 (.04)	0.53 (.05)	117.4	<.0001
Mastery3	0.81 (.02)	2.38 (.11)	-2.05 (.08)	-1.40 (.06)	-0.53 (.04)	0.44 (.04)	109.6	<.0001
Mastery4	0.84 (.02)	2.67 (.13)	-1.95 (.08)	-1.29 (.05)	-0.49 (.04)	0.42 (.04)	98.5	<.0001
Mastery5	0.78 (.02)	2.12 (.09)	-2.18 (.10)	-1.36 (.06)	-0.31 (.04)	0.75 (.05)	126.8	<.0001
Mastery6	0.69 (.03)	1.61 (.08)	-2.25 (.11)	-1.50 (.07)	-0.47 (.05)	0.65 (.05)	108.7	<.0001

Note. λ = full information factor loadings; a = item slope (discrimination); b_1-b_4 = thresholds, S-χ^2 = item-fit statistic, *p* = p value associated with the item-fit statistic. Values in parentheses are standard error estimates.

Table 4.30

Cross-Validation of Unidimensional Model Item Parameter Estimates for the 3-item Performance Approach Subscale of the PALS Classroom Goal Structures Scale

	Factor loading	Slope	Threshold				Item-fit statistic	
Item	Λ	A	b_1	b_2	b_3	b_4	S-χ^2	*p*
Papp1	0.80 (.03)	2.25 (.13)	-2.43 (.11)	-1.64 (.07)	-0.83 (.05)	0.08 (.04)	51.9	.0039
Papp2	0.78 (.03)	2.10 (.11)	-2.36 (.10)	-1.54 (.07)	-0.50 (.04)	0.61 (.04)	102.1	<.0001
Papp3	0.83 (.03)	2.53 (.15)	-2.17 (.09)	-1.49 (.06)	-0.67 (.04)	0.25 (.04)	86.7	<.0001

Note. λ = full information factor loadings; a = item slope (discrimination); b_1-b_4 = thresholds, S-χ^2 = item-fit statistic, *p* = p value associated with the item-fit statistic. Values in parentheses are standard error estimates.

Table 4.31

Cross-Validation of Unidimensional Model Item Parameter Estimates for the 4-item Performance Avoid Subscale of the PALS Classroom Goal Structures Scale (Pavoid5 is dropped due to LD issues)

	Factor loading	Slope	Threshold				Item-fit statistic	
Item	Λ	A	b_1	b_2	b_3	b_4	S-χ^2	*p*
Pavoid1	0.70 (.03)	1.67 (.10)	-1.61 (.09)	-0.65 (.06)	0.44 (.06)	1.40 (.09)	74.3	.0021
Pavoid2	0.75 (.03)	1.96 (.11)	-1.15 (.07)	-0.32 (.05)	0.58 (.05)	1.34 (.08)	94.7	<.0001
Pavoid3	0.73 (.03)	1.84 (.10)	-1.51 (.09)	-0.62 (.06)	0.41 (.05)	1.39 (.08)	97.1	<.0001
Pavoid4	0.82 (.03)	2.43 (.07)	-1.24 (.07)	-0.52 (.05)	0.23 (.05)	0.93 (.06)	70.6	.0020

Note. λ = full information factor loadings; a = item slope (discrimination); b_1-b_4 = thresholds, S-χ^2 = item-fit statistic, *p* = p value associated with the item-fit statistic. Values in parentheses are standard error estimates.

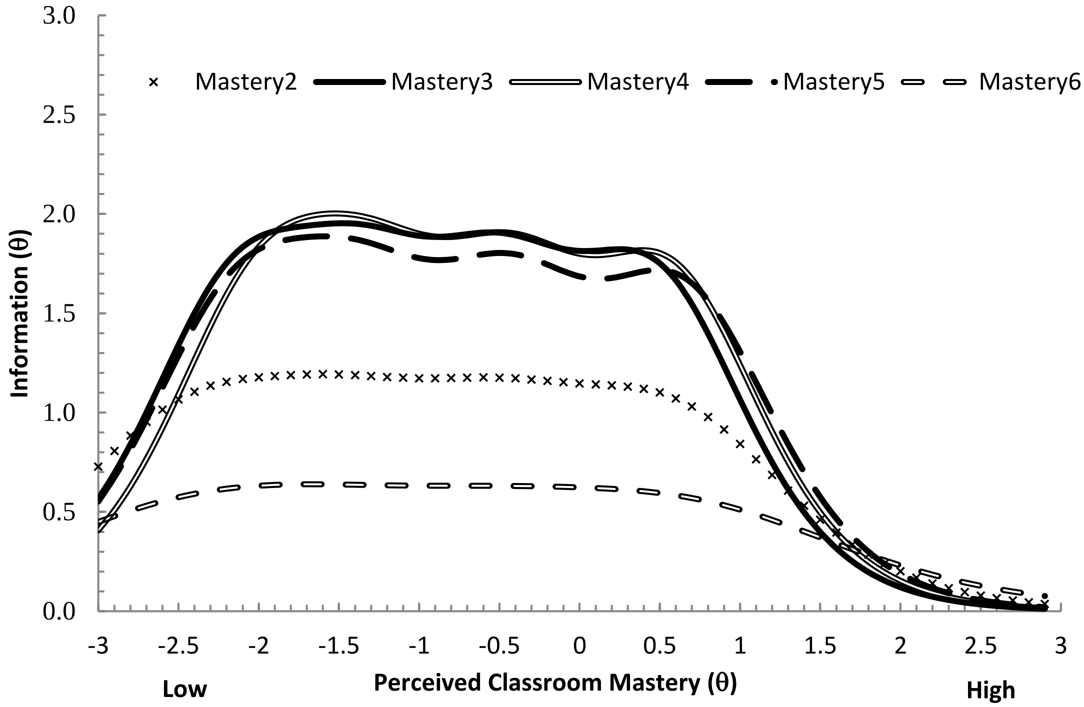

Figure 4.1. Item information functions for the 5 items composing the reduced unidimensional Mastery subscale of the PALS perceived classroom goal structures scale (Mastery1 was dropped due to positive LD issues with Mastery2).

Note. Each function represents the amount of information available from a given item across the potential range of thetas.

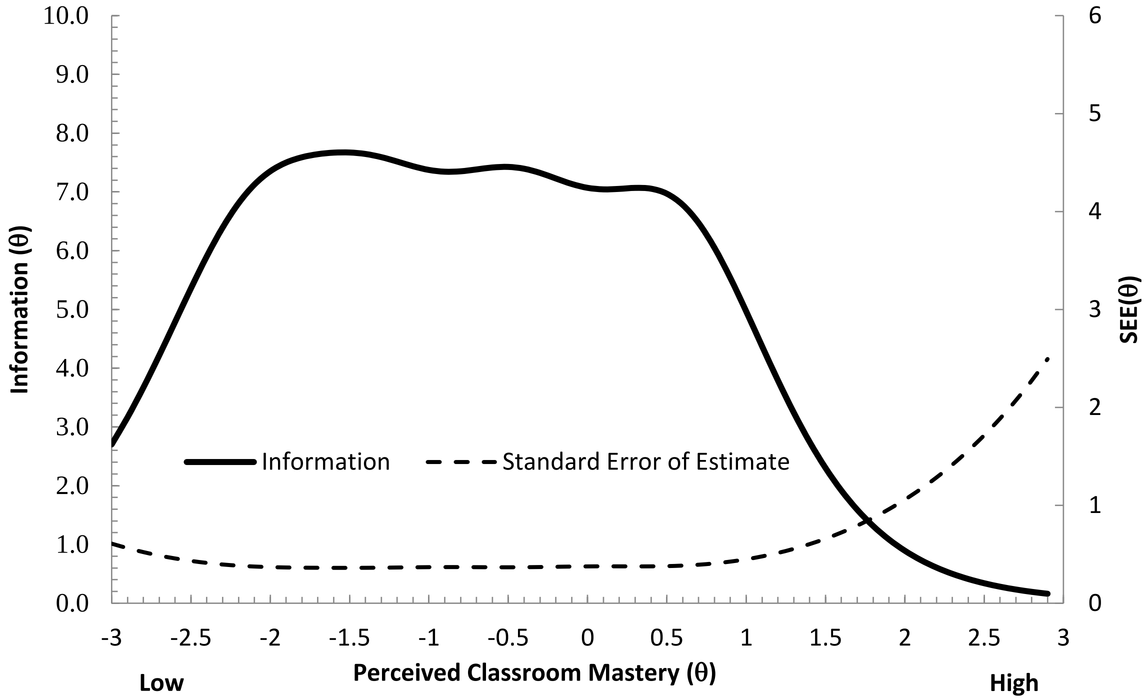

Figure 4.2. Total information function and expected standard error of the estimate [SEE] function for the reduced unidimensional Mastery subscale of the PALS perceived classroom goal structures scale.

Note. The SEE and information functions are related in that SEE is approximately equivalent to 1 divided by the square root of the Information at any given point on the latent trait continuum.

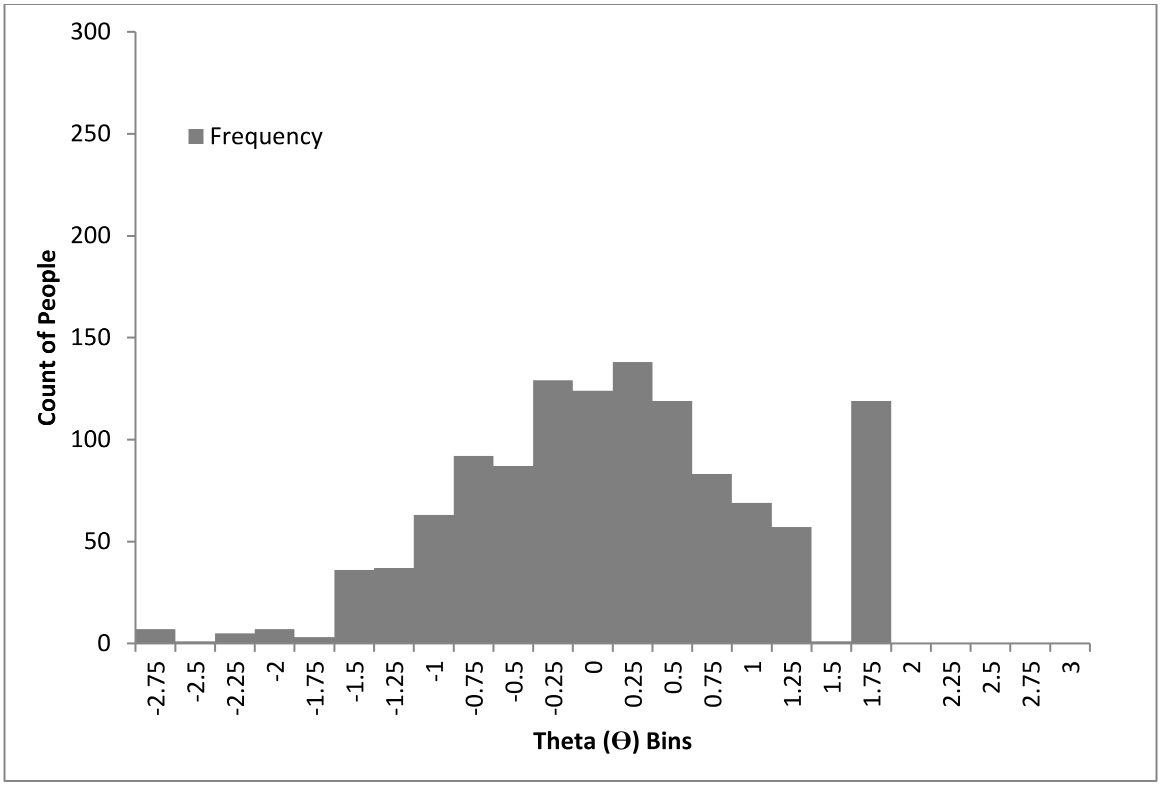

Figure 4.3. Frequency of people in sample A by estimated latent trait level on the reduced unidimensional mastery subscale of the PALS perceived classroom goal structures scale.

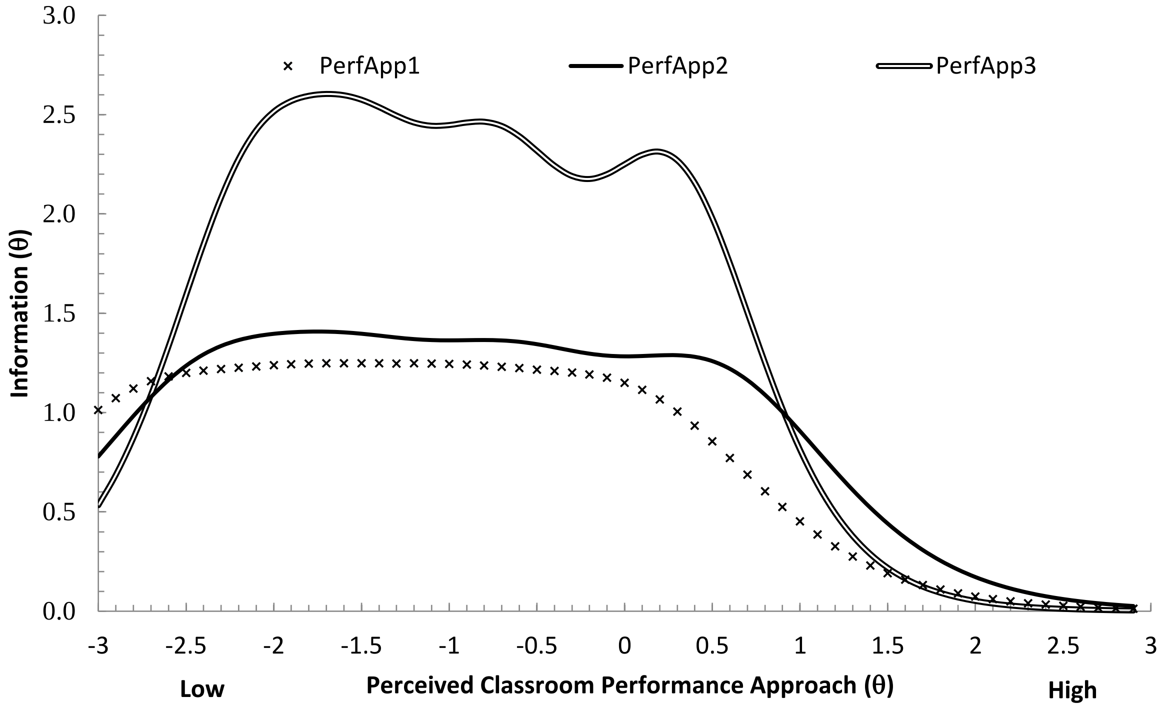

Figure 4.4. Item information functions for the 3 items composing the unidimensional Performance Approach subscale of the PALS perceived classroom goal structures scale.
Note. Each function represents the amount of information available from a given item across the potential range of thetas.

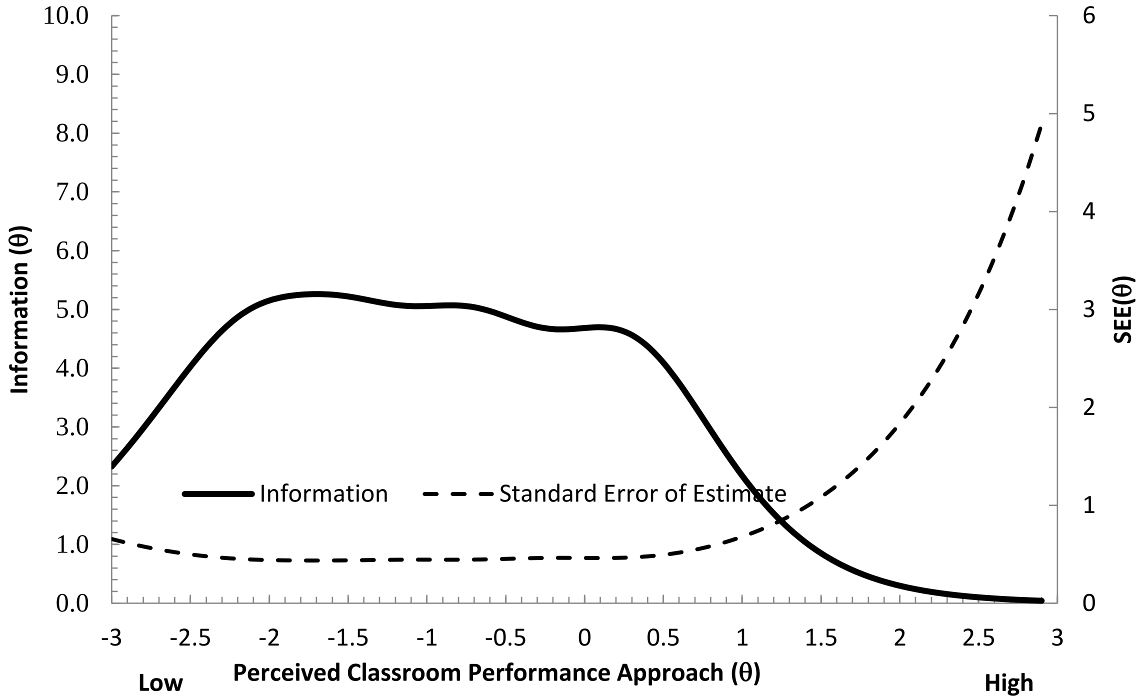

Figure 4.5. Total information function and expected standard error of the estimate [SEE] function for the unidimensional Performance Approach subscale of the PALS perceived classroom goal structures scale.

Note. The SEE and information functions are related in that SEE is approximately equivalent to 1 divided by the square root of the Information at any given point on the latent trait continuum.

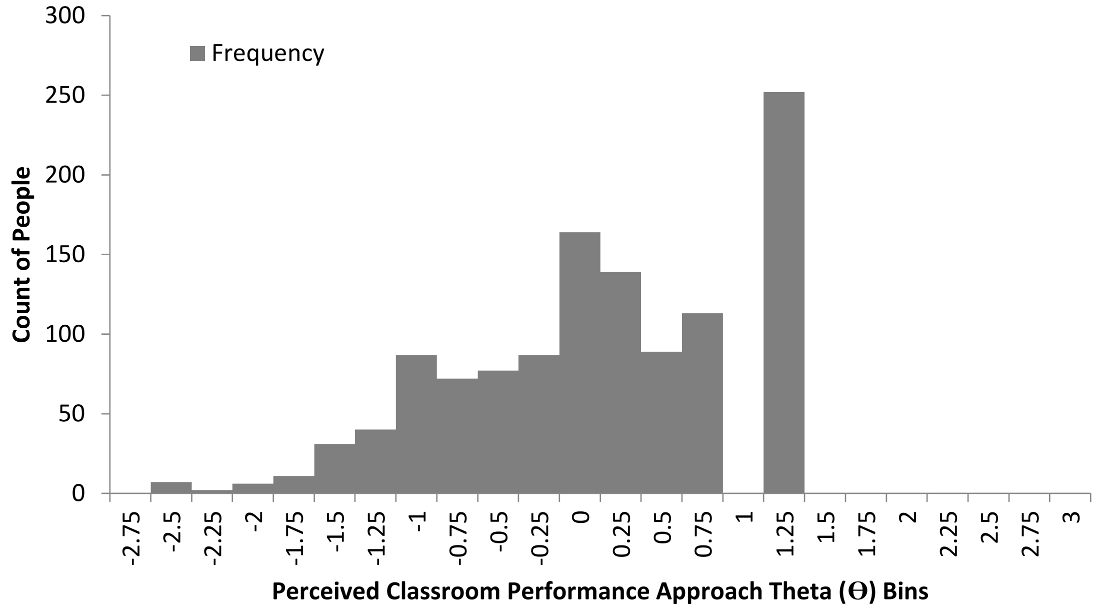

Figure 4.6. Frequency of people in Sample A by estimated latent trait level on the unidimensional perceived classroom Performance Approach subscale of the PALS perceived classroom goal structures scale.

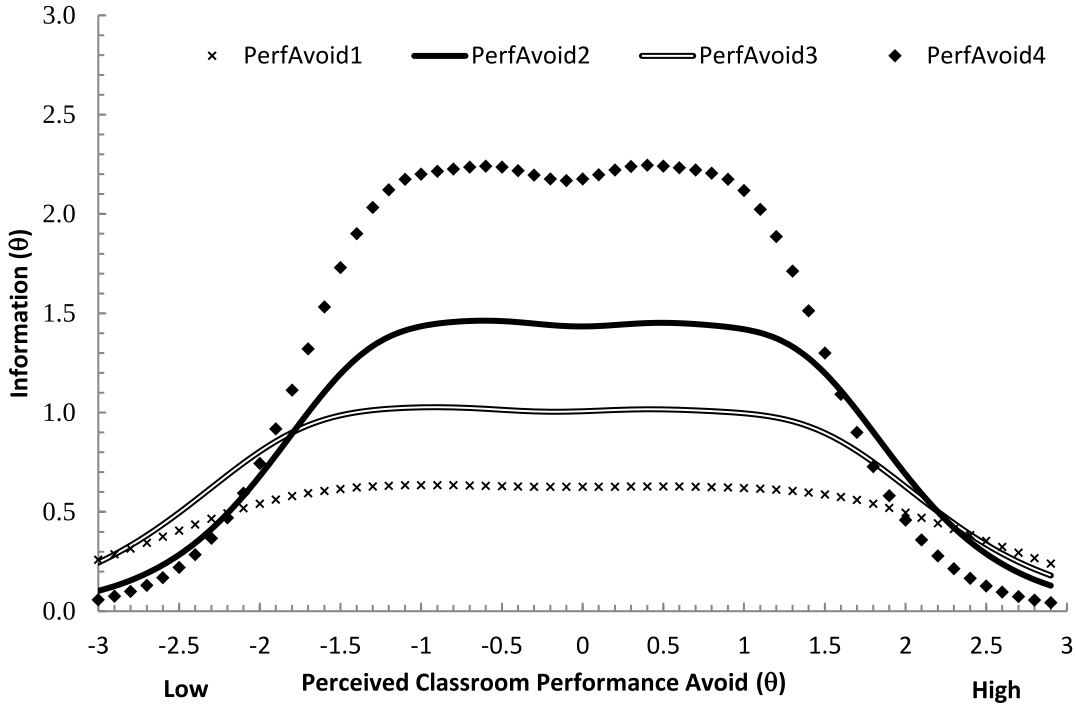

Figure 4.7. Item information functions for the 4 items composing the reduced unidimensional Performance Avoid subscale of the PALS perceived classroom goal structures scale (PAvoid5 was dropped due to positive LD issues with PAvoid1).
Note. Each function represents the amount of information available from a given item across the potential range of thetas.

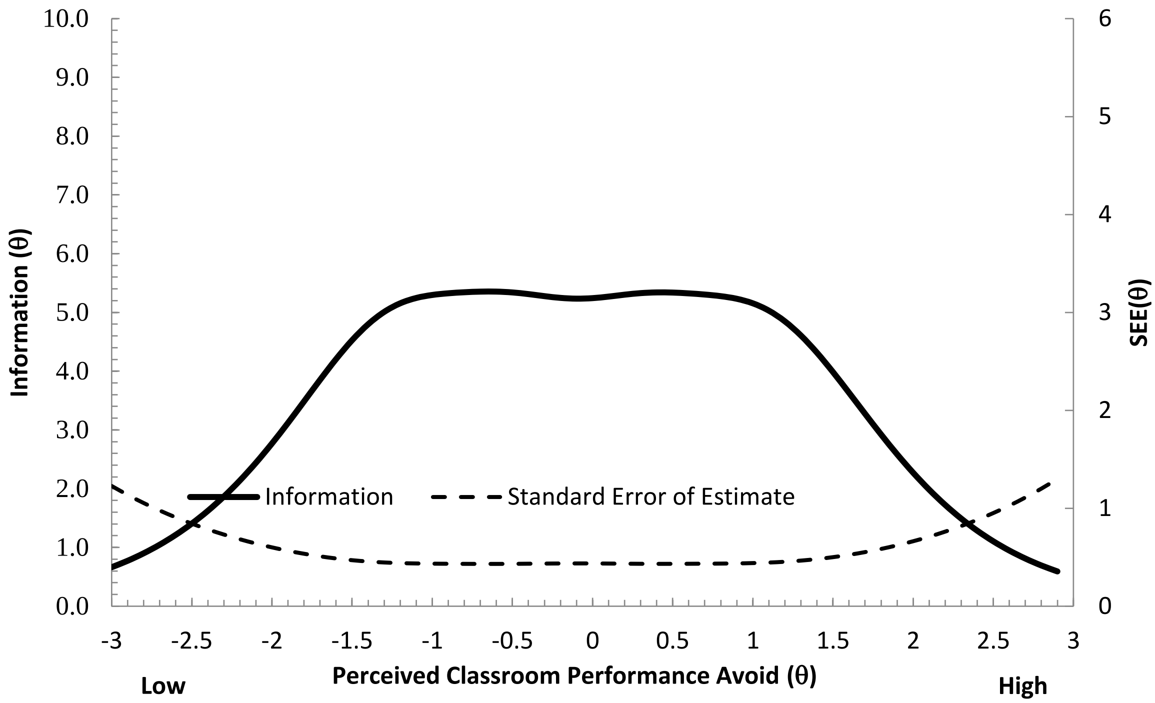

Figure 4.8. Total information function and expected standard error of the estimate [SEE] function for the reduced unidimensional Performance Avoid subscale of the PALS perceived classroom goal structures scale.
Note. The SEE and information functions are related in that SEE is approximately equivalent to 1 divided by the square root of the Information at any given point on the latent trait continuum.

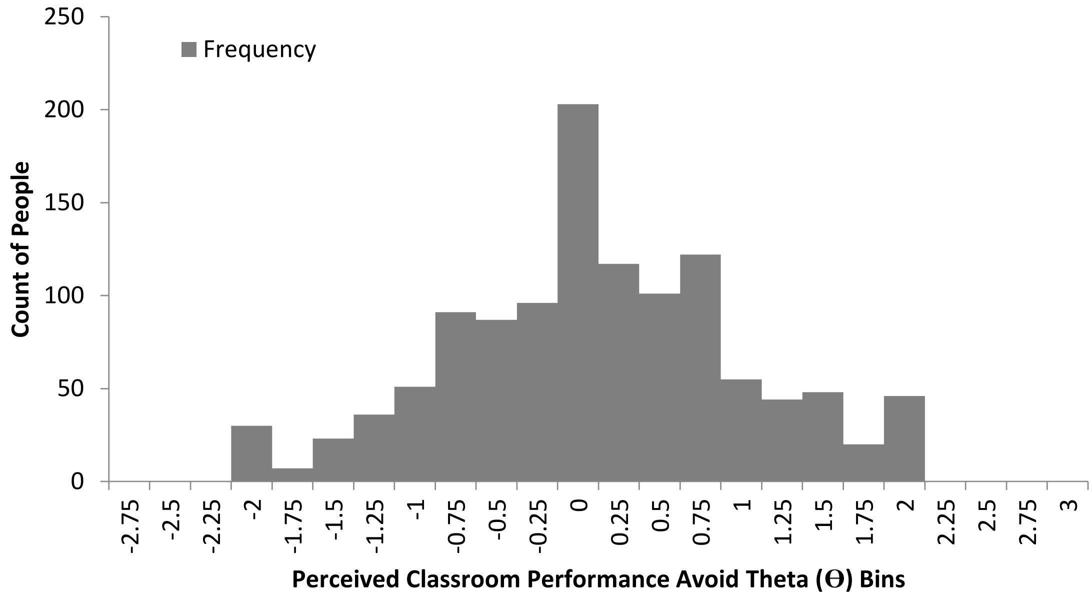

Figure 4.9. Frequency of people by estimated latent trait level on the reduced unidimensional perceived classroom Performance Avoid subscale of the PALS perceived classroom goal structures scale.
Note. Sample A was used to construct this figure.

Chapter Five: Discussion

Based on results from the initial and validation samples, a three dimensional correlated trait structure in line with current theory showed the best fit out of the plausible models. Although demonstrating best overall fit, the high construct intercorrelations displayed in the MIRT models and the within item multidimensionality apparent in the bifactor models are suggestive of problems that need addressing. The following discussion will first address the practical implications of the results found before moving to broader speculation on potential causes and implications of these research findings.

Practical Implications for the Classroom Goal Structures Scales

The drill-down method of moving from correlated traits, to bifactor, to unidimensional models allowed for more insight into both the functioning of individual items on these scales as well as the relationships between these scales. One of the clearest conclusions is that the performance approach and mastery subscales, while not tapping the same construct, possess a great deal of overlap. Model C, the three-factor correlated traits model estimated a correlation of approximately .82 between the latent traits for both samples. Although a correlation of this magnitude is possible while constructs remain distinct, the bifactor models suggest that within item multidimensionality is present. Particularly, mastery1, ‘trying hard is important,’ showed a strong relationship to the performance approach construct which helped to drive the general factor in the bifactor models. This same mastery item showed positive LD issues in both samples, suggesting that the item itself may be problematic. It seems plausible that ‘trying hard’ could be seen as a component of getting good grades as well as attempting to understand the material, thus tapping both mastery and performance approach latent traits.

Whether this item is dropped in future revisions to this scale or not, researchers using the scale in its current instantiation should prepare to handle this issue. One method would be including both performance approach and mastery scales in a study even if the hypothesis being tested is only referencing one of the two subscales. Much as socio-economic status is frequently used as a control for academic performance, performance approach and mastery need to both be assessed so that the analysis can partial out the influence of one construct when attempting to assess the independent influence of the other construct. Although the latent traits of performance approach and performance avoid had an estimated correlation of .46 in the initial sample and .49 in the validation sample, the bifactor models showed minimal and ignorable within-item multidimensionality. Performance avoid showed consistently stronger loadings on the specific factor across both samples. Thus, the inclusion of performance avoid as a control for either performance approach or mastery appears to be less essential.

Drilling down from overall construct level to item level revealed some further concerns with the overall scales. Although it is difficult to get reliable estimates of a construct with small item numbers, some tentative conclusions can be reached. Negative LD was prevalent between item pairs across all models, frequently occurring in more than 50% of item pairs in any given model. This pattern may be suggestive of item redundancy, or at least interpretation of the items by respondents as querying overly similar attitudes. Whether this can be remedied depends on whether the construct is broad enough to allow development of more unique items aligned with the construct of interest. It is plausible that the current similarity of items is due to development from a CTT perspective which rewards selecting items that have similar proportion of agreement through increased coefficient alpha values (Reise, Horan, & Blanchard, 2011). With IRT, scale construction is based on intended use through the utilization of item slopes and threshold values. If a scale is intended to provide precision across the full range of the latent trait spectrum, items are selected which have large discriminations (factor loadings) but target different areas of the latent trait spectrum. If a scale is intended to provide maximum precision at a desired cut score, items are selected with large discriminations, which target the same area of the latent trait spectrum to ensure maximum information at that point. The more common construction for attitudinal measures would be to create a uniform precision range across the theta spectrum

(Embretson & Reise, 2000). A revision to the perceived classroom goal structures scales, using IRT and expanding the number of items intended to tap each construct, could facilitate better coverage of the theta spectrum.

In terms of precision, two of the three scales were composed of items too easy to agree with for a large portion of respondents. Both of the approach scales, performance approach and mastery, showed optimal precision for those respondents who were average or below average on the latent trait. With the current structure of the subscales, a researcher could attempt to dichotomize respondents into categories based on the latent trait; however, any attempts to differentiate individuals who score towards the top of the spectrum on these motivational constructs would be potentially fraught with a lack of precision, making it difficult to detect effects which may exist. In order to increase confidence in score differentiations for those above average on these latent traits, items that are harder to agree will need to be added. One example of a harder item for the mastery subscale could be as follows: In this class, it is important to continue working on a problem even after getting a good score to ensure understanding of the solution. Although this item risks assessing an excess of mastery orientation when compared to performance orientation, such a juxtaposition may be needed to create items that are difficult to agree with.

Potential Implications for Classroom Goal Structures

Although the previous section details the specifics about handling the statistical and measurement issues inherent in the current scale, the theoretical implications of this research are also worth discussing. The large overlap in variance between mastery and performance approach could be symptomatic of several potential scenarios, including: 1) these constructs are truly distinct but the current scale is not reflective of this, 2) these constructs were once distinct, but are no longer distinct, or 3) these constructs never were distinct in a high school population. These scenarios will be examined in order followed by some final thoughts on the approach/avoid distinction and potential item phrasing effects.

Mastery and performance approach subscales at the classroom level may be truly distinct, but fail to reflect this distinction due to previously identified issues such as ceiling effects with the items being too easy to differentiate amongst high school students in this sample. A common ceiling effect present in both of these subscales could be masking differences that would be discernible if harder items were implemented, causing artificially inflated correlations and apparent within item multidimensionality. Another potential cause of shared variance that could distort the findings is social desirability. The main dimensionality validation of these scales was performed on sixth graders while this sample consisted of high school students (Midgley et al., 2000). High school students may have an enhanced awareness that both mastering a problem and achieving good grades is expected in school environments, exacerbating any social desirability effects found when beginning sixth grade.

Another plausible reason for the current overlap in variance found between mastery and performance approach is that these constructs were once distinct but are no longer perceived to be separate. Seventeen years have passed since publication of this version of the PALS scales and thirty years have passed since the inception of this theory. In this time period, we have seen classrooms attempt to integrate mastery-style frameworks while the Federal government has emphasized testing and performance with the No Child Left Behind Act (e.g., Ames, 1992). The emphasis on mastery while in classrooms may have been only a background shadow to the idea that the most important representation of learning is passing the test. These dual forces may have led to an intertwining of these two classroom goal structures that was not present at the inception of the theory.

A related possibility is that these constructs may never have been orthogonal by the time a student entered high school. High school students are probably aware that the decisive indicator of success is high school completion and potentially college acceptance, two goals which are ultimately determined by performance. Once a student reaches this point in his academic career, mastery may typically be the penultimate goal, serving as an intermediary to ensure strong test performance. Engaging with high school students on the reasons that they select different goal orientations may shed light on this process.

Another possibility is that the approach and avoid dimensions are more prominent than the mastery and performance dimensions. The results of the bifactor analyses show the strong possibility of an approach dimension which encompasses both mastery and performance approach. As mentioned previously, the approach/avoid dimension was added to the scale to clarify some of the common confusion found when researching performance goal orientations, which related to a host of both positive and negative outcomes depending on the research study and scales used; however, it has yet to be seriously contemplated whether the approach/avoid dimension is more potent at the classroom level than the initial theory it is being used to clarify. One component of goal orientations is affect in the form of transitory emotions, but selection of personal goal orientations or perception of classroom goal structures may be partially attributable to mood, a more stable and global emotional state. Mood, inherently a long-term version of affect, has been related to many of the same correlates encompassed within goal orientations such as behaviors, embodied through individual goal selection and persistence, and cognitions, embodied through working memory capacity and perceived self-efficacy (Langens, 2009). At the classroom level, a student may perceive an environment in which one is either approaching success or avoiding failure rather than selecting what success in that environment entails. This is a potential avenue for further research with implications for both classroom environment and practice.

A final possibility is that the variance overlap between mastery and performance approach is an artifact of item construction. Both mastery and performance approach items consist of positive statements whereas performance avoid items consist of negative statements. Previous research has suggested that positive and negative phrasing can alter the psychometric properties of an instrument: it is indeterminate whether this is because the construct is being tapped differently through phrasing, the actual construct being tapped through negative statements is different, or because people interpret the term "not" in a unique fashion (Benson & Hocevar, 1985; Spector, Van Katwyk, Brannick, & Chen, 1997). One option is to rephrase some performance avoid items in a positive fashion. For example, Pavoid5 could be modified to positive phrasing as follows: in this class, one of the main goals is to looking like you can do the work. Pavoid 3 could also be readily modified as follows: in this class, it's important to do as well as other students. Future research can explore whether the dimensionality found in this study is replicated when items have been modified to address phrasing issues. If the results hold even when phrasing effects have been eliminated, then the theoretical structure of classroom goal structures may need revisiting.

Future Measurement Directions

Overall, these scales performed in line with theory in terms of the three subscales being plausibly distinct. However, the high factor correlations and the potential within item multidimensionality revealed through the bifactor analyses suggest a need for further study. Using a cross-sectional snapshot of students as found in this sample allows one to determine that excess covariance is found between dimensions, but not the root cause. Multiple potential causes have been discussed and potential avenues for future research suggested to clarify the reason for this within item multidimensionality.

A caveat to the discussion thus far is the limitations of this study, by both data collection and currently available techniques. A multilevel IRT model would have been desirable to model the nature of students being nested within classrooms; however, the information linking students to classrooms was lacking in the data used for this analysis. Future research is needed to account for this potential nesting. It is possible that incorporating a multi-level model may shed light on the LD issues and suggest alternative revision directions or even that minimal revision is needed for these scales. Another future direction suggested by this study entails research into performance of the S-χ2 statistic for item-fit (Orlando & Thissen, 2000). Simulation studies frequently focus on dichotomous items and ability testing, limiting the generalizability to polytomous attitudinal measures where guessing is not a plausible scenario (eg, Orlando & Thissen, 2003). With the passage of the Every Student Succeeds Act, it seems likely that attitudinal measures will become increasingly high-stakes and demand strong psychometric evaluation.

Final Conclusions

There are some definitive avenues for exploration and more psychometric analysis seems necessary to determine the reason for the dimensionality results found in this study. Ultimately, classroom goal orientations does not currently perform in the expected fashion. Further investigation is needed to determine whether the previously described high dimension correlations and within-item multidimensionality should be attributed to item issues or theoretical issues. Further research is necessary to clarify the questions raised in this analysis.

Appendix A: Glossary of Terms

AGQ: Achievement Goal Questionnaire
AIC: Akaike Information Criteria

BIC: Bayesian Information Criteria

CFA: confirmatory factor analysis
CFI: Comparative Fit Index
CTT: classical test theory

ECV: explained common variance
EFA: exploratory factor analysis

FA: factor analysis or factor analytic

GFI: Goodness of Fit Index
GR: graded response (model)

IECV: item-based explained common variance
IFA: item factor analysis
IIF: item information function
IRF: item response function
IRT: item response theory

LD: local dependency
LI: local (or conditional) independence

LRT: likelihood ration test

MIRT: Multidimensional Item Response Theory

PALS: Patterns of Adaptive Learning Scales
PCA: principal components analysis
PL: parameter logistic

RMSEA: Root Mean Square of Error Approximation

SEM: structural equation modeling

TIF: total information function
TLI: Tucker-Lewis Index

References

Ames, C. (1984). Achievement attributions and self-instructions under competitive and individualistic goal structures. *Journal of Educational Psychology, 76*, 478-487. doi:10.1037/0022-0663.76.3.478

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology, 84,* 261-271. doi:10.1037//0022-0663.84.3.261

Ames, C. & Archer, J. (1988). Achievement goals in the classroom: Students' learning strategies and motivational processes. *Journal of Educational Psychology, 80,* 260-270. doi:10.1037/0022-0663.80.3.26

Ames, C. & Archer, J. (1987). Mothers' beliefs about the role of ability and effort in school learning. *Journal of Educational Psychology, 79,* 409-414. doi:10.1037/0022-0663.79.4.409

Anderman, E. M., Griesinger, T., & Westerfield, G. (1998). Motivation and cheating during early adolescence. *Journal of Educational Psychology, 90,* 84-93. doi:10.1037/0022-0663.90.1.84

Anderman, E. M., & Midgley, C. (2002). Methods for studying goals, goal structures, and Patterns of Adaptive Learning. In Midgley, C. (Ed.), *Goals, goal structures, and patterns of adaptive learning* (1-20). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Anderman, E. M., & Midgley, C. (1997). Changes in achievement goal orientations, perceived academic competence, and grades across the transition to middle-level schools. *Contemporary Educational Psychology, 22,* 269-298. doi:10.1006/ceps.1996.0926

Anderman, E. M., Urdan, T., & Roeser, R. (2003, March). *The Patterns of Adaptive Learning survey: History, development, and psychometric properties.* Paper presented at the Indicators of Positive Development Conference, Washington, D.C.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42 (Supplement 1)*, 7-16. Retrieved from http://www.jstor.org/stable/4640697

Benson, J, & Hocevar, D. (1985). The impact of item phrasing on the validity of attitude scales for elementary school children. *Journal of Educational Measurement, 22,* 231-240. Retrieved from http://www.jstor.org/stable/1435036

Birnbaum, A. (2008). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores*. United States of America: Information Age Publishing. (Original work published 1968)

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice, 16(10),* 21-32. doi:10.1111/j.1745-3992.1997.tb00605.x

Bong, M. (2004). Academic motivation in self-efficacy, task value, achievement goal orientations, and attributional beliefs. *Journal of Educational Research, 97,* 287-297. Retrieved from http://www.jstor.org/stable/27548044

Borsboom, D. (2005). *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics.* New York: Cambridge University Press.

Brophy, J. R. (1999). Toward a model of the value aspects of motivation in education: developing appreciation for particular learning domains and activities. *Educational Psychologist, 34(2)*, 75-85. doi:10.1207/s15326985ep3402_1

Button, S. B., Mathieu, J. E., & Zajac, D. M. (1996). Goal orientation in organizational research: A conceptual and empirical foundation. *Organizational Behavior and Human Decision Processes, 67,* 26-48. doi:10.1006/obhd.1996.0063

Cai, L. (2015). flexMIRT®: Flexible multilevel multidimensional item analysis and test scoring (version 3.0.3) [computer software]. Chapel Hill, NC. Vector Psychometric Group, LLC.

Cai L., du Toit S. H. C., Thissen D. (2011). *IRTPRO: User guide*. Lincolnwood, IL: Scientific Software International.

Cai, L., & Monroe, S. (2014). A new statistic for evaluating item response theory models for ordinal data. CRESST Report 839. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST).*

Chen W-H., Thissen D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22,* 265-289. doi:10.3102/10769986022003265

Church, M. A., Elliot, A. J., & Gable, S. L. (2001). Perceptions of classroom environment, achievement goals, and achievement outcomes. *Journal of Educational Psychology, 93,* 43-54. doi:10.1037//0022-0663.93.1.43

Chomsky, N. (1959). A review of BF Skinner's Verbal Behavior. *Language*, *35*(1), 26-58.

Coombs, C. (1950). Psychological scaling without a unit of measurement. *Psychological Review, 57,* 145-158. doi:10.1037/h0060984

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation*, *10(7)*, 1-9.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334. doi:10.1007/BF02310555

de Ayala, R. J. (2009). The theory and practice of item response theory. New York: The Guilford Press.

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. International Journal of Testing, 13, 354- 398. doi:10.1080/15305058.2013.799067

DeMars, C. E. (2006). Application of the bifactor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement, 43,* 145-168. Retrieved from http://www.jstor.org/stable/20461818

Deemer, S. (2004). Classroom goal orientation in high school classrooms: Revealing links between teacher beliefs and classroom environments. *Educational Research, 46,* 73-90. doi:10.1080/0013188042000178836

DeVellis, R. F. (2006). Classical test theory. *Medical Care, 44 (Supplement 3),* S50-S59. doi:10.1097/01.mlr.0000245426.10853.30

Diener D. I. & Dweck, C. S. (1978). An analysis of learned helplessness: Continuous changes in performance, strategy, and achievement cognitions following failure. *Journal of Personality and Social Psychology, 36,* 451-462. doi:10.1037/0022-3514.36.5.451

Duda, J. L. & Nicholls, J. G. (1992). Dimensions of achievement motivation in schoolwork and sport. *Journal of Educational Psychology, 84,* 290-299. doi:10.1037/0022-0663.84.3.290

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist, 41,* 1040-1048. doi:10.1037/0003-066X.41.10.1040

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review, 95,* 256-273. doi:10.1037//0033-295X.95.2.256

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Qual Life Res, 16 (Supplement 1),* 5-18. doi:10.1007/s11136-007-9198-0

Elliot, A. J. (1994). Approach and avoidance achievement goals: An intrinsic motivation analysis (doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. 9428343)

Elliot, A. J. (1999). Approach and avoidance motivation and achievement goals. *Educational Psychologist, 34,* 169-189. doi:10.1207/s15326985ep3403_3

Elliot, A. J. (2005). A conceptual history of the achievement goal construct. In Elliot, A. J., & Dweck, C. S. (Eds.), *Handbook of competence and motivation* (52-72). The Guilford Press: New York, NY.

Elliot, A. J., & Church, M. A. (1997). A hierarchical model to approach and avoidance achievement motivation. *Journal of Personality and Social Psychology, 72,* 218-232. doi:10.1037/0022-3514.72.1.218

Elliot, A. J., & Harackiewicz, J. M. (1996). Approach and avoidance achievement motivation goals and intrinsic motivation: A mediational analysis. *Journal of Personality and Social Psychology, 70,* 461-475. doi:10.1037/0022-3514.70.3.461

Elliot, A. J., & McGregor, H. (2001). A 2x2 achievement goal framework. *Journal of Personality and Social Psychology, 80,* 501-519. doi:10.1037//0022-3514.80.3.501

Elliot, E. S., & Dweck, C. S. (1988). Goals: An approach to motivation and achievement. *Journal of Personality and Social Psychology, 54,* 5-12. doi:10.1037/0022-3514.54.1.5

Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory*. Mahwah, New Jersey: Lawrence Earlbum, Inc.

Finney, S. J., Pieper, S. L., & Barron, K. E. (2004). Examining the psychometric properties of the achievement goal questionnaire in a general academic context. *Educational and Psychological Measurement, 64,* 365-382. doi:10.1177/0013164403258465

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9,* 466-491. doi:10.1037/1082-989X.9.4.466

Flora, D. B., LaBrish, C., & Chalmers, R. P. (2012). Old and new ideas for data screening and assumption testing for exploratory and confirmatory factor analysis. *Frontiers in Psychology, 3,* 1-21. doi:10.3389/fpsyg.2012.00055

Ford, J. K., MacCallum, R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel Psychology, 39,* 291-314. doi:10.1111/j.1744-6570.1986.tb00583.x

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods, 14,* 275-299. doi:10.1037/a0015825

Greene, B. A., Miller, R. B., Crowson, M., Duke, B. L., & Akay, K. L. (2004). Predicting high school students' cognitive achievement: Contributions of classroom perceptions and motivation. *Contemporary Educational Psychology, 29,* 462-482. doi:10.1016/j.cedpsych.2004.01.006

Gutman, L. M. (2006). How student and parent goal orientations and classroom goal structures influence the math achievement of African Americans during the high school transition. *Contemporary Educational Psychology, 31,* 44-63. doi:10.1016/j.cedpsych.2005.01.004

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9,* 139-150. Retrieved from http://www.jstor.org/stable/2086306

Harackiewicz, J. M., Barron, K. E., Pintrich, P. R., Elliot, A. J., & Thrash, T. M. (2002). Revision of achievement goal theory: Necessary and illuminating. *Journal of Educational Psychology, 94,* 638-645. doi:10.1037//0022-0663.94.3.638

Harackiewicz, J. M., Durik, A. M., Barron, K. E., Linnenbrink-Garcia, L., & Tauer, J. M. (2008). The role of achievement goals in the development of interest: Reciprocal relations between achievement goals, interest, and performance. *Journal of Educational Psychology, 100,* 105-122. doi:10.1037/0022-0663.100.1.105

Helgado-Tello, F. P., Chacon-Moscoso, S., Barbero-Garcia, I., & Vila-Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality and Quantity, 44,* 153-166. doi:10.1007/s11135-008-9190-y

Henning, G. (1989). Does the Rasch model really work for multiple-choice items? Take another look: A response to Divgi. *Journal of Educational Measurement, 26,* 91-97. Retrieved from http://www.jstor.org/stable/1434626

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement, 66,* 393-416. doi:10.1177/0013164405282485

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2,* 41–54.

Houts, C. R., & Cai, L. (2015). flexMIRT: Flexible Multilevel Multidimensional Item Analysis and Test Scoring User's Manual Version 3.0. *Chapel Hill, NC: Vector Psychometric Group, LLC.*

Hulleman, C. S., Schrager, S. M., Bodman, S. M., & Harackiewicz, J. M. (2010). A meta-analytic review of achievement goal measures: Different labels for the same constructs of different constructs with the same labels? *Psychological Bulletin, 136,* 422-449. doi:10.1037/a0018947

Jagacinski, C. M., & Duda, J. L. (2001). A comparative analysis of contemporary achievement goal orientation measures. *Educational and Psychological Measurement, 61,* 1013-1039. doi:10.1177/00131640121971626

Jones, L. V., & Thissen, D. (2007). A history and overview of psychometrics. In C.R. Rao, & S. Sinharay (Eds.) *Handbook of statistics vol. 26* (pp. 1-22). Oxford: North Holland.

Kang, T., & Chen, T. T. (2008). An investigation of the performance of the generalized S-X^2 item-fit index for polytomous IRT models. *Journal of Educational Measurement, 45,* 391-406. doi:10.1111/j.1745-3984.2008.00071.x

Kaplan, A., Gheen, M., & Midgley, C. (2002). Classroom goal structure and student disruptive behavior. *British Journal of Educational Psychology, 72,* 191-211. doi:10.1348/000709902158847

Kaplan, A., & Maehr, M. L. (2007). The contributions and prospects of goal orientation theory. *Educational Psychology Review, 19,* 141-184. doi:10.1007/s10648-006-9012-5

Kaplan, A., & Middleton, M. J., Urdan, T., & Midgley, C. (2002). Achievement goals and goal structures. In Midgley, C. (Ed.), *Goals, goal structures, and patterns of adaptive learning* (21-53). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Karabenick, S. A. (2004). Perceived achievement goal structure and college student help seeking. *Journal of Educational Psychology, 96,* 569-581. doi:10.1037/0022-0663.96.3.569

Kopp, J. P., Zinn, T. E., Finney, S. J., & Jurich, D. P. (2011). The development and evaluation of the Academic Entitlement Questionnaire. *Measurement and Evaluation in Counseling and Development, 44,* 105-129. doi:10.1177/0748175611400292

Langens, T. A. (2009). Striving for personal goals: The role of negative mood and the availability of mood-regulation strategies. In Wosnitza, M., Karabenick, S. A., Efklides, A., & Nenniger, P. (Eds.), *Contemporary motivation research: From global to local perspectives* (359-377). Cambridge, MA: Hogrefe and Huber Publishing.

Lau, S., & Nie, Y. (2008). Interplay between personal goals and classroom goal structures in predicting student outcomes: A multilevel analysis of person-context interactions. *Journal of Educational Psychology, 100,* 15-29. doi:10.1037/0022-0663.100.1.15

Linnenbrink, E. A. (2005). The dilemma of performance-approach goals: The use of multiple goal contexts to promote students' motivation and learning. *Journal of Educational Psychology, 97,* 197-213. doi:10.1037/0022-0663.97.2.197

Linnenbrink, E. A., & Pintrich, P. R. (2002). Achievement goal theory and affect: An asymmetrical bidirectional model. *Educational Psychologist, 37(2),* 69-78. doi:10.1207/S15326985EP3702_2

Lord, F. M. (1952). *A theory of test scores.* Psychometric Monograph, No. 7.

Lord, F. M., & Novick, M. R. (2008). *Statistical Theories of Mental Test Scores.* United States of America: Information Age Publishing. (Original work published 1968)

Meece, J. L., Anderman, E. M., & Anderman, L. H. (2006). Classroom goal structure, student motivation, and academic achievement. *Annual Review of Psychology, 57,* 487-503. doi:10.1146/annurev.psych.56.091103.070258

Middleton, M. J., & Midgley, C. (1997). Avoiding the demonstration of a lack of ability: An unexpected aspect of goal theory. *Journal of Educational Psychology, 89,* 710-718. doi:10.1037/0022-0663.89.4.710

Midgley, C., Kaplan, A., Middleton, M., Maerh, M. L., Urdan, T., Anderman, L. H., et al. (1998). The development and validation of scales assessing students' achievement goal orientations. *Contemporary Educational Psychology, 23,* 113-131. doi:10.1006/ceps.1998.0965

Midgley, C., Maehr, M. L., Hruda, L. Z., Anderman, E., Anderman, L., Freeman, K. E., et al. (2000). *Manual for the patterns of adaptive learning scales (PALS).* Ann Arbor: University of Michigan.

Miller, A. D., & Murdock, T. B. (2006). Modeling latent true scores to determine the utility of aggregate student perceptions as classroom indicators in HLM: The case of classroom goal structures. *Contemporary Educational Psychology, 32,* 83-104. doi:10.1016/j.cedpsych.2006.10.006

Muis, K. R., Winne, P. H., & Edwards, O. V. (2009). Modern psychometrics for assessing achievement goal orientation: A Rasch analysis. *British Journal of Educational Psychology, 79,* 547-576. doi:10.1348/000709908X383472

Muthen, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38,* 171-189.

Nicholls, J. G. (1975). Causal attributions and other achievement-related cognitions: Effects of task outcome, attainment value, and sex. *Journal of Personality and Social Psychology, 31,* 379-389. doi:10.1037/h0076505

Nicholls, J. G. (1984). Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance. *Psychological Review, 91*, 328-346. doi:10.1037/0033-295X.91.3.328

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. Psychometrika, 44, 443-460. doi:10.1007/BF02296207

Osborne, J. W., Costello, A. B., & Kellow, J. T. (2008). Best practices in exploratory factor analysis. In J. W. Osborne (Ed), *Best practices in quantitative methods* (pp. 205-213). Thousand Oaks, California: Sage Publishing.

Orlando M., Thissen D. (2000). Likelihood-based item fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24,* 50-64. doi:10.1177/01466216000241003

Orlando M., Thissen D. (2003). Further investigation of the performance of S-χ2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27,* 289-298. doi:10.1177/0146621603027004004

Payne, S. C., Youngcourt, S. S., & Beaubien, J. M. (2007). A meta-analytic examination of the goal orientation nomological net. *Journal of Applied Psychology, 92*, 128-150. doi:10.1037/0021-9010.92.1.128

Pintrich, P. R. (2000a). An achievement goal theory perspective on issues in motivation terminology, theory, and research. *Contemporary Educational Psychology, 25*, 92-104. doi:10.1006/ceps.1999.1017

Pintrich, P. R. (2000b). Multiple goals, multiple pathways: The role of goal orientation in learning and achievement. *Journal of Educational Psychology, 92,* 544-555. doi:10.1037//0022-0663.92.3.544

Pintrich, P. R., Conley, A. M., & Kempler, (2003). Current issues in achievement goal theory and research. *International Journal of Educational Research, 39*, 319-337. doi:10.1016/j.ijer.2004.06.002

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*, 667-696. doi:10.1080/00273171.2012.715555

Reise, S. P., Horan, W. P., & Blanchard, J. J. (2011). The challenges of fitting an item response theory model to the Social Anhedonia Scale. *Journal of Personality Assessment, 93*, 213-224. doi:10.1080/00223891.2011.558868

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scores. *Journal of Personality Assessment, 92*, 544-559. doi:10.1080/00223891.2010.496477

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcome measures. *Quality of Life Research, 16*, 19-31. doi:10.1007/s11136-007-9183-7

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). *Factor analysis and scale revision. Psychological Assessment, 12*, 287-297. doi:10.1037//1040-3590.12.3.287

Roeser, R. W., Midgley, C., & Urdan, T. C. (1996). Perceptions of the school psychological environment and early adolescents' psychological and behavioral functioning in school: The mediating roles of goals and belonging. *Journal of Educational Psychology, 88*, 408-422. doi:10.1037/0022-0663.88.3.408

Rupp, A. A., & Zumbo, B. D. (2004). A note on how to quantify and report whether IRT parameter invariance holds: When Pearson correlations are not enough. Educational and *Psychological Measurement, 64*, 588-599. doi:10.1177/0013164403261051

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores (psychometric monograph no. 17)*. Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment, 29*, 304-321. doi:10.1177/0734282911406653

Schunk, D. H., Pintrich, P. R., & Meece, J. L. (2008). *Motivation in education: Theory, research, and applications (3rd ed.).* New Jersey: Pearson Education, Inc.

Sharkness, J., & DeAngelo, L. (2011). Measuring student involvement: A comparison of classical test theory and item response theory. *Research in Higher Education, 52,* 480-507. doi:10.1007/s11162-010-9202-3

Sideridis, G. D. (2009). Normative vs. non-normative performance goals: Effects of behavioral and emotional regulation in achievement situations. In Wosnitza, M., Karabenick, S. A., Efklides, A., & Nenniger, P. (Eds.), *Contemporary motivation research: From global to local perspectives* (321-338). Cambridge, MA: Hogrefe and Huber Publishing.

Skinner, B. F. (1953). *Science and human behavior*. New York: The Free Press.

Spector, P. E., Van Katwyk, P. T., Brannick, M. T., & Chen, P. Y. (1997). When two factors don't reflect two constructs: How item characteristics can produce artifactual factors. *Journal of Management, 23,* 659-677. doi:10.1016/S0149-2063(97)90020-9

Sperry, R. W. (1993). The impact and the promise of cognitive revolution. *American Psychologist, 48,* 878-885. doi:10.1037/0003-066X.48.8.878

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103,* 677-680. doi:10.1126/science.103.2684.677

Steinberg, L. & Thissen, D. (2013). Item response theory. In J.S. Comer & P.C. Kendall (Eds.). *The Oxford Handbook of Research Strategies in Clinical Psychology.* New York, NY: Oxford University Press.

Streiner, D. L. (2010). Measure for measure: New developments in measurement and item response theory. *The Canadian Psychiatric Association Journal, 55,* 180-186. Retrieved from http://web.ebscohost.com/ehost/pdfviewer/pdfviewer?sid=02527cd5-0c41-468c-9c73-65ce843cd34d%40sessionmgr111&vid=2&hid=119

Stark, S. (2001). MODFIT: A computer program for model-data fit. Unpublished manuscript. University of Illinois: Urbana-Champaign.

Stark, S. (2002). MODFIT: Plot theoretical item response functions and examine the fit of dichotomous or polytomous IRT models to response data [computer program]. Department of Psychology, University of Illinois at Urbana-Champaign.

Stucky, B. D., & Edelen, M. O. (2015). Using hierarchical IRT models to create unidimensional measures from multidimensional data. In S. P. Reise & D. A. Revicki (Eds.) *Handbook of item response theory modeling: Applications to typical performance assessment.* New York, NY: Rutledge.

Ten Berge, J.M.F. & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69*, 613-625. doi:10.1007/BF02289858

Takane, Y., & deLeeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. Psychometrika, 52, 393-408. doi:10.1007/BF02294363

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*, 567-577. doi:10.1007/BF02295596

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 16, 433-451. doi:10.1037/h0073357

Thurstone, L. L. (1928). Scale construction with weighted observations. *Journal of Educational Psychology, 19*, 441-453. doi:10.1037/h0074895

Toland, M. D. (2014). Practical guide to conducting an item response theory analysis. *The Journal of Early Adolescence, 34*, 120-151. doi:10.1177/0272431613511332

Toland, M. D., Sulis, I., Giambona, F., Porcu, M., & Campbell, J. M. (2017). Introduction to bifactor polytomous item response theory analysis. *Journal of School Psychology, 60*, 41-63. doi:10.1016/j.jsp.2016.11.001

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11*, 1-13. doi:10.1007/BF02288894

Urdan, T. C. (2004). Predictors of academic self-handicapping and achievement: Examining achievement goals, classroom goal structures, and culture. *Journal of Educational Psychology, 96,* 251-264. doi:10.1037/0022-0663.96.2.251

Urdan, T. C. (1997). Achievement goal theory: Past results, future directions. In Maehr, M. L., and Pintrich, P. R. (Eds.), *Advances in motivation and achievement* (Vol. 10, 99-141). Greenwich, CT: Jai Press, Inc.

Urdan, T., & Midgley, C. (2001). Academic self-handicapping: What we know, what more there is to learn. *Educational Psychology Review*, *13,* 115-138. doi:10.1023/A:1009061303214

Urdan, T., & Ryan, A. M., Anderman, E. M., & Gheen, M. H. (2002). Goals, goal structures, and avoidance behaviors. In Midgley, C. (Ed.), *Goals, goal structures, and patterns of adaptive learning* (55-84). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Urdan, T., & Schoenfelder, E. (2006). Classroom effects on student motivation: Goal structures, social relationships, and competence beliefs. *Journal of School Psychology, 44,* 331-349. doi:10.1016/j.jsp.2006.04.003

Vandewalle, D. (1997). Development and validation of a work domain goal orientation instrument. *Educational and Psychological Measurement, 57,* 995-1015. doi:10.1177/0013164497057006009

Zickar, M. J. & Broadfoot, A. B. (2009). The partial revival of a dead horse: Comparing classical test theory and item response theory. In C.E. Lance, & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legend (p. 37-60).* New York: Routledge.

Weiner, B. (2010). The development of an attribution-based theory of motivation: A history of ideas. *Educational Psychologist*, *45*, 28-36. doi:10.1080/00461520903433596

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current research and future directions. *Psychological Methods*, *12,* 58-79. doi:10.1037/1082-989X.12.1.58

Wolters, C. A. (2004). Advancing achievement goal theory: Using goal structures and goal orientations to predict students' motivation, cognition, and achievement. *Journal of Educational Psychology, 96,* 236-250. doi:10.1037/0022-0663.96.2.236

Vita

Angela Kristi Tombari

EDUCATION

B. S., **Psychology**, Texas A&M University, College Station, TX. Graduated 2006, summa cum laude.

PERTINENT EMPLOYMENT EXPERIENCE

Research Analyst, KCEWS, January 2015 to present
Webmaster, College of Education Information Systems, Fall 2014
Statistical Consultant, Evaluation Center, University of Kentucky, 2013 to 2014
Statistician and Webmaster, Institute for Educational Research, University of Kentucky, 2010 to 2013
Statistics Consultant, Grockit, San Francisco, CA, 2009 to 2010

TEACHING EXPERIENCE

Guest Lecturer, KHP Intro to Research Methods, "Introduction to Survey Methods", University of Kentucky, 2013-2014
Teaching Assistant, Item Response Theory, University of Kentucky, Spring 2014
Teaching Assistant, Research Methods, University of Kentucky, Fall 2008/Fall 2013/Spring 2014
Teaching Assistant, Intermediate Statistics, Spring 2009
Instructor EDP203, Teaching Students with Exceptionalities, University of Kentucky, 2007-2009
Private Tutor, Lexington, KY, 2008-2014
Mathematics Instructor, Tutoring Club, Lexington, KY, Summer 2008
Instructor Middle School Mathematics, Hearne Junior High, Hearne, TX, 2006-2007
Mathematics Instructor, Sylvan Learning Center, College Station, TX, 2006-2007
Trumpet Instructor, Music Consulting Unlimited, College Station, TX, 2006-2007
Supplemental Instruction Leader, Texas A&M University, College Station, TX, 2005-2006
Mathematics Tutor, Kingwood High School, Kingwood, Texas, 2001-2002

PUBLICATIONS & PRESENTATIONS

Boggs, B., Kirchner, T., Moore, M., Peach, H, Ryan, C., Schroeder, G., Sharma, M., **Tombari, A. K.**, & Walker, A (March, 2017*).* *Acting as one: The power of a state-wide EPP and cross-agency collaborative*. Presentation at the annual conference of the American Association of Colleges for Teacher Education, Tampa, FL.

Secamiglio, S., Mensah, R. K., Cunningham, J., **Tombari, A. K.**, and McGhee, D. (October, 2016). *Benefits of the Kentucky longitudinal data system on education and workforce outcomes.* Presentation at the annual conference of the Mid-west Educational Research Association, Evanston, Illinois.

Tombari, A. K., Cunningham, J., & Mensah, R. K. (October, 2016). *Evaluating the measurement properties to guide revision of a new teacher feedback survey.* Presentation at the annual conference of the Mid-west Educational Research Association, Evanston, Illinois.

Tombari, A. K., Secamiglio, S., Cunningham, J., & Eifler, B. (2016, July). *Inner workings and products of the Kentucky longitudinal data system.* Presentation at the annual National Center for Educational Statistics STATS-DC conference, Washington DC.

Morales, A, Toland, M. D., Little II, D. L., Weisenhorn, D. A., **Tombari, A. K.,** Li, Z., & Rostosky, S. S. (2014). *First item response theory analysis: A new measure of perceived discrimination for sexual minority Latinas*. Submitted 08-20-2014

Bradley, K. D., Snyder, E., & **Tombari, A. K** (in preparation*). Assessing psychometric properties of higher education end-of-course evaluations.* Journal of Studies in Educational Evaluation.

Tombari, A. K., & Henchy, A. M. (2014, August). *Survey research: Making sure that you are hearing what the students are really saying.* Presentation at the annual Student Affairs Assessment and Research Conference, Columbus, OH.

Danner, F., **Tombari, A. K.,** & Toland, M. D. (2011, October). *Associations between the onset of regular smoking and trajectories of depressive symptoms from adolescence through young adulthood*. Paper to be presented at the 5th conference on Emerging Adulthood, Providence, RI.

Tombari, A. K., Toland, M. D., & Hastings, E. H. (2011, April). *Parallel analysis with ordinal data under complex structure.* Paper presented at the annual conference of the National Council on Measurement in Education, New Orleans, LA.

Li, Z., Toland, M. D., & **Tombari, A. K.** (2011, April). *Comparison of the CES-D factor structure using three estimation methods: Bayesian, MLM, and WLSMV*. Paper presented at the annual Spring Research Conference, Cincinatti, OH.

Henchy, A. M., Toland, M. D., **Tombari, A. K.**, & Piercey, R. R. (2010, August). *A review and assessment of reliability generalization research*. Poster presented at the annual convention of the American Psychological Association, San Diego, CA.

Shukla, S. Y., **Tombari, A. K.**, Toland, M. D., & Danner, F. W. (2015). At-home parental support for learning and high school students' academic motivation and persistence in mathematics. *Journal of Educational and Developmental Psychology, 5(1),* 44-56. doi:10.5539/jedp.v5n1p44.