Brigham Young University
BYU ScholarsArchive

Theses and Dissertations

2021-04-02

A Psychometric Analysis of the Precalculus Concept Assessment

Brian Lindley Jones
Brigham Young University

Follow this and additional works at: https://scholarsarchive.byu.edu/etd

 Part of the Education Commons

BYU ScholarsArchive Citation
Jones, Brian Lindley, "A Psychometric Analysis of the Precalculus Concept Assessment" (2021). *Theses and Dissertations*. 8918.
https://scholarsarchive.byu.edu/etd/8918

This Dissertation is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact ellen_amatangelo@byu.edu.

www.manaraa.com

A Psychometric Analysis of the Precalculus Concept Assessment

Brian Lindley Jones

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Richard R. Sudweeks, Chair
Joseph A. Olsen
Ross A. A. Larsen
Dawn Teuscher

Educational Inquiry, Measurement, and Evaluation

Brigham Young University

Copyright © 2021 Brian Lindley Jones

All Rights Reserved

www.manaraa.com

ABSTRACT

A Psychometric Analysis of the Precalculus Concept Assessment

Brian Lindley Jones
Educational Inquiry, Measurement, and Evaluation, BYU
Doctor of Philosophy

The purpose of this study was to examine the psychometric properties of the Precalculus Concept Assessment (PCA), a 25-item multiple-choice instrument designed to assess student reasoning abilities and understanding of foundational calculus concepts (Carlson et al., 2010). When this study was conducted, the extant research on the PCA and the PCA Taxonomy lacked in-depth investigations of the instruments' psychometric properties. Most notably was the lack of studies into the validity of the internal structure of PCA response data implied by the PCA Taxonomy. This study specifically investigated the psychometric properties of the three reasoning constructs found in the PCA taxonomy, namely, Process View of Function (R1), Covariational Reasoning (R2), and Computational Abilities (R3).

Confirmatory Factor Analysis (CFA) was conducted using a total of 3,018 pretest administrations of the PCA. These data were collected in select College Algebra and Precalculus sections at a large private university in the mountain west and one public university in the Phoenix metropolitan area. Results showed that the three hypothesized reasoning factors were highly correlated. Rival statistical models were evaluated to explain the relationship between the three reasoning constructs. The bifactor model was the best fitting model and successfully partitioned the variance between a general reasoning ability factor and two specific reasoning ability factors. The general factor was the dominant factor accounting for 76% of the variance and accounted for 91% of the reliability. The omegaHS values were low, indicating that this model does not serve as a reliable measure of the two specific factors.

PCA response data were retrofitted to diagnostic classification models (DCMs) to evaluate the extent to which individual mastery profiles could be generated to classify individuals as masters or non-masters of the three reasoning constructs. The retrofitting of PCA data to DCMs were unsuccessful. High attribute correlations and other model deficiencies limit the confidence in which these particular models could estimate student mastery.

The results of this study have several key implications for future researchers and practitioners using the PCA. Researchers interested in using PCA scores in predictive models should use the General Reasoning Ability factor from the respecified bifactor model or the single-factor model in conjunction with structural equation modeling techniques. Practitioners using the PCA should avoid using PCA subscores for reasoning abilities and continue to follow the recommended practice of reporting a simple sum score (i.e., unit-weighted composite score).

Keywords: factor analysis, Diagnostic Classification Models (DCMs), calculus, mathematics education

www.manaraa.com

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to Dr. Richard Sudweeks my dissertation chair, mentor, and friend. This dissertation would not have been possible without his guidance and insightful questioning. I feel privileged to have worked with Dr. Sudweeks in many capacities while in the EIME program. As I move forward in life, I hope to emulate his insatiable appetite to learn and grow. I also extend my sincere thanks to the rest of my committee. To Dr. Joseph Olsen for his Socratic teaching and deep knowledge of factor analysis that helped me when I got stuck. To Dr. Ross Larsen for his encouraging demeanor and ability to make the most complicated statistical methods understood by all. To Dr. Dawn Teuscher for introducing me to the PCA, providing data, and helping me navigate the domain of Mathematics Education. I gratefully acknowledge the influence of all the faculty that have taught me over the years. Equally influential were my classmates and graduate cohorts. I miss our lively discussions in the classroom and lab. Thank you for inviting me to see and approach the world in a new way.

If I were allowed, I would add my wife and four children's names to this project and degree. They have only ever known a husband and father who is also a student. One of our young children once said, "I want to be an engineer, or maybe a clown when I grow up, but if that doesn't work, I'll just do what dad does and go to school." One of my most cherished memories will be kneeling next to a child and hearing them pray the words "bless dad to pass his tests." Someday my children may know that many of my tests were not the kind administered by a teacher but by life. When they do, I hope they know that God answered their prayers and sustained me throughout this journey. I undoubtedly feel a profound sense of gratitude for the "widow's mite" and all those who have sacrificed to support this institution, particularly the EIME program. Their sacrifice has resulted in one of my most reverenced blessings.

www.manaraa.com

TABLE OF CONTENTS

www.manaraa.com

www.manaraa.com

www.manaraa.com

LIST OF TABLES

www.manaraa.com

LIST OF FIGURES

www.manaraa.com

LIST OF EQUATIONS

www.manaraa.com

CHAPTER 1

Introduction

The Precalculus Concept Assessment (PCA) is a 25-item multiple-choice instrument designed to assess a student's reasoning abilities and understanding of foundational calculus concepts (Carlson et al., 2010). Carlson and her colleagues developed the instrument through a series of research studies, which resulted in the PCA Taxonomy formation (Appendix A). This taxonomy identifies reasoning abilities and conceptual understandings essential for a student's success in learning calculus. Items on the PCA instrument were developed to align with the PCA Taxonomy. Carlson et al. identified several possible uses for the PCA, including "(a) assessing student learning in college algebra and precalculus, (b) comparing the effectiveness of various curricular treatments, and (c) determining student readiness for calculus." (2010, p. 113). The primary focus of this research was to examine the factor structure of the reasoning abilities portion of the PCA Taxonomy.

The original publication describing the PCA (Carlson et al., 2010) details an admirable instrument development process. The iterative process of item development and refinement provides strong evidence for the content validity of the instrument. It is no surprise that the majority of publications citing the PCA reference the underlying theory of the PCA based on the PCA Taxonomy. The current *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA] et al., 2014), which from this point forward will be referred to as the *Standards,* list five general types of validity evidence that may be used to build a case for the validity of test use and interpretation. These five pieces include:

- Evidence based on test content
- Evidence based on response processes

www.manaraa.com

- Evidence based on the internal structure
- Evidence based on relations to other variables
- Evidence for the consequences of testing

The *Standards* emphasized that each of the above forms of evidence is not required in all settings. However, that evidence should be selected based on the evidence's appropriateness to support test interpretation and use. The *Standards* noted that "a sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses." (AERA et al., 2014, p. 21). As the *Standards* suggested, specific uses and intended interpretations of PCA results are required to build a robust validity argument. It may be inferred from Carlson et al. (2010) that PCA scores are intended to represent the degree to which students have mastered a "composite effect of the reasoning and understandings on student abilities" (p. 137) outlined in the PCA Taxonomy. As previously cited, Carlson et al. (2010) suggested that PCA scores could be used to (a) assess student learning, (b) investigate curricular interventions, and (c) gauge calculus readiness. Before writing this dissertation, the primary evidence to support the validity of these interpretations and uses of PCA results has rested solely on instrument content-based evidence.

Statement of Problem

The Precalculus Concept Assessment (PCA) is well-developed by many test development standards (AERA et al., 2014; Haladyna, 2004; Lane, Raymond, & Haladyna, 2016; Miller et al., 2013). Carlson et al. (2010) should be commended on developing the PCA Taxonomy and its use for carefully guiding and defining concepts assessed by the PCA. Furthermore, the item writing and refinement based on student interviews supplied substantial evidence for the test's content

www.manaraa.com

validity (AERA et al., 2014, p. 82; M. C. Rodriguez, 2016). Multiple-choice response options were well researched through the preliminary use of open-ended questions to identify common misconceptions. Plausible distractors were developed from this research, and response options were further refined or eliminated based on an in-depth distractor analysis.

Although items were developed with great care, Carlson et al. (2010) noted "that there are significant and complex interactions among the subcategories [of the PCA Taxonomy] so that no one subcategory can be completely isolated" (p. 119). They further described an individual student's score on the PCA as "a broad indicator of reasoning abilities and understandings relative to the PCA Taxonomy" (p. 137). An in-depth analysis of articles citing Carlson et al. (2010) found no published articles that empirically investigated the dimensionality of the test to provide evidence to support the extent to which the "complex interactions among subcategories" are manifested in the data.

An empirical investigation of the internal structure of PCA response data has the potential to produce empirical evidence to support the calculation and use of a single score or multiple subscores (Standards 1.13, 1.14, 1.15; AERA et al., 2014). The common practice of limiting PCA results to a simple sum score (i.e., unit-weighted composite score) may be less appropriate and informative if psychometric analyses yield evidence that the PCA is multidimensional.

Statement of Purpose

There were two overarching purposes for conducting this research. First, this research empirically investigated the dimensionality of PCA response data in relation to the three reasoning abilities specified in the PCA Taxonomy. The purpose of this portion of the research was to examine empirical evidence of the internal structure to inform the methods by which the

www.manaraa.com

PCA scores are interpreted and reported (i.e., single total score or multiple subscores). Second, this research investigated the use of diagnostic classification modeling techniques as a means for providing fine-grained diagnostic information (i.e., mastery profile) for each student regarding their mastery of the three reasoning abilities.

Research Questions

This study focused on the following research questions:

1. To what extent does a confirmatory factor analysis of PCA pretest data provide evidence that supports the validity of the three-factor structure implied by the PCA Taxonomy?
2. If the three first-order factors are found to be highly correlated, to what extent do rival models (i.e., a single-factor model, a second-order factor model, or a bifactor model) fit better than the three first-order factors model and illuminate the interrelationships among the three first-order factors?
3. How successfully can the PCA response data be retrofitted for an analysis using a general diagnostic classification model (DCM)?
4. How does the adequacy of a DCM model based on the factor structure implied by the PCA Taxonomy compare with a DCM model based on the CFA results?

www.manaraa.com

CHAPTER 2

Review of Literature

The literature review presented in this dissertation consists of a synthesis of research on (a) a general approach to test development, (b) the Precalculus Concept Assessment (PCA), (c) the application of factor analysis for the use of instrument evaluation, and (d) the use of diagnostic classifications models for the analysis of assessment data.

General Test Development

Test development that supports the validity of score uses and interpretation requires a systematic approach. The Handbook of Test Development contains a detailed description of the test development process. This development process can be summarized in a framework of 12 coordinated components. "Each of these 12 components can be used to provide a framework for collecting and organizing evidence to support the psychometric quality of the test and the validity of the test score interpretations and uses" (Lane, Raymond, Haladyna, & Downing, 2016, p. 3).

Twelve Test Development Components:

- Overall Plan
- Domain Definition and Claims Statements
- Content Specifications
- Item Development
- Test Design and Assembly
- Test Production
- Test Administration
- Scoring
- Cut Scores
- Test Score Reports
- Test Security
- Test Documentation

www.manaraa.com

The overall plan serves as a guide for navigating the other components of the development process. The plan should clearly articulate the test objectives and outline the steps for gathering validity evidence required to support the intended score interpretations and uses. Defining the test's domain and content is one of the significant considerations for test development, particularly achievement tests. This component of test development may be considered the keystone of the test development process. Lane, Raymond, Haladyna, and Downing (2016) asserted that

> The effectiveness of all other test development activities relies on how well the domain is defined, and claim statements are delineated. The validity of test score interpretations and uses rest on the adequacy and defensibility of the methods used to define the domain and claim statements and the successful implementation of procedures to systematically and sufficiently sample the domain (p. 6).

Miller et al. (2013) emphasized that in addition to defining the test domain in general terms, a specific focus should be given to portions of the domain related to the identified goals and objectives of the test. Priority should be given to salient portions of the domain, reflecting the importance of goals and objectives.

Test content specifications include both the content to be sampled from the domain of interest as well as the response format (i.e., "mechanism that a test taker uses to respond to a test item"; AERA et al., 2014, p. 223). The *Standards* (AERA et al., 2014, p. 15) noted that other aspects of the test content, including cognitive processes and response type, should be included as part of the content specifications. Miller et al. (2013) recommended using a table of specifications to specify the test's content. The table of specification establishes the relationship between the subject-matter content and the instructional objectives. Many tables of specification

www.manaraa.com

form a two-way table with rows representing the subject-matter content and the columns representing various levels of cognitive complexity (e.g., Bloom's Taxonomy). The entries in the two-way table cells specify the number of test items designed to measure the intersection of content and cognitive complexity. The table of specifications can be used to furnish evidence that the test adequately samples content from the domain of interest.

As part of the test content specifications, identifying salient constructs helps to ensure appropriate construct representation. Identifying salient constructs also aids in identifying the presence of factors that may be ancillary or irrelevant to the constructs of interest. A thorough test development process will employ specifications that will limit construct underrepresentation and construct-irrelevant variance. Evidence for construct validity is produced mainly in the test development phase. Miller et al. (2013) suggested a focused approach to gathering evidence for construct validity by emphasizing evidence that is reasonable to gather and is relevant to the specified uses and interpretations of test scores.

Item development includes both the item type and item writing. Item types should be selected to adequately measure test constructs with enough fidelity to stand as validity evidence. Selecting item types is often influenced by extraneous factors related to cost, scoring, and administration time. Lane, Raymond, Haladyna, and Downing (2016) recommended that "the methods and procedures used to produce effective items are a major source of validity evidence for all testing programs. Complete documentation of these steps is essential" (p. 8).

Subject-matter experts commonly write item content. Test developers often take the role of guiding item writers to ensure the quality of items. Miller et al. (2013) listed eight recommendations for item writing:

1. Use test and assessment specifications as a guide.

www.manaraa.com

2. Write more items and tasks than needed.
3. Write the items and tasks well in advance of the testing date.
4. Write each test item and assessment task so that the task to be performed is clearly defined, and it calls forth the performance described in the intended learning outcome.
5. Write each item or task at an appropriate reading level.
6. Write each item or task not to provide help in responding to other items or tasks.
7. Write each item so that the answer is one that would be agreed on by experts or, in the case of assessment tasks, the responses judged excellent would be agreed on by experts. (pp. 164–165)

Following these general recommendations can help avoid some of the more prominent errors in item writing. However, more specific recommendations related to specific item types can be found in the test development literature. Poorly constructed items introduce construct-irrelevant variance. An external item review by experts not involved in item development can supply valuable feedback for improving item quality. Piloting items is another practice used to improve item quality. Item pilots, or field tests, allow the test developer to (a) perform preliminary item analysis of item difficulty, (b) discrimination, (c) differential item functioning, and (d) investigate relationships with other items. Other forms of item review include cognitive interviews or think-aloud protocols.

The test development process also includes the development and design of scoring and reporting procedures. As noted by Lane, Raymond, Haladyna, and Downing (2016), “Reporting test results is considered to be one of the most essential activities of the test development process because the way in which results are reported can either enhance or jeopardize valid score

www.manaraa.com

interpretations and uses." (p. 15). The purpose for which a test is administered and the score reporting should be aligned. For example, a classroom test developed to help students identify knowledge gaps should report test results in a way that clearly articulates these gaps.

In contrast, a normative test designed to compare student performance should clearly communicate student performance relative to other students. If subscores are used, the *Standards* (AERA et al., 2014) have noted the need to provide rational and relative evidence for the use and interpretation of subscores. Likewise, the use of a composite score should also be supported by empirical evidence for the composite score's interpretation and use.

Concept Inventories in the Sciences

A concept inventory is a test explicitly developed to assess student conceptual understanding in or working knowledge of a particular domain. Concept inventories are occasionally referred to as tests of misconception. These tests often consist of a set of multiple-choice items where distractors are associated with domain-specific misconceptions. There are several varying approaches to developing a concept inventory; however, most approaches follow these general steps:

1. Research and development of taxonomies of conceptions and the identification of common misconceptions are developed.
2. Open-ended questions are constructed.
3. Open-ended questions are refined through cognitive interviews.
4. Multiple-choice distractors are developed.
5. Verification is conducted using with additional item testing using pilot items and cognitive interviews.

www.manaraa.com

The foundation of a concept inventory is the research supporting the identification and pertinent concepts and the associated misconceptions. The validity of a concept inventory may quickly come into question without adequate foundational research.

The Force Concept Inventory (Hestenes et al., 1992) is perhaps one of the most cited examples of a concept inventory. At the time of its publication, the Force Concept Inventory was the culmination of several years of research and the development of the Mechanics Diagnostic Test (Hestenes & Halloun, 1995; Hestenes & Wells, 1992). At the time, the approach used by Hestenes et al. (1992) was an innovation in the science field because it did not focus on student problem-solving skills. They proposed that the Force Concept Inventory be used as a general measure of Newtonian and non-Newtonian thinking. They also suggested the instrument could be used to diagnose misconceptions and to evaluate instructional practices.

An abundant number of concept inventories in the sciences were developed after the publication and popularization of the Force Concept Inventory. These inventories not only vary in their content domain but also in the quality of their development process. Table 1 provides a non-exclusive list of other concept inventories in the sciences. The instrument of focus for this dissertation, the Precalculus Concept Assessment (Carlson et al., 2010), followed the rigorous development process modeled by the Force Concept Inventory. Carlson et al. (2010) cited the Force Concept Inventory and the associated development process as the inspiration for the PCA's development. Although they did not follow the same naming convention, the PCA could be considered in the same class as other concept inventories.

Review of Precalculus Concept Assessment (PCA)

The Precalculus Concept Assessment (PCA) is an instrument developed by Carlson et al. (2010). The PCA was developed to measure students' reasoning abilities and understandings,

www.manaraa.com

Table 1

Examples of Concept Inventories in the Sciences

Field	Instrument Name	Reference
Biology	Conceptual Inventory of Natural Selection	Anderson et al., 2002
	Osmosis and Diffusion Conceptual Assessment	Fisher et al., 2011
	The Genetics Concept Assessment	Smith et al., 2008
	Genetics Literacy Assessment Instrument	Bowling et al., 2008
	Host Pathogen Interactions	Marbach-Ad et al., 2010
Physics and Astronomy	Star Properties Concept Inventory	Bailey et al., 2012
	Astronomy and Space Science Concept Inventory	Sadler et al., 2009
	Force Concept Inventory	Hestenes et al., 1992
	Statistics Concept Inventory	Steif & Dantzler, 2005
	Lunar Phases Concept Inventory	Lindell & Olsen, 2002
	Digital Logic Concept Inventory	Herman, 2011
Chemistry	Test to Identify Student Conceptualizations	Voska & Heikkinen, 2000
	The Mole Concept	Krishnan & Howe, 1994
	Chemistry Concepts Inventory	Mulford & Robinson, 2002
	ACID I	McClary & Bretz, 2012
	Enzyme–Substrate Interactions Concept Inventory	Bretz & Linenberger, 2012
	A Chemistry Concept Reasoning Test	Cloonan & Hutchinson, 2011
	Stereochemistry Concept Inventory	Leontyev, 2016
	Understanding Acids and Bases	Cetin-Dindar & Geban, 2011

Note. This table was adapted from (Leontyev, 2016)

www.manaraa.com

which comprise a foundation for learning calculus. This review of the PCA will first address the development of the PCA instrument, particularly item development and validation. Second, published articles citing the Carlson et al. (2010) article will be reviewed to ascertain the extent to which psychometric analyses of PCA response data have been conducted and the general use of the PCA in published research.

PCA Instrument Development

A thorough description of the PCA development was published by Carlson et al. (2010). They attributed their inspiration for developing the PCA to similar instruments from the domain of physics education research. The genesis of the PCA was original research on the essential knowledge and skills required for learning calculus. From this literature, Carlson et al. (2010) identified salient features of student reasoning abilities and mathematical understanding that contributed to students' success in calculus. These features were classified into reasoning abilities (i.e., process view of functions, covariational reasoning, computational abilities) and understandings (i.e., the meaning of function concepts, the growth rate of function types, and function representations). The PCA Taxonomy (Appendix A) provides a detailed organization of these classifications. The development of the PCA Taxonomy served as the foundation for the development of the PCA items.

A four-phase approach was used for the development of the PCA. Phase 1 and 2 spanned ten years of original research to identify understandings and reasoning abilities required for students to be successful in beginning calculus courses. These phases were initiated by Carlson's research on students' understandings of functions (1995), in which an initial taxonomy was developed to categorize students' function reasoning and understanding. The refinement and development of this taxonomy ultimately resulted in the PCA Taxonomy (Appendix A). This

www.manaraa.com

taxonomy was a guide for developing the PCA instrument. The taxonomy, similar to a table of specifications, provides evidence that the test adequately samples the content of the domain of interest.

The PCA Taxonomy highlights how each item was associated with reasoning ability, understanding, or a combination of the two. Very few items on the PCA are associated with a single construct from the taxonomy. However, all items measure a different combination of constructs from the taxonomy. Carlson et al. (2010) noted the complexities of the taxonomy, stating, "there are significant and complex interactions among the subcategories so that no one subcategory can be completely isolated" (p. 119). The initial set of 34 questions written following the PCA Taxonomy were open-ended questions used to further develop the PCA items.

Item refinement followed a process of administering the original 34 open-ended items to groups of students and conducting cognitive interviews. The analysis of the open-ended responses, coupled with the cognitive interviews, provided validity evidence to support the interpretation and use of scores from these items. These early activities also yielded a base of understanding for the students' common misunderstandings and incorrect responses. This information was then used in the development of multiple-choice distractors.

The primary focus of Phase 3 was on the development and validation of multiple-choice items. Eight cycles of data collection and refinement took place using an initial set of 25 multiple-choice PCA items. During this validation process, students were instructed to document their problem-solving processes. The documented student work was then used as data for qualitative analyses of student problem-solving processes. Researchers also conducted over 300 interviews to understand the cognitive processes used in conjunction with each item.

www.manaraa.com

Researchers' understanding of these cognitive processes informed item revisions, including multiple-choice distractors, and revisions to the PCA Taxonomy. A quantitative distractor analysis was conducted to select and refine multiple-choice distractors by calculating the percent of students who selected each item response option. Options selected by less than 5% of the student population were revised or removed. These test development activities provided additional evidence for internal content validity.

The final version of the PCA included 25 multiple-choice items with five response options for each item. Many of the items were context-dependent items that relied on a figure or graph to provide the context in which the student is to think. Each graph or figure was unique to each item in all but one item (i.e., no item sets). Several items were presented as a *story* or a *word* problem that tasked students with solving mathematical equations in the context of a specific situation. Other items presented students with a mathematical equation to solve.

The fourth and final phase of the PCA development involved examining the meaning of PCA scores. Due to the complexities of the PCA Taxonomy and the relatively short length of the assessment, Carlson et al. (2010) noted:

> The PCA assess the composite effect of the identified reasoning abilities and understandings on students' abilities. This approach is consistent with viewing the complex interactions among categories in the PCA Taxonomy as producing an emergent effect (Cohen et al., 1990) related to important precalculus reasoning abilities, rather than trying to establish independent uni-dimensional measures of underlying latent variables. (p. 137)

The reporting of a single composite score limits the use and interpretation of PCA scores to a general measure of student reasoning and understanding. Carlson et al. (2010) further emphasize

www.manaraa.com

that “it would not be appropriate to draw inferences about the abilities of an individual student relative to PCA subscores” (p. 137).

The primary source of evidence that the PCA measures the taxonomy constructs was supplied in the item development process and over 300 cognitive interviews. However, Carlson et al. (2010) also explored the relationship between PCA scores and course performance. They reported a correlation of 0.47 between students’ pretest PCA score and their final grade in first-semester calculus. They also reported that 77% of students with a score higher than 12 received a passing grade of an A, B, or C in their first-semester calculus course. These relationships between the PCA and calculus course outcomes produced evidence that PCA scores may be considered for use as a general predictor of future success. However, the generality of the composite score limits the ability to provide targeted feedback systematically. Carlson et al. (2010) did not report the use of other statistical methods such as factor analysis or diagnostic classification modeling as methods used to investigate the potential for reporting subscores. If supported by empirical evidence, the reporting of PCA subscores has the potential to provide more targeted formative feedback to students and faculty.

Publications Citing the PCA

A citation analysis was conducted to understand how researchers have used the PCA in their academic work and assess the degree to which validity evidence of the internal structure has been investigated. The citation analysis began by identifying publications that cited the article describing the PCA development (Carlson et al., 2010).

Several databases were used to identify citations including, Crossref, Web of Science, Scopus, and Google Scholar. At the time of writing this dissertation, a total of 82 unique citations were identified from sources ranging from peer-reviewed academic journals, records of

www.manaraa.com

conference proceedings, and works published on the web. Each citing article was reviewed to determine the nature of their reference(s) to the Carlson et al. (2010) article. Each section containing a reference to the PCA was read, and the nature of the reference was classified. A full list of citations and classifications can be found in Appendix B.

The citation analysis resulted in five general categories of references. The distribution of the 82 articles across these five categories is shown in Table 2. It is important to note that it was common for articles to have multiple references to the PCA, which allowed articles to be classified into multiple categories. The first category was *Reference to Theory*. An article was considered part of the *Reference to Theory* category if the author made one or more references to the theory discussed in the original PCA article. These references were either references to a specific theory or the more encompassing theory illustrated in the entire PCA Taxonomy. For example, Carlson et al. (2010) were referenced by LaRue (2017) as part of broad research on student's understanding of functions, and Bannerjee (2017) referenced the broad reasoning and understandings needed for success in calculus.

The second category was *Instrument Reference*. An article was categorized as an *Instrument Reference* if it made any reference to the PCA as an instrument. For example, Marfai (2016) referenced the PCA as an instrument when discussing a teacher's content mastery. Mejia-Ramos et al. (2017) referenced the PCA as an instrument in their literature review of undergraduate mathematics education assessments and as an exemplar in assessment construction.

The third category was the *Type of Test* category. Articles in the *Type of Test* category referenced the PCA as a type of test such as a test of conceptions or misconceptions. For example, Stanhope et al. (2017) wrote about developing a biological science quantitative

www.manaraa.com

reasoning exam, which referenced the PCA as a type of test used to measure students' reasoning abilities. M. Thomas and Lozano (2012) referenced the PCA as one of the many types of concept inventories.

The fourth category was *PCA Data*. Articles were placed in the *PCA Data* category if data from the PCA was used. This category typically manifested itself when the PCA was used as an outcome measure in a research study. For example, Cousino (2013) used PCA data as the outcome variable in a series of Bayesian models. Cromley et al. (2017) used a subset of PCA items from the *Understand Function Representations* from the PCA Taxonomy. They used these items to investigate students' abilities to coordinate multiple representations within the broader context of their investigation of the relationship between spatial skills and calculus proficiency.

The fifth category was the *Psychometric* category. Articles appearing in this category made references to one or more psychometric properties of the PCA. Only one article referenced the psychometric properties of the PCA. Zahner et al. (2017) referenced the reliability of the *Understand Functions Representations* item subset. Although they reported a Cronbach alpha of 0.692, they did not reference any analysis to test the assumptions needed for the appropriate use of Cronbach's alpha.

The extent to which Carlson et al. (2010) was referenced for the PCA Taxonomy's underlying theory speaks to the high quality of research done to establish the PCA's theoretical underpinnings. The references to the PCA as an exemplar in test development offered additional support to the quality of instrument development. However, this citation analysis revealed an unexpected finding that limited psychometric work had been conducted using PCA response data. Specifically, statistical modeling, such as factor analysis and diagnostic classification modeling, did not appear in any literature. The literature also lacked empirical studies that

www.manaraa.com

investigated the instrument's dimensionality and the appropriateness of reporting a single composite score or multiple subscores.

Table 2

Distribution of Articles Across Reference Categories

Theme	Count of Articles	Percent of Articles
Reference to Theory	53	65%
Instrument Reference	36	44%
Type of Test	18	22%
PCA Data	22	27%
Psychometric	1	1%

Note. The percent of articles in each category exceeds 100% due to single articles being classified into multiple categories.

Confirmatory Factor Analysis

Factor analysis (FA) is a type of analytical method with classes such as exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) that can be used to evaluate latent constructs (i.e., factors) associated with a given set of observable variables. FA is commonly used in the development and evaluation of measurement instruments. FA may also be used to evaluate the dimensionality of a test empirically. Fabrigar and Wegener (2012) noted that more classical approaches to instrument development placed a heavy emphasis on using estimates of reliability (e.g., Cronbach's alpha) to conclude that instruments were unidimensional. However, Fabrigar and Wegener (2012) emphasized the potential for reliability to be high even when scales are by nature multidimensional. Even when subscales are used, they noted that

> subscales might or might not tap distinct constructs or the constructs might not group in the way expected by the researcher. Fortunately, factor analysis provides a clear method

www.manaraa.com

> for testing the dimensionality of a set of items and determining which items appropriately belong together as part of the same scale or subscale. (p. 23)

In addition to identifying which items are associated with a latent construct, FA can also evaluate the strength of the item to factor relationship.

The application of FA methods for test development and evaluation often places the practitioner in a position to make several, sometimes subjective, decisions. The literature on FA is replete with recommendations to guide practitioners through the process of conducting FA. Several articles and books have been published to synthesize these recommendations to provide holistic guides to FA (Beavers et al., 2013; Brown, 2015; Fabrigar & Wegener, 2012; Gaskin & Happell, 2014; Harrington, 2009; Schmitt, 2011; Worthington & Whittaker, 2006).

Confirmatory factor analysis (CFA) is a class of factor analytic methods often used for (a) psychometric evaluation of an instrument, (b) construct validation, (c) investigation of method effects, and (d) measurement invariance. CFA is considered under the larger umbrella of structural equation modeling (SEM) techniques and constitutes the measurement model (Brown, 2015; Kline, 2011; Raykov & Marcoulides, 2006; Wang & Wang, 2012). CFA necessitates the specification of all aspects of the model before parameter estimation, including (a) the number of factors, (b) the item to factor relationships, and (c) error variances. For this reason, CFA is often used to test models based on theoretical hypotheses or hypothesized models derived from the data-driven EFA process (Haig, 2005).

CFA model parameter estimation is a subset of the broader common factor model. Maximum likelihood and weighted least squares estimations and their associated variants comprise the more common CFA estimation methods (Wang & Wang, 2012). Specifying a CFA model is typically based on theory or prior research. Specifying a CFA model necessitates the

www.manaraa.com

consideration of what parameters to include in the model estimation process and what to exclude from the model.

Data Adequacy and Sampling

In terms of the adequacy of the number of items necessary for estimating statistical models, MacCallum et al. (1999) recommended the final set of items contain at least three to four items per factor (construct) to support the data demands for estimating FA models adequately. However, in general, they stated that "it is desirable that the number of [items] be at least several times the number of factors" (1999, p. 90). The literature review by DiStefano and Hess (2005) echoed the recommendation for at least three variables per factor and reported a median of four items per factor from the 84 models they reviewed. They also noted that "a latent factor measured by one variable (e.g., one subscale) is not optimal and may lead to problems with estimation as well as with construct interpretation" (DiStefano & Hess, 2005, p. 227). It is important to note that these recommendations pertain to the final number of items associated with a factor and not the initial number of items created during the item writing process.

Two considerations for sampling in the context of FA include the sample characteristics and sample size. In the context of FA and test development, it is more desirable that the sample be representative of high and low scores rather than representative of other demographic variables (Worthington & Whittaker, 2006). Fabrigar and Wegener (2012) noted that a sample that is too narrowly drawn (e.g., only university students) might, depending on the construct being measured, produce a sample of data with a low variance, which could be problematic for FA estimation. They wrote (2012):

www.manaraa.com

> Reduced variance on the measured variables will then tend to attenuate the correlations among measured variables. Such attenuated correlations will, in turn, tend to attenuate the factor loadings and the correlations among factors. (p. 27)

Worthington and Whittaker (2006) also caution that homogeneity in the sample scores may cause problems that may persist, even when dealing with large sample sizes.

Several recommendations have been given to guide practitioners in determining adequate sample sizes. Ratios ranging from five to ten participants for every item have been widely used for sample sizes under 300 participants (Tinsley & Tinsley, 1987). However, some research indicates that these ratios may be an oversimplification to determining the adequacy of sample sizes (Fabrigar & Wegener, 2012; Worthington & Whittaker, 2006). Item communalities have been found to be an important element in the adequacy of sample size. Worthington and Whittaker (2006) synthesized sample size recommendations to include the element of communalities. They offered four broad guidelines:

> (a) Sample sizes of at least 300 are generally sufficient in most cases, (b) sample sizes of 150 to 200 are likely to be adequate with data sets containing communalities higher than .50 or with 10:1 items per factor with factor loadings at approximately |.4|, (c) smaller samples sizes may be adequate if all communalities are .60 or greater or with at least 4:1 items per factor and factor loadings greater than |.6|, and (d) samples sizes less than 100 or with fewer than 3:1 participant-to-item ratios are generally inadequate (p. 817).

Brown (2015) noted that the use of dichotomous items necessitates a larger sample size than the use of continuous data. According to Flora and Curran (2004), the need for drastically larger sample sizes of dichotomous items can be reduced using a robust WLS estimator (e.g., WLSMV). However, they recommend caution when estimating large models with small sample

www.manaraa.com

sizes as they found such circumstances produced slightly biased test statistics and standard errors.

Considerations of both sample characteristics and sample size are made to guard against overfitting a statistical model and limiting the generalizability of findings. A cross-validation sample may also be used to protect against an overfitted model. DeVellis (2012) stated that "replicating a factor analytic solution on a separate sample may be the best means of demonstrating its generalizability" (p.158). Floyd and Widaman (1995) recommended that the cross-validation sample be randomly drawn from a larger population to help ensure the generalizability between model development samples and the cross-validation sample. This process permits the comparison of model fit from the sample used to develop the model with a presumably equivalent sample that was not used in the model development process. The obtainment of similar model fit indices with both samples provides evidence for the generalizability of the model. Floyd and Widaman also noted that the process of cross-validation could be used to test the generalizability from the sample used for model generation and a sample with differing characteristics (e.g., clinical and non-clinical samples). However, the evaluation of known differences between groups is generally reserved for analyses of measurement invariance and is an essential part of developing scales for use with multiple groups (Borsboom et al., 2008; Guenole & Brown, 2014; Lubke et al., 2003).

Dichotomous Items

Estimating FA models with dichotomous data requires specialized estimation procedures to account for the violation of the assumptions common to many estimation procedures. Brown (2015) reviewed several issues that arise when conducting FA with dichotomous data. He noted that the violation of the approximately interval-level data assumption has consequences such as

www.manaraa.com

(a) attenuated estimates, (b) factors representing item difficulty rather than latent constructs, (c) incorrect parameter estimates, and (d) inconsistent test statistics and standard errors. Brown and others (Wang & Wang, 2012) recommended using the WLSMV Mplus estimation procedure (Muthén & Muthén, 2017). The WLSMV estimator uses a latent response variable framework to estimate a normally distributed continuous latent variable (y^*) underlying the observed dichotomous variable. A threshold is estimated, which links the amount of the latent variable y^* needed to respond positively to the item (y = 1). Restated, the threshold for a dichotomous variable is the point on y^* where y = 1 if the threshold is exceeded (similar to a 2PL IRT model) (Brown, 2015, p. 355; Wang & Wang, 2012, p. 68). Brown also noted that the WLSMV estimator's use causes the observed variances of the items not to be analyzed. Mplus provides the delta and theta parameterization methods for scaling. The delta parameterization fixes the item variances to 1.0, which causes the residual variances to be unidentified.

> Thus the measurement errors (θ) of the CFA model with [dichotomous] indicators are not free parameters, but instead reflect the remainder of 1 minus the product of the squared factor loading and factor variance (or simply 1 minus the squared completely standardized factor loading). (Brown, 2015, p. 356)

With the theta parameterization, the item residual variances are fixed to unity. Brown (2015) explains:

> the variances of y* are computed as the sum of the residual variance plus the variance due to the latent variable (where θ = 1 for all indicators)…this method is useful when the structure of the residual variances may be an important aspect of the measurement model (e.g., to obtain fit diagnostic information regarding the possible presence of method effects). (p. 356)

www.manaraa.com

When appropriately accounted for, data from dichotomously scored items can be subjected to psychometric analysis using FA techniques to determine an instrument's dimensionality.

Model Identification

An identified CFA model is a model in which it is mathematically possible to obtain a unique estimate for each parameter in the model. That is, the number of freely estimated parameters does not exceed the amount of information provided by the data. A model is considered *under-identified* when the number of parameters to be estimated exceeds the amount of available information from the data. An under-identified model cannot be estimated and will fail to converge upon a solution for estimating parameters. A model is considered to be *just-identified* when the number of freely estimated parameters equals the amount of available information. A *just-identified* model only allows for a single solution of parameter estimates that perfectly fit the data. Although a just-identified model can be estimated, the results of a just-identified model have limited practical utility because their *perfect fit* does not allow for model comparisons. An *over-identified* model is when the number of freely estimated parameters is less than the amount of information provided by the data. Specifying an over-identified model holds more practical utility as it permits the use of many fit statistics and indices, which can be used for model comparison and validation. A single-factor model requires at least three items with uncorrelated measurement errors to be identified. For models with multiple factors, having three items per factor is a general recommendation to achieve an over-identified model (Brown, 2015; DiStefano & Hess, 2005; Wang & Wang, 2012).

Model Evaluation

Models can be broadly defined as an abstraction of reality or a simplified representation of a phenomenon. As a simplification of reality, models are not a perfect representation and

www.manaraa.com

always contain a degree of error. As previously alluded to, model comparison is a key component to factor analysis and particularly CFA. Generally, model comparison is an evaluative endeavor that seeks to draw on multiple sources of evidence to judge the quality of a model and determine the degree to which model predicted variances equal observed variances. In many instances, competing models are estimated and evaluated to determine which model (a) has the least amount of error, and (b) presents the most parsimonious representation of reality. Model fit statistics and indices are often used for model comparison. Both relative and absolute fit indices have been used to compare CFA models. These methods are used to evaluate the models as a whole. Some of these methods take into account the number of parameters specified in the model compared to the possible number of parameters. In this way, these methods of model comparison penalize an overspecified model. Brown (2015) recommended that researchers consider and report at least one indicator of model fit from each of these categories. West et al. (2012) recommended using CFI, TLI, SRMR, and RMSEA indicators.

Absolute model fit indices assess how well the relationships specified CFA model, as represented in the *model implied covariance matrix* (Σ), reproduce the relationships found in the data, as represented in the *observed covariance matrix* (S) (Brown, 2015). The chi-square goodness-of-fit statistic (χ^2) (Joreskog, 1969) was the first model fit index used to evaluate absolute model fit. The χ^2 statistic tests the difference between the model implied covariance matrix (Σ) and the observed covariance matrix (S), where a statistically significant result implies a statistically significant difference between the two matrices. As such, a statistically significant χ^2 result is typically considered undesirable (Wang & Wang, 2012), indicating that the model does not match the data. Although many researchers report χ^2, this statistic has several limitations that reduce its utility to applied researchers. These limitations include:

www.manaraa.com

- High sensitivity to sample size; larger sample sizes are more likely to produce a statistically significant χ^2
- Fitting functions often produce distributions that do now follow the χ^2 distribution due to small sample size or violations of multivariate normality
- χ^2 increases with the number of variables included in the model
- The assumption that S = Σ may be too strict and reject solutions where an *acceptable* approximation has been found

(Brown, 2015; Wang & Wang, 2012)

Many scholars recommend the use and reporting of the standardized root mean square residual (SRMR) and the root mean square error of approximation (RMSEA) as indicators of absolute model fit (Bentler, 2007; West et al., 2012; Worthington & Whittaker, 2006). Brown described the SRMR conceptually as "the average discrepancy between the correlations observed in the input matrix and the correlations predicted by the model" (2015, p. 70). Hu and Bentler (1999) suggested that an SRMR value of < .08 would be considered a good fit, while others consider this to be a less demanding standard (Kline, 2011). Yu (2002) found that the SRMR did not perform well in simulation studies when indicators were dichotomous. Yu recommended using the weighted root mean square residual (WRMR) for dichotomous items with a cutoff value of ≤ 1.0, indicating a good model fit.

The root mean square error of approximation (RMSEA) represents the degree of data-to-model discrepancy and is sometimes classified as a *parsimony correction index* (Brown, 2015) because RMSEA values are adjusted based on the degrees of freedom present in the model. That is, more parsimonious models (i.e., more degrees of freedom) are expected to have lower RMSEA values. However, Kline (2011) noted that models with more degrees of freedom are not

www.manaraa.com

inherently favored over those with less because the parsimony correction diminishes with increased sample size. RMSEA values range from 0 to 1, where values closer to zero are indicative of model-data fit. Some of the more widely cited guidelines for evaluating RMSEA values are those set forth by Browne and Cudeck (1993) and Hu and Bentler (1999). Browne and Cudeck (1993) suggested that an RMSEA value of 0 = perfect fit; < 0.05 = close fit; 0.05-0.08 = fair fit; 0.08-0.10 = mediocre fit; and > 0.10 = poor fit. Hu and Bentler (1999) suggested using RMSEA values < 0.06 to indicate a good model fit.

Relative model fit indices, sometimes referred to as comparative fit indices, describe the relative improvement of the specified model to a more restricted baseline model, which often is a *null* model assuming no covariance among all indicators (Hu & Bentler, 1998). Values for these indices typically range from 0 to 1, where values closer to one are indicative of better model-data fit. The Comparative Fit Index (CFI) (Bentler, 1990) is conceptually the ratio of improvement moving from the null model to the specified model. CFI values "close to 0.95" (Hu & Bentler, 1999, p. 27) are considered a good fit. The Tucker-Lewis Index (TLI) (Tucker & Lewis, 1973) is another widely used relative model fit index. One way in which the TLI differs from the CFI is by incorporating a penalty for model complexity. TLI values close to 0.95 are also considered an indicator of a good model fit (Hu & Bentler, 1999). Wang and Wang (2012) noted that a TLI value < 0.90 is indicative of a model that needs to be respecified.

Information criteria indices are commonly used to compare two specified models, particularly non-nested models. These indices include but are not limited to the Akaike Information Criterion (AIC) (Akaike, 1974, 1987), Bayes Information Criterion (BIC) (Schwarz, 1978), and the sample size adjusted BIC (ABIC) (Sclove, 1987). Each of these indices is a parsimony corrected fit index. Lower values of AIC, BIC, and ABIC are indicative of a good

www.manaraa.com

model fit. Therefore, when using these indices to compare models, the model with the smaller value would be considered the better fitting model.

As previously mentioned, the evaluative nature of examining CFA results necessitates considering multiple sources of evidence to judge the model's quality. Researchers who focus on a single piece of evidence run the risk of misjudging the model's quality. Even while considering multiple sources of evidence, Marsh et al. (2004) counseled researchers not to interpret guidelines, or *rules of thumb*, in the literature as *golden rules*. Wang and Wang (2012) further cautioned that a model with strong evidence from fit indices does not indisputably conclude that it is the *correct* model. They noted that "the model evaluation is not entirely a statistical matter. It should also be based on sound theory and empirical findings. If a model makes no substantive sense, it is not justified even if it statistically fits the data very well" (Wang & Wang, 2012, p. 22).

Model Comparison and Revision

The estimation of a CFA model is less commonly used in a strictly confirmatory approach where the specified model structure is categorically accepted or rejected. In these situations, modifications are not made to the model. More frequently, modifications to CFA models are undertaken to improve the model parsimony and interpretation and improve model fit. Initial CFA models based on theory or empirical findings often do not fit the data very well. These initial ill-fitting models are often used as the starting point for revisions based on the initial model parameter estimates. Raykov and Marcoulides (2006) emphasized this point:

> The starting point of CFA is a very demanding one, requiring that the complete details of a proposed model be specified before it is fitted to the data. Unfortunately, in many

www.manaraa.com

> substantive areas this may be too strong a requirement since theories are often poorly developed or even nonexistent. (p. 117)

Common model revisions include (a) the number of factors, (b) item to factor relationships, and (c) modifications to the error theory. Kline (2011) noted that model revision is often more challenging than the initial model specification because of the vast number of changes that could be made to the model. A methodical approach based on substantive evidence should guide the model revision process. MacCallum (2003) cautioned against the practice of modifying a model to the point of overspecifying the model. He recommended that model modification "focus on identifying and correcting gross misspecification" (2003, p. 129). MacCallum further emphasized the importance of cross-validation.

> Of critical importance is that when a model is modified and eventually found to fit the data well, that model must be validated on new data. That is, a model cannot be supported by a finding of good fit to data when that model has been modified so as to improve its fit to that same data. (p. 129)

According to the review of CFA reporting practices by Jackson et al. (2009), modifications are often either unreported or provide limited to no details regarding the nature of modifications. They recommended clearly distinguishing between proposed or theoretical models and models resulting from post hoc modifications. They also recommended that the post hoc modification process be well documented.

Modifications to the number of factors in a model should be rare when the initial CFA model is based on substantive theory and empirical data (e.g., EFA modeling). Modifying the number of factors is primarily undertaken to resolve instances where higher-order factors or the collapsing of multiple factors are merited and where correlated errors (i.e., residuals) could better

www.manaraa.com

account for an item to factor relationships (Brown, 2015). Factor correlations can be used to provide evidence to support the extent to which multiple factors should be reduced to fewer factors. It is generally expected that factors be at least moderately correlated. However, large correlations that exceed .85 are often problematic (Brown, 2015). These high correlations, combined with substantive theory, could provide evidence that the model should be modified to collapse the highly correlated factors into a single factor or that a higher-order factor is specified to account for the relationship between factors.

Additional modifications to the number of factors in a model could be justified through a simultaneous modification to the model error theory. That is, the estimation of a correlation between item error parameters could be added to a model to account for shared item variance while permitting items to load on separate factors. This practice is commonly applied to modify the number of factors while accounting for measurement method effects (Harrington, 2009).

Modifications to item factor relationships can provide another potential for model modifications. Low factor loadings may indicate that the item does not measure the factor well. Such items could be considered for revision or removal from the model. Items loading on multiple factors where one loading is relatively higher than another may indicate the model should be adjusted to associate the item with a single factor.

Standardized errors/residuals can be conceptually described as the "number of standard deviations by which the fitted residuals differ from the zero-value residuals associated with a perfectly fitting model" (Brown, 2015, p. 98). The evaluation of the absolute value of standardized residuals can help identify localized areas of model strain. Positive standardized residuals may indicate an underspecified model, while negative values indicate an overspecified model. Raykov and Marcoulides (2006) noted that standardized residuals less than an absolute

www.manaraa.com

value of 2 are not a cause for concern. However, the size of standardized residuals tends to be inversely related to sample size. Therefore, some researchers recommended a larger |2.58| guideline (Harrington, 2009; Kline, 2011; Wang & Wang, 2012). Raykov and Marcoulides (2006) also noted that the standardized residual distribution shape could be useful for evaluating model performance. That is, a uniform distribution would suggest that the model is performing equally across all items. At the same time, more isolated cases of large absolute standardized residuals would be indicative of a localized area of model strain.

The identification of large standardized residuals can provide evidence for the modification to the model error theory. Models can be modified by specifying the model to estimate a correlation between item error parameters. Choosing which correlated error terms to add to the model is an evaluative process that, like all model revisions, should be based on substantive information. Positively and negatively worded items and other item characteristics such as common item stems or wording can cause *method effects* (i.e., a common systematic error between items caused by common item characteristics). Correlated errors can be added to the model to account for hypothesized method effects. Modification indices can also be used to identify correlations among item errors. Wang and Wang (2012) emphasized that the addition of correlated errors to any model should be "substantially meaningful" (p. 40) and not added solely based on modification indices.

The modification index can be used as a tool to identify potential model revisions. A modification index value is the estimated reduction in the model's chi-square for a given model modification. A modification index of 3.84 with 1 *df* would constitute a statistically significant change. Considering the chi-square test's limitations, the relative size of the modification index is often given more consideration than statistical significance (Brown, 2015; Raykov &

www.manaraa.com

Marcoulides, 2006; Wang & Wang, 2012). It is important to note that modification indices assess the potential effect of freely estimating a currently fixed model parameter. The omission of key elements from the model may contribute to poor model fit that modification indices will not detect. As Brown (2015) wrote:

> Modification indices can point to problems with the model that are not the real source of misfit. Again, this underscores the importance of an explicit substantive basis (both conceptual and empirical) for model (re)specification. (p. 142)

Using the modification index and other parameters as a guide to modify a model is often referred to as a *model specification search*. MacCallum et al. (1992) noted several limitations to conducting a data-driven model specification search. The results of their research highlighted the importance of cross-validation samples when conducting a model specification search. They found that sequential specification searches can be highly unstable with smaller sample sizes. They wrote:

> when a sequential specification search is conducted in practice using data from a single sample, researchers cannot have great confidence that the specific model modifications would generalize beyond that sample. Unless sample size is very large, modifications may be quite idiosyncratic to that particular sample. (MacCallum et al, 1992, p. 501)

The inconsistency of the modification searches to generalize to different populations prompted MacCallum et al. (1992) to recommend that modification searches not be conducted when a model fits well. If a model specification search is to be conducted, it should have a theoretical underpinning and should not constitute major model changes with one model change investigated at a time (Brown, 2015; Kline, 2011; Raykov & Marcoulides, 2006; Wang & Wang, 2012).

www.manaraa.com

Diagnostic Classification Models (DCMs)

Diagnostic classification models (DCMs) are a class of statistical models that share many of the same underlying computational foundations of factor analysis (Rupp & Templin, 2008b). While factor analysis is often used to investigate an instrument's dimensionality, it can also be used to provide a score for each dimension in the model. However, many of the inherent characteristics of factor analysis make it difficult to interpret and report these scores to students. The well-established research base devoted to DCMs has sought to develop a modeling technique that overcomes the inherent challenges associated with reporting factor analysis scores.

Applying DCMs to multidimensional assessments provides several advantages over many traditional assessment practices. Primarily, students can be provided with a mastery profile that communicates diagnostic information for each student about the mastery of the fine-grained dimensions of the assessment. This approach differs from a more traditional assessment approach where students are given a single total test score without further information on their mastery of specific skills tested. More traditional assessment approaches require students and instructors to review individual item responses and intuit the strengths, weaknesses, and general concept mastery. This process can be both difficult and time-consuming. The application of DCMs seeks to overcome these difficulties by providing a straightforward approach to assess students and clearly communicate results.

Several specific advantages of DCMs noted in the literature are (a) the simultaneous measurement of multiple attributes (i.e., multidimensionality), (b) estimation of student mastery profiles, (c) opportunities for more complex item structures, (d) higher reliability with fewer items, and (e) fewer data demands. In addition to summative assessment, DCMs have great

www.manaraa.com

potential for providing formative feedback on student skill mastery. These formative results may be used to inform educational interventions. For example, diagnostic assessments may be administered before instruction begins to help identify students' strengths and weaknesses. Students may then be provided with differentiated learning resources that have the potential to augment the teaching and learning process. The diagnostic assessment results may also support students' self-regulated learning by more clearly communicating their current understanding and providing actionable information to support students in making study plans.

General Research on Diagnostic Classification Models

Research and applications of diagnostic assessments in various forms began to be published in the early 1980s (Tatsuoka, 1983). However, a strong resurgence of interest occurred in the late 1990s and has continued until the present time. A variety of statistical models have been presented in the literature (Rupp et al., 2010).

Few comprehensive critical literature reviews have been published in the emerging field of diagnostic assessments and their statistical counterparts – Diagnostic Classification Models (DCMs). To the best of the author's knowledge, the review by Rupp and Templin (2008b) constituted the first comprehensive review of the DCM literature. The main objective of this review was to:

> raise awareness about the unique characteristics of DCM vis-à-vis popular scaling alternatives for contexts that call for the analysis of data from diagnostic assessments in a certain discipline. It also serves to address the resulting advantages and disadvantages of DCM by focusing on statistical as well as substantive considerations. (p. 220)

It appears that Rupp and Templin's review did much to focus the academic discourse on DCMs. Their review puts a heavy emphasis on informing readers of the potential advantages and

www.manaraa.com

disadvantages of the various statistical models falling under the broader DCM framework and the potential for future research – which was much needed at the time. Since that time, a plethora of methodological and a small amount of applied research has been conducted to develop the potentialities of DCMs further.

In their literature review, Ravand and Baghaei (2019) provided an overview of recent developments and practical issues in the DCM literature. They argued that although a considerable amount of DCM research has been published, applied DCM research has been stifled by:

> (1) their lack of accessibility to a broad audience interested in their application, (2) fast growth of the models which makes it hard for practitioners to keep up with the latest developments, and (3) unresolved issues such as sample size in DCMs, which hinder their applications (p. 3)

They also noted that "to keep up with the latest developments in DCMs, interested readers must review many articles in diverse sets of journals" (p. 3).

To help make the DCM literature more accessible, Ravand and Baghaei (2019) reviewed various DCMs (Table 3). Several general DCMs have been proposed, including the General Diagnostic Model (GDM; von Davier, 2008), Log-linear Cognitive Diagnosis Model (LCDM; Henson et al., 2009), and the Generalized DINA (G-DINA; de la Torre, 2011). These general, or saturated, DCMs share many of the same characteristics. More restrictive DCMs can be considered special cases of these more general models (Rupp et al., 2010; von Davier, 2014). As such, the emerging practice is an iterative process of (a) fitting a fully saturated model, (b) evaluating the model, (c) fitting rival models, and (d) reevaluating. The flexibility of general

www.manaraa.com

DCMs affords the ability for individual assessment items to take on unique modeling characteristics as opposed to forcing all items to use an identical parameterization.

More restrictive DCMs impose strict assumptions about the item to attribute relationships. For example, noncompensatory DCMs assume that the probability of a correct response is conditional on mastering all attributes associated with an item. That is, noncompensatory DCMs do not permit the strength of one attribute to compensate for the weakness of another attribute in the estimation of the probability of a correct response. Examples of noncompensatory models include the Rule Space Model (RSM; Tatsuoka, 1983), Deterministic Input Noisy "And" gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001), Noisy Input Deterministic "And" gate (NIDA; Junker & Sijtsma, 2001), and the Noncompensatory Reparametrized Unified Model (NC-RUM; Hartz, 2002). In contrast, compensatory DCMs do not assume that a respondent must master all attributes associated with an item to obtain a high probability of a correct response. Compensatory models permit the mastery of one attribute to compensate for not mastering another attribute. Examples of compensatory models include the Deterministic Inputs, Noisy, “Or” gate (DINO; Templin & Henson, 2006), Noisy Inputs, Deterministic “Or” gate (NIDO; Templin & Henson, 2006), and the Compensatory Reparametrized Unified Model (C-RUM; Hartz, 2002). Due to the general nature of the LCDM, attributes are permitted to function as both compensatory and noncompensatory based on the nature of the response data.

Q-Matrices Design

Results obtained using DCMs are only as good as the diagnostic assessment's underlying theory due to an explicit linkage between the assessment items and a domain-specific theory.

www.manaraa.com

Table 3

Categorization of Common Diagnostic Classification Models

Type	Examples
Disjunctive	Deterministic-input, noisy-or-gate model (DINO) Noisy input, deterministic-or-gate (NIDO)
Conjunctive	Deterministic-input, noisy-and-gate model (DINA) Noisy inputs, deterministic-and-gate (NIDA)
Additive	Additive CDM (ACDM) Compensatory reparametrized unified model (C-RUM) Noncompensatory reparametrized unified model (NC-RUM) Linear logistic model (LLM)
Hierarchical	Hierarchical DINA (HO-DINA) model Hierarchical diagnostic classification model (HDCM)
General (Disjunctive, Conjunctive, and Additive)	General diagnostic model (GDM) Log-linear CDM (LCDM) Generalized DINA (G-DINA)

Note. This table was adapted from Ravand and Baghaei (2019, p. 6).

Therefore, it constitutes one of the most critical steps in DCM development (Gorin, 2009). These theoretical linkages must occur before the estimation of the statistical model and are specified using a Q-matrix. The process for developing a Q-matrix first begins by using domain-specific theories to identify and define salient attributes or concepts for which the assessment will be used to classify respondents as masters or non-masters. Second, items that presume to measure the theory-based attributes are developed. The Q-matrix represents the structural item to attribute relationship. As presented in Table 4, a Q-matrix is typically constructed with attributes being represented by columns and individual items by rows. Items that measure a given attribute are indicated by a 1 in the matrix. The Q-matrix specification result is a different vector q for each item i and attribute A such that $q_i = [q_{i1}, q_{i2}, \ldots, q_{iA}]$. For example, item 2 in Table 4 has a q vector $q_2 = [1,0,1]$, indicating that this item measures Attribute 1 and Attribute 3, but not

www.manaraa.com

Attribute 2. Items presumed to measure a single attribute are said to have a *simple attribute structure* (e.g., item 1). Items presumed to measure more than one attribute are said to have a *complex attribute structure* (e.g., item 2). The Q-matrix is used in a confirmatory nature during the estimation of the DCM. This process is similar to using the factor-pattern matrix in confirmatory factor analysis (Bradshaw, 2017; Templin & Bradshaw, 2013).

Table 4

Sample Q-Matrix

Item	Attribute 1	Attribute 2	Attribute 3
1	1	0	0
2	1	0	1
3	0	1	0
4	0	1	1
5	0	0	1

Kunina-Habenicht et al. (2012) emphasized the importance of Q-matrix specification in the process of applying DCMs. They wrote:

> The development or derivation of one or several competing Q-matrices is a critical (and potentially the most challenging) step in the analysis [of DCMs]. Any change in the Q-matrix redefines, at least slightly, the substantive interpretations of the set of user-specified attributes, even if their labels remain the same. Different Q-matrix specifications reflect different theoretical hypotheses about the structure of the diagnostic assessment. (p. 60)

The misspecification of the Q-matrix can have several adverse effects on model estimation quality (Kunina-Habenicht et al., 2012; Rupp & Templin, 2008a). Misspecification of the Q-matrix can appear in the form of an under-specification (i.e., omitting correct item to attribute associations) or over-specification (i.e., including an incorrect item to attribute associations).

www.manaraa.com

Q-matrix design refers to the number of attributes, items, and the prevalence of simple and complex item structures. Simulation studies have shown that a Q-matrix design with one or more simple structure items resulted in higher classification accuracy. Conversely, increasing the number of complex items or complexity of items (i.e., the number of attributes measured by that item) decreased classification accuracy (Madison & Bradshaw, 2015). Madison and Bradshaw's (2015) simulation studies also revealed that DCMs often struggle to isolate distinct attributes when they are always measured in conjunction with another attribute. They recommended combining attributes that are always measured together. The issue of item complexity introduces the need to balance the specificity of attributes and the Q-matrix design complexity. Increased specificity is desirable from a diagnostic perspective, but increased specificity is potentially undesirable from a modeling standpoint because of the added model complexity. Ravand and Baghaei (2019, p. 16) noted that most DCM studies specify up to five attributes.

Several different methods have been explored in the literature to validate Q-matrix design (Y. Chen et al., 2015; Chiu, 2013; de la Torre, 2008; DeCarlo, 2012; J. Liu et al., 2013). However, many of these methods are limited to being implemented within restricted DCMs, such as the DINA model. The inconclusive nature of the literature on Q-matrix validation necessitates estimating and evaluating multiple competing Q-matrix structures in applied DCM work. It is important to note that part of this evaluative process should include considering the interpretability of attributes and, subsequently, the classifications produced by the model (Lei & Li, 2016).

Log-Linear Cognitive Diagnosis Model (LCDM) Parameterization

The general purpose of DCMs is to estimate the probability of a correct response based on the mastery of predetermined attributes. The LCDM (Henson et al., 2009) is a general DCM

www.manaraa.com

parameterized as a linear model akin to an ANOVA. The general form of the LDCM can fit an infinite number of attributes and accommodate an item that measures any combination of these attributes defined as:

$$P(Y_{ri} = 1|\alpha_r) = \frac{\exp\left(\lambda_i^T h(q_i,\alpha_r)\right)}{1+\exp\left(\lambda_i^T h(q_i,\alpha_r)\right)}. \quad (1)$$

This general form of the LCDM can be simplified into unique linear equations to model the log-odds of a correct response conditional on a respondent's mastery of attributes associated with that item. Bradshaw (2017) provided an example of how an item response function for an item measuring two attributes (α_a, α_b) can be derived from the general form of the LCDM. The log-odds of a correct response is modeled conditional on the respondent's attribute mastery profile as:

$$Logit(Y_{ri} = 1|\alpha_r) = \lambda_{i0} + \lambda_{i1(a)}(\alpha_{ra}) + \lambda_{i1(b)}(\alpha_{rb}) + \lambda_{i2(a*b)}(\alpha_{ra})(\alpha_{rb}). \quad (2)$$

The intercept for the linear equation is noted as λ_{i0} and represents the log-odds of a correct response for respondents who have not mastered attribute *a* or *b.* The main effects for having mastered an attribute are noted by $\lambda_{i1(a)}$ and $\lambda_{i1(b)}$ and represent the increase in the log-odds of a correct response given the mastery of the respective attribute. The two-way interaction is noted by $\lambda_{i2(a*b)}$ and represent the increase in the log-odds of a correct response given the mastery of both attributes *a* and *b*.

The LCDM permits the estimation of a fully saturated statistical model (i.e., a model which contains the maximum possible number of parameters). The saturated LCDM subsumes many of the more restrictive DCMs found in the literature. Fitting a saturated model enables researchers to evaluate the nature of the relationship between items and attributes to help curtail

www.manaraa.com

the pitfall of model misspecification. However, the LCDM is often constrained by removing non-significant item parameters for the sake of parsimony.

Model Evaluation of Diagnostic Classification Models

Several methods have been developed to evaluate the quality of DCMs of which model fit plays a central role. Like other statistical models, the model's usefulness hinges on the degree to which the model fits the data. Evaluation of model fit for DCMs includes estimates of both absolute and relative model fit.

Absolute model fit refers to the extent to which the model, as a whole, fits the data. Several methods for assessing absolute fit have been researched. Templin and Henson (2006) wrote about a Monte Carlo resampling technique for estimating model fit while others (Sinharay & Almond, 2007) have proposed a Bayesian posterior predictive model checking. However, Y. Liu et al. (2016) noted that these two approaches to model fit are more challenging to estimate in terms of time and computational requirements. Hansen et al. (2016) researched the appropriateness of applying the limited-information M2 fit statistic (Maydeu-Olivares & Joe, 2006) to DCMs. Hansen et al. (2016) used simulation studies to investigate the M2 statistic's sensitivity to detect testlet effects, misspecification of higher-order structures, Q-matrix misspecification, and misspecification of DCM (C-RUM or DINA). They found the M2 statistic was sensitive to detecting underspecification and over-specification of attributes in the Q-matrix and the omission of an attribute from the Q-matrix. However, the M2 statistic was not sensitive to the detection of an extraneous attribute (i.e., adding an irrelevant attribute to the Q-matrix). Jurich (2015) noted similar findings when applying the M2 statistic specifically to the LCDM.

Relative model fit statistics are used to compare model fit between two or more models. Several commonly used relative model fit statistics can be applied to DCMs including, Akaike

www.manaraa.com

Information Criterion (AIC; Akaike, 1974, 1987), Bayesian Information Criterion (BIC; Schwarz, 1978), and the -2 log-likelihood (-2LL) used to test the difference between two nested models. Sen and Bradshaw (2017) researched the AIC, BIC, and SABIC performance in the context of the LCDM. They found that AIC, BIC, and SABIC's ability to identify the true simulated model depended largely on item quality and base rates. Item quality is defined as the discrimination power of the item. They wrote:

> [H]igher quality items have greater item discrimination, meaning that the items are better at separating masters and nonmasters of the measured attributes. For DCMs, item discrimination is defined as the difference in the probability of a correct response for two groups of students (Sen & Bradshaw, 2017, p. 9).

Simple structure items (i.e., items measuring a single attribute) were simulated to have medium item quality with a discrimination value of 0.60 and 0.64 for high-quality items. Complex structure items (i.e., items measuring multiple attributes) had discrimination values simulated at 0.60 for medium-quality items and 0.83 for high-quality items. The differences in item quality can also be described in terms of item parameters. Simple structure items had a main effect of 2.84 for medium-quality items and 3.0 for high-quality items. Complex structure items had main effects of 1.3 and a two-way interaction of 0.24 for medium-quality items. High-quality items had a main effect of 2.0 and a two-way interaction of 1.0.

The literature on model fit indices for DCMs is currently inconclusive. Several model fit indices have been researched in simulation studies with mixed results. In their review, Ravand and Baghaei (2019) reported several fit indices found in the literature (Table 5). There are currently no research-based guidelines for using these fit indices, which makes their application more difficult for practitioners. Another limiting factor to the evaluation of model fit is the

www.manaraa.com

availability of software with the capacity to compute model fit indices. As such, several fit indices and multiple types of fit should be used to provide multiple sources of evidence for the quality of model fit (Sessoms & Henson, 2018).

Table 5

Model Fit Indices Used With DCMs

Index	Description	References
X^2	Chi-square test statistic	W.-H. Chen & Thissen, 1997 Rupp et al., 2010
MADcor	Mean absolute difference for the item-pair correlations	DiBello et al., 2007
MADRES	Mean residual covariance	McDonald & Mok, 1995
Q3	Measure of local dependence	Yen, 1984
RMSEA	Root mean square error	Browne & Cudeck, 1993
SRMSR	Standardized root mean squared residual	Maydeu-Olivares, 2013

Diagnostic Classification Models in Practice

The review of DCM literature by Sessoms and Henson (2018) captures the developments that have taken place in applying the DCM literature to practical applications. They reviewed 36 papers published since 2009 in 27 various peer-reviewed journals. Of the constructs being measured, 47% were math, and 39% were reading (Sessoms & Henson, 2018, p. 5). Other constructs did not appear in more than one study. This lack of diversity highlights the unproven ability of DCMs to be directly applicable to a wide variety of content areas. The number of attributes measured may serve as a proxy for the complexity of DCM. In their review, Sessoms and Henson (2018, p. 6) found that the number of attributes measured ranged from four to twenty-three (M = 8.19, median = 6.5, SD = 4.95). Sessoms and Henson suggested that "DCM

www.manaraa.com

technical research often does not align with DCM applications. Thus, DCM simulation research may need to expand the number of attributes assessed to increase generalizability to applied research." (Sessoms & Henson, 2018, p. 9).

Guidelines for sufficient sample sizes have yet to be established for the many variations of DCMs. The use of DCMs to provide diagnostic information to classroom teachers has often been cited as one of the great potentials of DCMs. Some have criticized DCM research for using large sample sizes that do not approximate classroom settings (Henson, 2009; Huff & Goodman, 2007). However, large sample sizes may only be necessary for the initial parameterization of the model. Once appropriate model parameters have been established, these parameters could be used in a formulaic approach to produce diagnostic results with small sample sizes. However, large sample sizes are generally recommended for the initial estimation of model parameters. Roughly 61% of articles reviewed by Sessoms and Henson (2018) had a sample size of more than 1,000. Surprisingly, four studies reviewed had a sample size between 50 and 150.

The most popular DCMs variants were the DINA, general models, and the RUM. The distribution of attribute classifications is often used to evaluate the quality of the DCM. A common practice is to report the proportions of attribute masters and non-masters as well as mastery profiles. Attribute associations are also commonly used to evaluate the relationship between attributes. Sessoms and Henson (2018) found that these attribute correlations were often .90 or larger. These high correlations may bring into question each attribute's distinctiveness and, consequently, the DCM's ability to appropriately differentiate between masters and non-masters (Kunina-Habenicht et al., 2009). It is important to note that at least one study which developed diagnostic assessments within the DCM framework had attribute correlations below .80 (Bradshaw et al., 2014). This finding suggests that perhaps highly correlated attributes are a

www.manaraa.com

more complex issue involving domain theories, assessment development, and diagnostic modeling.

The *Standards for Educational and Psychological Testing* (AERA et al., 2014) stated that evidence for the reliability and validity of test scores should be reported in conjunction with a description of how the results will be used. Some have criticized the application of DCMs due to the lack of reliability and validity evidence (Sinharay & Haberman, 2009). Reliability evidence that can be provided for DCMs is distinct from the evidence that typically accompanies CTT or IRT methods. Templin and Bradshaw (2013) conceptualized reliability in the context of DCMs as "how consistent an examinee's estimate from a DCM will be over hypothetically repeated observations. As such, the calculation of the DCM reliability measure is enabled by simulating repeated testing occasions through repeated draws from an examinee's posterior distribution." (p. 258). This definition of reliability is similar to the more traditional test-retest reliability. They developed a three-stepped method for providing reliability evidence in the context of DCMs. First, the probability of attribute mastery is estimated for each examinee. Second, a replication contingency table of attributes is created. Third, attribute reliability is calculated using a tetrachoric or polychoric correlation of attributes. Sessoms and Henson (2018) found that few studies reported reliability evidence (36%) and that only 61% utilized DCM-specific approaches to estimating reliability.

The literature on validity evidence in the context of DCMs has not been well developed. As such, it is not surprising that Sessoms and Henson (2018) found sparse reporting of validity evidence. They found that only 22% of studies reported any type of validity evidence and that this evidence was rarely formulated into a validity argument. They suggested practitioners using

www.manaraa.com

DCMs should construct a validity argument, which may include a treatment of (a) construct representativeness, (b) internal validity, (c) external validity, and (d) results use and utility.

Retrofitting Assessment Data to Diagnostic Classification Models

Retrofitting is commonly defined as the post hoc analysis of non-diagnostic assessment response data to DCMs. A common retrofitting practice is to take an existing unidimensional assessment and attempting to tease out multiple dimensions through the use of DCMs. As expected, this practice often results in suboptimal model fit and classifications. In contrast, the development of a diagnostic assessment with the specific purpose for fitting response data to a DCM is often lauded as the ideal in diagnostic assessment and modeling. However, this practice requires a substantial amount of effort, and there are few published examples for practitioners to follow (Bradshaw et al., 2014; Sessoms & Henson, 2018). Many authors have acknowledged the limitations of retrofitting and argued for the use of purposefully designed multidimensional diagnostic assessments (Bradshaw et al., 2014; Gierl & Cui, 2008; Leighton, 2008; R. Liu et al., 2017; Ravand & Baghaei, 2019; Rupp & Templin, 2009). Rupp and Templin (2009) strongly asserted that "we need to stop retrofitting DCMs to unidimensional assessments (p. 116).

R. Liu et al. (2017) noted that:

> retrofitting multidimensional DCMs [to unidimensional assessments] can introduce a conundrum with respect to dimensionality that may not be easily resolved. However, retrofitting provides a way to attempt to reap the benefits of DCM in the current landscape in which not many tests have been designed to assess multidimensional skills, and it will be a number of years before that situation changes given the time intensive nature of developing such assessments. Therefore, it is possible that retrofitting may be a primary source of DCM applications for the near future until the test construction

www.manaraa.com

> processes for multidimensional assessments become more ingrained in practice. We support the notion that retrofitting should not be encouraged as a standard approach to a measurement endeavor and not all assessments are suitable for retrofitting. But it also may be justifiable to recognize the struggle between the urging needs of diagnostic information and limited resources to develop and administer new diagnostic assessments. (p. 359)

Theoretically based multidimensional assessments developed and used outside the context of DCMs may be more favorable candidates for retrofitting than their unidimensional counterparts.

One criticism of retrofitting is a lack of a theoretical foundation for assessment dimensions (i.e., attributes) and the ad-hoc association of items to cognitive processes (Leighton, 2008). However, in rare circumstances, non-diagnostic assessments have been developed based on substantive cognitive theories using sound assessment development practices; the Precalculus Concept Assessment (PCA; Carlson et al., 2010) is one such assessment. Many of the practices recommended for the development of uniquely diagnostic assessments for the use with DCMs (Bradshaw et al., 2014; Bradshaw, 2017) were followed in the PCA development. As such, PCA response data is positioned well for an ad-hoc retrofit analysis using DCMs.

Frameworks for retrofitting assessment data to DCMs have been established to guide practitioners in applying DCMs (R. Liu et al., 2017; Ravand & Baghaei, 2019). The focal point of the retrofitting process is the identification of attributes and their association with assessment items (i.e., the construction of the Q-matrix). The remaining process of fitting DCMs and assessing model quality is not unique to the retrofitting process and follows recommendations for applying DCMs to diagnostic assessments.

www.manaraa.com

CHAPTER 3

Method

The methods described in this section were formulated to address each of the study's four research questions, namely:

1. To what extent does a confirmatory factor analysis of PCA pretest data provide evidence that supports the validity of the three-factor structure implied by the PCA Taxonomy?
2. If the three first-order factors are found to be highly correlated, to what extent do rival models (i.e., a single-factor model, a second-order factor model, or a bifactor model) fit better than the three first-order factors model and illuminate the interrelationships among the three first-order factors?
3. How successfully can the PCA response data be retrofitted for an analysis using a general diagnostic classification model (DCM)?
4. How does the adequacy of a DCM model based on the factor structure implied by the PCA Taxonomy compare with a DCM model based on the CFA results?

Confirmatory factor analysis (CFA) procedures were used to address the first and second research questions. The third and fourth research questions were addressed using diagnostic classification methods (DCMs). The remainder of this section describes the method by which these questions were addressed, including the data collection and analysis procedures.

Data Collection and Instrumentation

This study was conducted using pre-existing data collected using the Precalculus Concept Assessment (PCA) instrument. Each PCA item consists of an item stem and five response options. The majority of the PCA items are context-dependent because they rely on interpreting a

www.manaraa.com

graph or figure. The PCA does not employ the use of testlets. That is, each context-dependent item is independent of the others since it relies on a unique graph or figure. Other items include story problems or mathematical equations presented in isolation.

Data for this study were initially collected in select College Algebra and Precalculus sections at a large private university in the mountain west and one public university in the Phoenix metropolitan area. The 25-question PCA was administered to students in these sections both at the beginning and end of the semester. A sample of 3,018 pretest administrations was selected for this study.

PCA student pretest response data were split into two data subsets: the *Primary* (n = 1,509) and *Cross-validation* (n = 1,509) samples. These samples were created using systematic random sampling. This technique assigned all students a number representing the order in which their PCA pretest was scored. Students with an even number were assigned to the *Primary* sample, and odd numbers were assigned to the *Cross-validation* sample. This sampling technique was selected to ensure equal representation from course sections and semesters, recognizing that students were not assigned to take the test in any systematic order.

The *Primary* subset was used in confirmatory processes, such as testing the model implied by the PCA Taxonomy and making modifications to that model. The *Primary* subset was also used to test the extent to which alternative models fit better than the implied model due to high correlations between the three first-order factors implied by the PCA Taxonomy. The *Cross-validation* subset was used to evaluate the degree to which model fit is consistent when estimated using data that was not part of the model development process. All available pretest data were used in the DCM specification due to the large data demands associated with this method.

www.manaraa.com

Factor Analysis Procedures

Factor analysis procedures were used to address the first and second research questions. Mplus 8.4 (Muthén & Muthén, 2017) was used to conduct confirmatory factor analyses. Due to the dichotomous nature of the PCA response data, the WLSMV estimator was used for all model estimation procedures (Brown, 2015; Wang & Wang, 2012).

Missing data in testing situations is a common occurrence. These missing data generally occur in two situations where (a) students intentionally skip questions or (b) students leave questions unanswered because they run out of time allocated for testing. The first situation is the only situation that applied to the PCA data because the test was not a timed test. Missing data were inspected to detect any systematic patterns of missingness. The results of this inspection found that there were no clear patterns in missing data. Missing data in these samples accounted for less than 0.12% of the data. One limitation of the WLSMV estimator in Mplus is that it uses the less desirable pairwise method for handling missing data. Considering the limited amount of missing data, the limitations of the pairwise deletion technique presumably did not impact this study's results negatively.

The factor analyses for this study were executed in a systematic, iterative process using preset guidelines for model evaluation and revision. This process consisted of two phases of model specification, estimation, and evaluation. The first phase was conducted using the *Primary* data subset in which the theoretical model of the PCA Taxonomy was tested. This phase also focused on making refinements to the model to arrive at an acceptable model fit. The first phase concluded by evaluating alternative models due to high correlations between the three first-order factors. The second phase of factor analysis procedures was a multi-group analysis using the *Cross-validation* data subset to cross-validate the Phase 1 results.

www.manaraa.com

Confirmatory Factor Analysis of the Implied Theoretical Model

The first phase of factor analysis used CFA methods to test the theorized three first-order factor model derived from the PCA Taxonomy. This model consists of the following item to factor loadings (Figure 1):

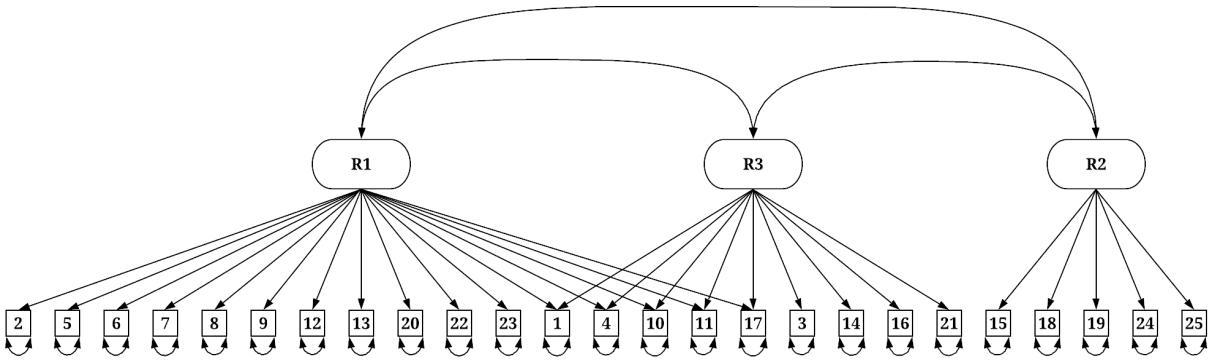
- Process View of Function (R1); items 1,2, 4, 5-13, 17, 20, 22,23
- Covariational Reasoning (R2); items 15, 18, 19, 24,25
- Computational Abilities (R3); items 1, 3,4, 10,11, 14, 16,17, 21

The CFA model implied by the PCA Taxonomy and other potential rival models were evaluated using the CFI, TLI, RMSEA, and SRMR fit indices (West et al., 2012; Yu, 2002). The following recommendations of Hu and Bentler (1999) were used as a guide: CFI and TLI ≥ 0.95, RMSEA ≤ 0.06, and SRMR $< .08$.

Modifications to the model implied by the PCA Taxonomy were made to arrive at a good fitting model that aligns with the previously mentioned guidelines. Modifications were based on a combination of (a) fit statistics, (b) factor loadings, (c) factor correlations, and (d) modification indices.

Figure 1

Model for Three First-Order Model



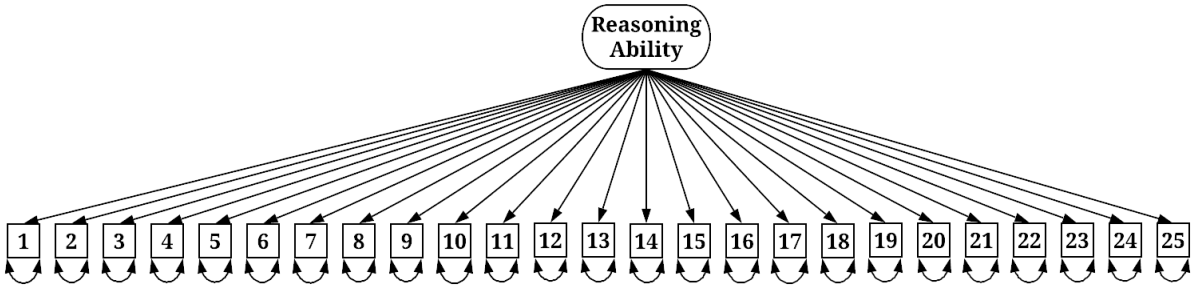
www.manaraa.com

Standardized factor loadings were evaluated based on a minimum weak factor loading of |.30| and desired factor loading ≥ |.40|. Item cross-loadings with a difference ≤ .15 from an item's largest factor loading were considered not to provide evidence for a simple structure (i.e., loading on a single factor). Factor correlations ≥ .85 were reviewed and considered for being specified with an alternative model structure. Localized areas of model strain were evaluated by examining individual standardized errors and the overall distribution of errors. Items with large standardized errors were reviewed. Modification indices were used as a final step in model modification. The relative size of modification indices was considered and compared to substantive theory to justify any recommended model modifications.

After a good fitting model was achieved, the factor correlations were examined. Alternative models were evaluated due to factor correlations greater than or equal to .85 (Brown, 2015). Alternative models evaluated included a single-factor model (Figure 2), a second-order factor model (Figure 3), and a bifactor model (Figure 4). Nested models were compared empirically using the adjusted chi-square difference test using the Mplus "DIFFTEST" function.

Figure 2

Alternative Single-Factor Model



www.manaraa.com

Figure 3

Alternative Second-Order Factor Model

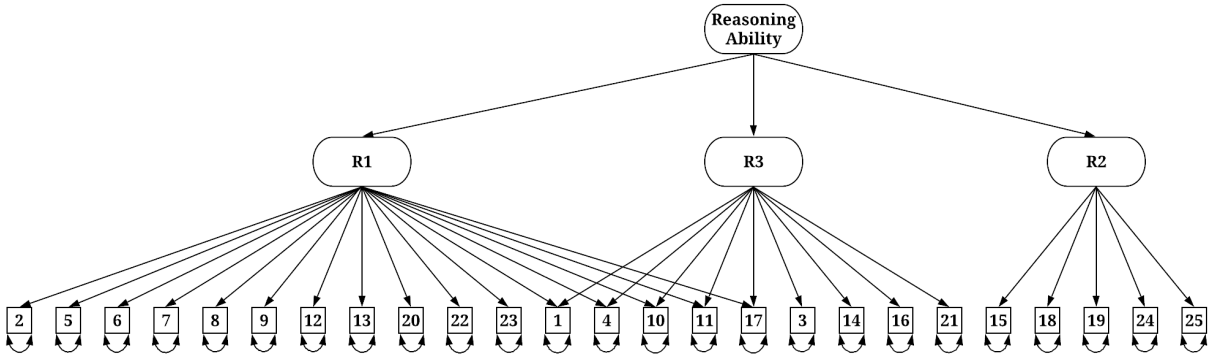
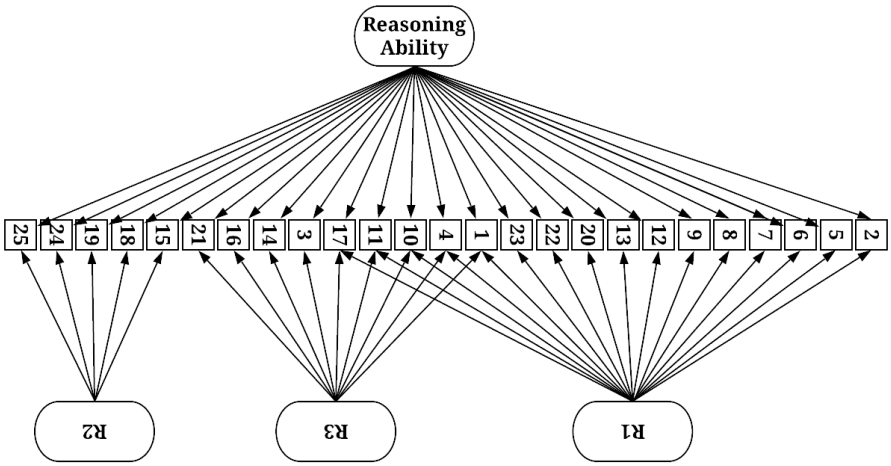


Figure 4

Alternative Bifactor Model



Cross-Validation of Confirmatory Factor Analysis Results

The purpose of the second and final phase of factor analytic procedures was to cross-validate the final models from Phase 1 and to compare these models to the initial models implied by the PCA Taxonomy. Model estimation and evaluation were replicated from Phase 1. However, models in Phase 2 were estimated using the *Cross-validation* data subset. Model fit using the *Cross-validation* data subset was compared to model fit using the *Primary* data subset.

www.manaraa.com

The degree to which model fit varied between these subsets provided evidence to support these models' generalizability.

Multi-group modeling procedures were used to compare models using each of the data subsets using the Mplus model = configural and scalar command with the WLSMV estimator and Theta parameterization. Muthén and Muthén (2017) wrote that the configural model with these estimation settings has "factor loadings and thresholds free across groups, residual variances fixed at one in all groups, and factor means fixed at zero in all groups" (p. 542-543). Muthén and Muthén (2017) also wrote that the scalar model has "factor loadings and thresholds constrained to be equal across groups, residual variances fixed at one in one group and free in the other groups, and factor means fixed at zero in one group and free in the other groups" (p. 543).

Evaluating the degree to which the configural and scalar models differ in model fit is a topic that has been written and researched much (F. F. Chen, 2007; Cheung & Rensvold, 2002; Meade et al., 2008). However, limited research has been conducted on how to approach these comparisons using dichotomously scored items. Based on simulation studies, Sass et al. (2014) reported that using changes in approximate fit indexes (ΔAFI) under these circumstances (i.e., dichotomous data with WLMSV estimator) is often problematic and cautioned against their use.

Considering the lack of empirical evidence for an alternative approach, this study used the Mplus "DIFFTEST" to test the differences between the configural and scalar models. The final models resulting from the two phases of factor analyses were used to provide empirical evidence to support recommendations for how PCA scores could be interpreted and reported. In addition, the final models were used to inform the subsequent DCM procedures by providing the basis for an alternative Q-matrix structure to the Q-matrix structure implied by the PCA Taxonomy.

www.manaraa.com

Diagnostic Classification Modeling Procedures

Diagnostic Classification Models (DCMs) were used to address the third and fourth research questions. PCA student response data were fitted to the Log-linear Cognitive Diagnostic Model (LCDM) using the C-RUM parameterization. The model estimation followed the published recommendations for estimating the LCDM using the statistical software Mplus (Fager et al., 2019; Templin & Hoffman, 2013). The item-to-attribute relationship from the Reasoning Ability portion of the PCA Taxonomy was adopted as the Q-matrix for the initial LCDM (see Table 6). The rival factor structure from the FA procedures was used as the basis for an alternative Q-matrix structure. Rival models were compared using (a) model fit indices, (b) attribute classification reliabilities, (c) attribute profile and mastery proportions, and (d) attribute correlation matrices.

Table 6

Q-Matrix Based on the PCA Taxonomy

Item	Process View of Function (R1)	Covariational Reasoning (R2)	Computational Abilities (R3)
1	1	0	1
2	1	0	0
3	0	0	1
4	1	0	1
5	1	0	0
6	1	0	0
7	1	0	0
8	1	0	0
9	1	0	0
10	1	0	1
11	1	0	1
12	1	0	0
13	1	0	0
14	0	0	1
15	0	1	0
16	0	0	1
17	1	0	1
18	0	1	0
19	0	1	0
20	1	0	0
21	0	0	1
22	1	0	0
23	1	0	0
24	0	1	0
25	0	1	0

www.manaraa.com

CHAPTER 4

Results

The results of this study are presented in the context of factor analysis, addressing Research Questions 1 and 2, and diagnostic classification models, addressing Research Questions 3 and 4.

Factor Analysis

The first analysis used CFA to fit the three-factor model implied by the PCA Taxonomy, shown in Figure 1, using the WLSMV estimator in Mplus. The initial estimation of this model converged normally but failed to estimate standard errors. Subsequently, a three-step specification search was conducted to arrive at a model that most closely approximated the model implied by the PCA Taxonomy.

The first step in the specification search identified items with factor loadings < .100. Items meeting this criterion were removed from the model. Although these items were fixed to not load on the originally specified factors, they were not removed from the analysis to retain a comparable data structure throughout the various steps of the model specification search. As a result, items 1, 4, 10, 13, and 23 were detached from the R1 factor, and items 14 and 17 were detached from the R3 factor.

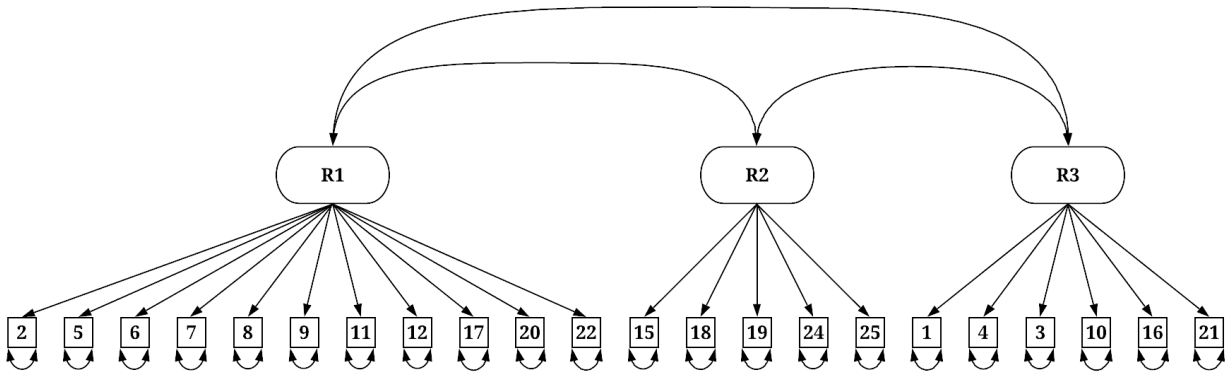
The second step in the specification search estimated the respecified model. The analysis terminated normally and produced standard errors, which in turn permitted statistical tests. Standardized factor loadings from this model were then examined. All factor loadings were statistically significant except for item 11 on the R3 factor. As a result, item 11 was detached from the R3 factor.

www.manaraa.com

The third and final stage of the specification search estimated the respecified model with item 11 removed from the R3 factor. The specification search ended with this respecified model, which successfully estimated the requested parameters, standard errors, and fit statistics. The path diagram for this respecified model is shown in Figure 5. In total, items 1, 4, 10, 13, and 23 were detached from the R1 factor, and items 11, 14, and 17 were detached from the R3 factor. In the final model, items 13, 14, and 23 were removed entirely from the analysis as they no longer had any association with one of the three factors. The respecified model resulting from the model specification search is referred to as the *three-factor model* in the subsequent sections.

Figure 5

Three-Factor Model



Analysis of the Three-Factor Model

The analysis of the three-factor model (Figure 5) terminated normally. The fit statistics (CFI = 0.952, TLI = 0.946, RMSEA = 0.026, SRMR = 0.049) indicated that the model fit the data well, meeting all predetermined criteria specified in the method section. The resulting standardized factor loadings are reported in Table 7. The

www.manaraa.com

correlations between the R3 factor and the other two factors were high, as reported in Table 8.

Table 7

Standardized Factor Loadings for the Three-Factor Model

	Factor loadings		
PCA Item	Process View of Function (R1)	Covariational Reasoning (R2)	Computational Abilities (R3)
2	.524		
5	.724		
6	.764		
7	.256		
8	.386		
9	.529		
11	.158		
12	.655		
17	.252		
20	.283		
22	.337		
15		.490	
18		.505	
19		.569	
24		.625	
25		.190	
1			.506
3			.577
4			.524
10			.230
16			.745
21			.241

Table 8

Correlations Among Factors in the Three-Factor Model

Factor	R1	R2	R3
Process View of Function (R1)	_		
Covariational Reasoning (R2)	.725	_	
Computational Abilities (R3)	.812	.861	_

www.manaraa.com

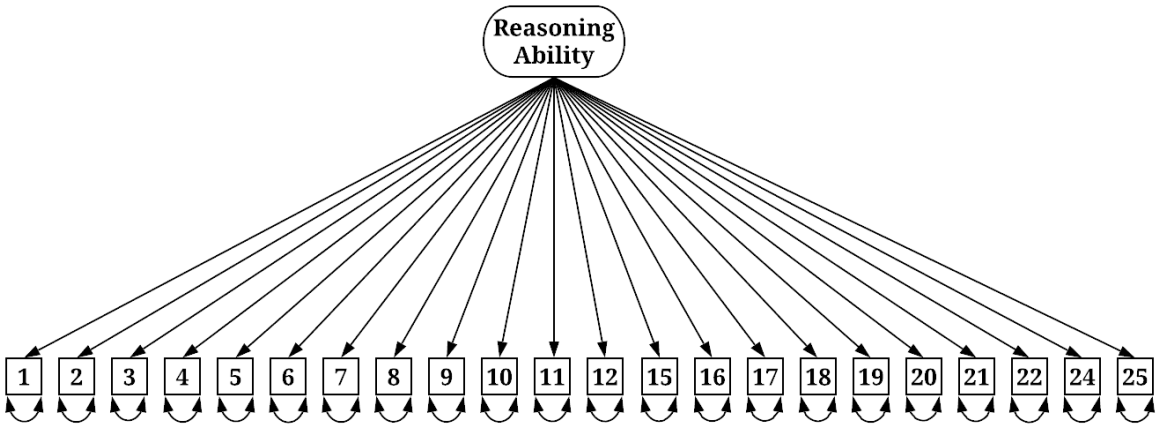
Analysis of Rival Models

The high factor correlations found in the three-factor model prompted the analysis of several rival models. These models included (a) single-factor model, (b) second-order factor model, and (c) bifactor model. An analysis of these models was conducted to understand better the relationship between the three factors (Brown, 2015). The results from these models are presented in the following sections.

The Single-Factor Model. A single-factor model (Figure 6) was estimated and converged normally. The fit statistics for the single-factor model (CFI = 0.934, TLI = 0.928, RMSEA = 0.031, and SRMR = 0.054) met only two of the four predetermined criteria. The Mplus Chi-Square Test for Difference Testing (DIFFTEST) was used to test the difference between the three-factor model and the single-factor model. The result of the DIFFTEST was statistically significant (χ^2 (16, 1509) = 109.524, $p < 0.001$), providing additional evidence that the single-factor model fit the data worse than the three-factor model.

Figure 6

Single-Factor Model



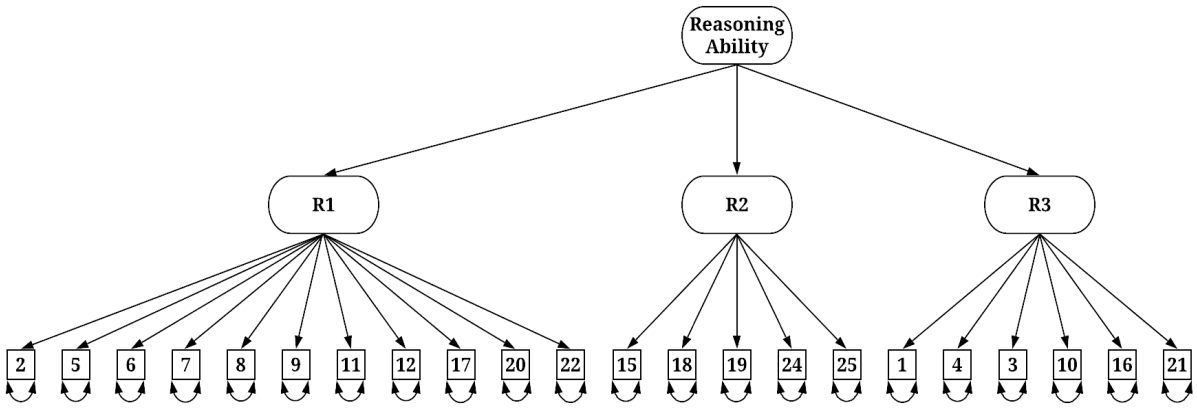
www.manaraa.com

The Second-Order Factor Model. Analysis of the second-order factor model (Figure 7) terminated normally. As expected, the fit statistics for this locally just-identified model were equivalent to the fit statistics of the three-factor model. However, the purpose of estimating this model was not to empirically compare the model fit of the second-order factor model with the three-factor model. Instead, the purpose was to gain additional insight into the relationship between the R1, R2, and R3 factors. This additional evidence was obtained by inspecting residual variances for the R1, R2, and R3 factors and their R^2 values.

A statistically significant residual variance was not found for the R3 factor (Std. Residual = 0.037, p = 0.558), indicating that the R3 factor does not account for a statistically significant portion of the variance. That is, the second-order Reasoning Ability factor presumably accounts for almost all of the variance of R3. The R^2 values for this model were all statistically significant with values ranging from 0.684 to 0.963 (R1 = 0.684, R2 = 0.769, and R3 = 0.963). The high R^2 value for R3 suggests that the R3 factor may be the main contributor to the higher-order Reasoning Ability factor.

Figure 7

Second-Order Factor Model



www.manaraa.com

The Bifactor Model. A bifactor model with one general factor and three specific factors was analyzed and terminated normally (Figure 8). However, the software failed to estimate standard errors for the parameters in this model. Based on evidence from the high correlations in the three-factor model and the residual variance and R^2 values for R3 in the second-order factor model, it was hypothesized that the specific factor R3 in the bifactor model had empirically collapsed (Brown, 2015, p. 303; F. F. Chen et al., 2006). Consequently, the bifactor model was then respecified to exclude R3 as a specific factor while retaining the items corresponding to R3 as part of the general factor. The revised bifactor model then included a general factor and only two specific factors as shown in Figure 9.

The analysis of this respecified bifactor model terminated normally and produced estimates of the standard errors. The fit statistics indicated that this model fit the data well (CFI = 0.953, TLI = 0.943, RMSEA = 0.026, SRMR = 0.048). The ability to make a direct empirical comparison between the three-factor model and the respecified bifactor model was limited. The collapsing of the R3 specific factor caused the respecified bifactor model to not be nested in the three-factor model and therefore eliminated the ability to use the DIFFTEST function as planned. The use of the respecified bifactor model successfully partitioned the unique variances of the items due to the influence of the specific and the general Reasoning Ability factors providing additional empirical evidence related to the unique contributions of R1 and R2.

An inspection of the standardized factor loadings in Table 9 revealed four items (7, 20,18, and 25), which had specific factor loadings that were not statistically significant. These non-significant factor loadings show that specific factors R1 and R2 do not account for a significant amount of item variance beyond the general factor. That is,

www.manaraa.com

the general Reasoning Ability factor accounts for almost all of the variance for these items.

Figure 8

Bifactor Model

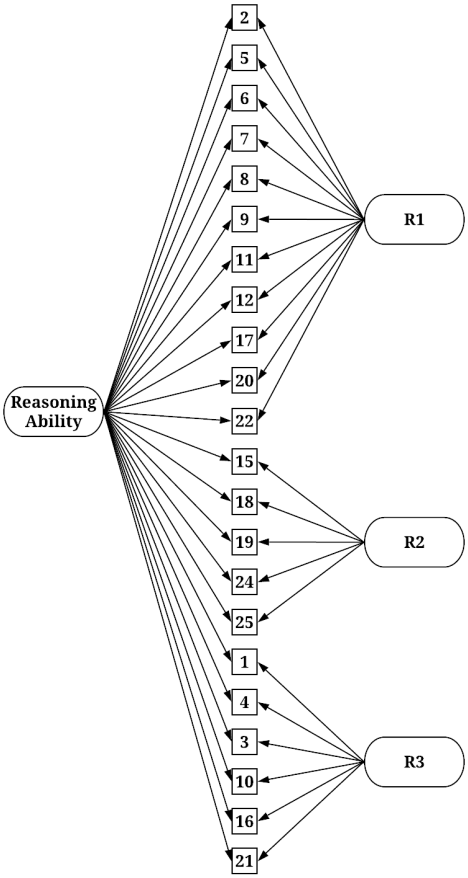
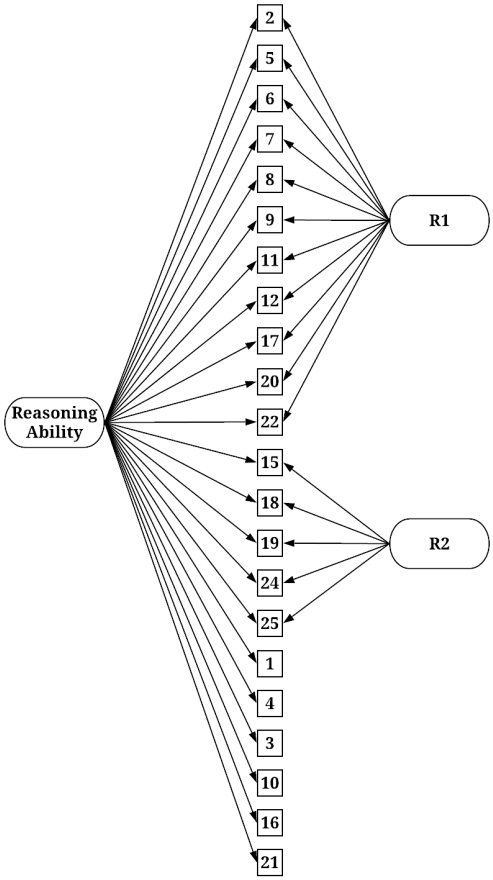


Figure 9

Respecified Bifactor Model



www.manaraa.com

Table 9

Respecified Bifactor Model Standardized Factor Loadings

	Factor loadings		
PCA Item	General Reasoning Ability	Process View of Function (R1)	Covariational Reasoning (R2)
2	.448*	.247*	
5	.568*	.503*	
6	.623*	.438*	
7	.233*	.078	
8	.328*	.184*	
9	.468*	.194*	
11	.108*	.150*	
12	.504*	.482*	
17	.209*	.134*	
20	.258*	.083	
22	.294*	.149*	
15	.416*		.386*
18	.462*		.084
19	.492*		.395*
24	.557*		.201*
25	.180*		-.004
1	.501*		
3	.572*		
4	.520*		
10	.227*		
16	.737*		
21	.240*		

* $p < 0.05$

Maximum Likelihood Estimation

Using the maximum likelihood with robust standard errors (MLR) estimator in Mplus produces the relative fit statistics AIC, BIC, and Sample-Size Adjusted BIC. These fit statistics allow for the comparison of non-nested models with smaller values indicating a better fit (Wang & Wang, 2012). The single-factor and respecified bifactor models were estimated with the MLR to facilitate model comparison using relative fit

www.manaraa.com

statistics. Small differences were found between the relative fit of the three models as shown in Table 10.

Table 10

Relative Fit Statistics Using MLR Estimator

Model	Number of Parameters Estimated	Degrees of Freedom	AIC	BIC	Sample-Size Adjusted BIC
Respecified bifactor	60	193	37290	37609	37419
Three factor	47	206	37340	37590	37441
Single factor	44	209	37406	37640	37500

Reliability Estimates

Reliability analysis using omega (ω) and omega-hierarchical (omegaH or ω_H) was conducted to examine the extent to which the specific factors R1 and R2 have substantive meaning above and beyond the general Reasoning Ability factor (S. P. Reise et al., 2013; A. Rodriguez et al., 2016a, 2016b). Omega estimates the proportion of variance attributed to all systematic sources of variance in the model (i.e., both specific and general factors). High omega values represent the high reliability of scores from the entire multidimensional model. OmegaH, however, estimates only the percent of variance directly attributed to the general factor. The "comparison of omega to omegaH is useful in revealing the degree to which an estimate of reliability is inflated due to multidimensionality" (S. P. Reise et al., 2013, p. 133). In contrast to omegaH, omega hierarchical subscale (omegaHS or ω_{HS}) estimates a single subscale's reliability.

www.manaraa.com

Omega, omegaH, and omegaHS were all estimated using the respecified bifactor model. The computation of the omega reliability statistic resulted in a value of .841. The reliability of the general Reasoning Ability factor was found to have an omegaH value of .763, meaning that 76% of the variance in PCA total scores can be attributed to the general Reasoning Ability factor. The ratio of omegaH to omega is .908, meaning that approximately 91% of the reliable variance can be attributed to the general Reasoning factor. Therefore, only 9% of the total variance can be attributed to the R1 and R2 specific factors. The computation of the omegaHS value further highlights the reliability of this 9% of the variance. The omegaHS values for R1 and R2 were .220 and .122, respectively, indicating that the variances partitioned by the R1 and R2 specific factors had little unique reliable variance.

Explained common variance (ECV; S. Reise et al., 2010) is an indicator that is used to assess unidimensionality. Conceptually, the ECV represents the amount of variance that can be attributed to the general factor out of the total common variance in a bifactor model. ECV values range from 0 to 1, with higher ECV values being indicative of a model with a strong general factor compared to the strength of the specific factors. The ECV value for the respecified bifactor model was .77. A. Rodriguez et al. (2016a) suggested that when ECV values are greater than .70, "the factor loadings obtained from a unidimensional model might approximate well (i.e., be unbiased) the factor loadings on the general factor obtained from a bifactor solution" (p. 231).

The percent of uncontaminated correlations (PUC) "is the number of unique correlations in a correlation matrix that are influenced by a single factor divided by the total number of unique correlations" (A. Rodriguez et al., 2016b, p. 146). The PUC is

www.manaraa.com

another metric that can be used to judge the dimensionality of an instrument. Higher PUC values are indicative of a bifactor model where the general factor can be considered e*ssentially unidimensional*. The PUC for the General Reasoning Ability factor in the revised bifactor model was .72. A PUC of .72 in conjunction with an ECV value of .77 provides additional evidence that the PCA is *essentially unidimensional* (A. Rodriguez et al., 2016a).

Cross-Validation

An analysis was conducted to compare the model fit using the primary (n = 1509) and cross-validation (n = 1509) data subsets using the Mplus DIFFTEST and GROUPING functions. The single-factor model was estimated, and no statistically significant differences were found between the configural and scalar models (χ^2 (20, 3018) = 12.075, p = 0.913). Similar results were found with the three-factor model (χ^2 (16, 3018) = 10.276, p = 0.851) and the respecified bifactor model (χ^2 (32, 3018) = 26.539, p = 0.739). Together, these results provide evidence that the respecified models were not over-specified to the specific subset of data used for their development.

Diagnostic Classification Modeling

Diagnostic Classification Models (DCMs) were used to address the third and fourth research questions. A Log-linear Cognitive Diagnostic Model (LCDM) was fitted to the PCA student response data. The models were estimated using the C-RUM parameterization using the Q-matrix structure implied by the PCA Taxonomy (Table 6). The analysis of this model resulted in a non-positive definite first-order derivative product matrix. The Q-matrix was then respecified based on the results of the factor analysis specification search (Table 11).

www.manaraa.com

Table 11

Respecified Q-Matrix Structure

Item	Process View of Function (R1)	Covariational Reasoning (R2)	Computational Abilities (R3)
1	0	0	1
2	1	0	0
3	0	0	1
4	0	0	1
5	1	0	0
6	1	0	0
7	1	0	0
8	1	0	0
9	1	0	0
10	0	0	1
11	1	0	0
12	1	0	0
15	0	1	0
16	0	0	1
17	1	0	0
18	0	1	0
19	0	1	0
20	1	0	0
21	0	0	1
22	1	0	0
24	0	1	0
25	0	1	0

Analysis of the Respecified Q-Matrix Structure

The analysis using the respecified Q-matrix terminated normally. However, Mplus was unable to estimate chi-square statistics due to the size of the frequency table for latent class portion of the model. The relative fit statistics for this model were; AIC = 70148.39, BIC = 70455.02, and Sample-Size Adjusted BIC = 70292.97. The attribute classification reliabilities (Templin & Bradshaw, 2013) were; R1 = .957, R2 = .788, and R3 = .957. Of the eight possible mastery profiles, this model only classified students into the three mastery profiles shown in Table 12. The attribute correlations ranged from .814 to .973 as shown in Table 13. An inspection of the model parameter estimates revealed that the main effect for R1 on item 12 approached extreme values. As a consequence,

www.manaraa.com

Mplus fixed these main effects to stabilize model estimation. The main effect for R2 on item 25 was not statistically significant (Table 14). All other main effect parameters were found to be statistically significant.

Table 12

Mastery Profiles for Respecified Q-Matrix Structure

Process View of Function (R1)	Covariational Reasoning (R2)	Computational Abilities (R3)	Student Count (%)
Non-master	Non-master	Non-master	2126 (70.44%)
Master	Non-master	Master	194 (6.43%)
Master	Master	Master	698 (23.13%)

Table 13

Attribute Correlations for Revised Q-Matrix Structure

Factor	R1	R2	R3
Process View of Function (R1)	–		
Covariational Reasoning (R2)	.814	–	
Computational Abilities (R3)	.973	.907	–

www.manaraa.com

Table 14

Item Parameters for DCM With Revised Q-Matrix Structure

Item	Intercept	Main Effect (Attribute)	Increase in Log-odds	Increase in Probability
1	-1.42	1.64* (R3)	3.06	0.96
2	-0.41	1.50* (R1)	1.91	0.87
3	-1.47	1.41* (R3)	2.88	0.95
4	-1.69	2.74* (R3)	4.43	0.99
5	-1.91	2.87* (R1)	4.78	0.99
6	-0.82	0.55* (R1)	1.37	0.80
7	-2.59	1.09* (R1)	3.68	0.98
8	-0.99	2.03* (R1)	3.03	0.95
9	-1.25	0.59* (R1)	1.84	0.86
10	-1.65	0.46* (R3)	2.11	0.89
11	-1.76	2.68* (R1)	4.44	0.99
12	-1.85	X (R1)	X	X
15	-2.52	0.67* (R2)	3.19	0.96
16	-1.02	1.19* (R3)	2.21	0.90
17	0.39	4.21* (R1)	3.82	0.98
18	-2.21	1.15* (R2)	3.36	0.97
19	-0.16	1.82* (R2)	1.98	0.88
20	-0.38	1.56* (R1)	1.94	0.87
21	-0.84	0.83* (R3)	1.67	0.84
22	-1.28	0.71* (R1)	1.99	0.88
24	-1.81	1.48* (R2)	3.29	0.96
25	-2.45	0.12 (R2)	2.57	0.93

* $p < 0.05$, X = parameter fixed by Mplus to stabilize model estimation

Analysis of the CFA Bifactor Derived Q-Matrix Structure

The rival Q-matrix structure shown in Table 15 was specified based on the results of the final bifactor model. The analysis of the bifactor derived Q-matrix structure terminated normally. The relative fit statistics for this model were AIC = 69582.10, BIC = 69984.92, and Sample-Size Adjusted BIC = 69772.04. The attribute classification reliabilities for this model were G1 = .873, R1 = .919, and R2 = .908. This model classified students into six of the eight possible mastery profiles (Table 16). The attribute correlations ranged from .325 to .993, as shown in Table 17. An inspection of the item

www.manaraa.com

parameters found that the main effects for attribute R1 on item 20 and R2 on item 19 were not statistically significant (Table 18). Mplus fixed three other parameters to stabilize model estimation.

Table 15

Bifactor Derived Q-Matrix Structure

Item	General Reasoning Ability (G1)	Process View of Function (R1)	Covariational Reasoning (R2)
1	1	0	0
2	1	1	0
3	1	0	0
4	1	0	0
5	1	1	0
6	1	1	0
7	1	1	0
8	1	1	0
9	1	1	0
10	1	0	0
11	1	1	0
12	1	1	0
15	1	0	1
16	1	0	0
17	1	1	0
18	1	0	1
19	1	0	1
20	1	1	0
21	1	0	0
22	1	1	0
24	1	0	1
25	1	0	1

www.manaraa.com

Table 16

Mastery Profiles for Bifactor Derived Q-Matrix Structure

General Reasoning Ability (G1)	Process View of Function (R1)	Covariational Reasoning (R2)	Student Count (%)
Non-master	Non-master	Non-master	1750 (57.99%)
Non-master	Master	Non-master	4 (0.13%)
Non-master	Master	Master	411 (13.62%)
Master	Non-master	Non-master	444 (14.71%)
Master	Master	Non-master	2 (0.07%)
Master	Master	Master	407 (13.49%)

Table 17

Attribute Correlations for Bifactor Derived Q-Matrix Structure

Factor	G1	R1	R2
General Reasoning Ability (G1)	–		
Process View of Function (R1)	.325	–	
Covariational Reasoning (R2)	.415	.993	–

www.manaraa.com

Table 18

Item Parameters for DCM With Bifactor Derived Q-Matrix Structure

Item	Intercept	Main Effect (G1)	Main Effect (R1)	Main Effect (R2)	G1 Increase in Log-odds (Prob.)	R1 Increase in Log-odds (Prob.)	R2 Increase in Log-odds (Prob.)
1	-0.25	1.41*	-----	-----	1.66 (0.84)	-----	-----
2	-1.53	0.69*	1.46*	-----	2.22 (0.90)	2.99 (0.95)	-----
3	-0.52	2.44*	-----	-----	2.96 (0.95)	-----	-----
4	-0.93	1.16*	-----	-----	2.09 (0.89)	-----	-----
5	-0.62	2.30*	0.29*	-----	2.92 (0.95)	0.91 (0.71)	-----
6	-1.66	1.40*	0.70*	-----	3.06 (0.96)	2.36 (0.91)	-----
7	-2.07	0.88*	3.35*	-----	2.95 (0.95)	5.42 (1.00)	-----
8	-2.21	1.47*	2.66*	-----	3.68 (0.98)	4.87 (0.99)	-----
9	-0.90	0.45*	0.41*	-----	1.35 (0.79)	1.31 (0.79)	-----
10	-1.29	0.81*	0.31*	-----	2.10 (0.89)	1.60 (0.83)	-----
11	-2.84	1.41*	1.29*	-----	4.25 (0.99)	4.13 (0.98)	-----
12	-1.23	1.72*	-----	-----	2.95 (0.95)	-----	-----
15	-1.85	X	-----	X	X	-----	X
16	-1.92	1.30*	-----	-----	3.22 (0.96)	-----	-----
17	-1.30	0.54*	0.27*	-----	1.84 (0.86)	1.57 (0.83)	-----
18	-2.57	0.00*	-----	0.64*	2.57 (0.93)	-----	3.21 (0.96)
19	-1.18	1.59*	-----	0.18	2.77 (0.94)	-----	1.36 (0.80)
20	-1.75	0.57*	0.22	-----	2.32 (0.91)	1.97 (0.88)	-----
21	-2.43	0.00*	-----	-----	2.43 (0.92)	-----	-----
22	-2.07	1.10*	2.83*	-----	3.17 (0.96)	4.9 (0.99)	-----
24	0.18	1.64*	-----	X	1.46 (0.81)	-----	X
25	-2.30	0.63*	-----	0.50*	2.93 (0.95)	-----	2.80 (0.94)

* $p < 0.05$, ----- = parameter not specified in model, X = parameter fixed by Mplus to stabilize model estimation

www.manaraa.com

CHAPTER 5

Discussion

The PCA is a widely used and highly regarded instrument designed to assess the constructs specified in the PCA Taxonomy. The overarching purpose for conducting this research was to investigate the internal structure of PCA response data to examine empirical validity evidence (Standards 1.13; AERA et al., 2014). Such validity evidence is essential to support the selection of a score reporting method (i.e., single total score or multiple subscores) (Standards 1.14, 1.15; AERA et al., 2014). A secondary purpose was to investigate the appropriateness of retrofitting PCA response data to diagnostic classification models (DCMs) to produce diagnostic profiles for individual students regarding their mastery and nonmastery status on each trait. The following sections will discuss the research results in the context of this study's four research questions.

Evidence Supporting the Three-Factor Structure

The first research question asked, to what extent does a confirmatory factor analysis of PCA pretest data provide evidence that supports the validity of the three-factor structure implied by the PCA Taxonomy. The fit statistics from the three-factor model analysis provided evidence that the model implied by the PCA Taxonomy fit the data well (Table 19). This evidence supports the validity of the implied multidimensionality of the PCA Taxonomy. However, the high correlations between factors brought into question the three-factor model's ability to adequately differentiate between each of the three reasoning ability constructs. The inability of this model to clearly distinguish between constructs may be influenced by several underlying characteristics, including but not limited to (a) the inherent nature of the constructs, (b) the quality of the items representing these constructs, or (c) inadequacies of the statistical model.

www.manaraa.com

Forcing all crossloadings in the three-factor model to zero may have inflated the factor correlations to the point where the model produces limited evidence of discriminant validity. The purpose for investigating rival statistical models (i.e., a single-factor model, a second-order factor model, and a bifactor model) was to explore the extent to which these models illuminated the interrelationships among the three reasoning constructs.

Table 19

Fit Statistics for Three Rival CFA Models

Model	Number of Parameters Estimated	Chi-square	Degrees of Freedom	CFI	TLI	RMSEA	RMSEA 90% C.I.	SRMR
Respecified bifactor	60	393.310	193	.953	.943	.026	.023 --.030	.048
Three factor	47	410.029	206	.952	.946	.026	.022 -- .029	.049
Single factor	44	503.214	209	.934	.928	.031	.027 -- .034	.054

Evidence Supporting Rival Model Structures

The second research question was contingent upon highly correlated factors in the three-factor model implied by the PCA Taxonomy. The presence of a correlation greater than .85 lead to the investigation of rival models. Because the three-factor model had favorable fit statistics, the investigation of rival models was primarily done to illuminate the interrelationships among the three first-order factors. Each rival model tested a hypothesized alternative internal structure than the structure implied by the PCA Taxonomy. The first rival model tested was the single-factor model. Fitting this model permitted the evaluation of the hypothesis that the three reasoning factors implied by the taxonomy were all representations of a single construct as opposed to three distinct constructs.

www.manaraa.com

The fit statistics for the single-factor model (Table 19) met only two of the four predetermined criteria which indicated it was not an adequate representation of the structure underlying the data. As shown in Table 10 and 20, a comparison of AIC, and BIC fit statistics using the MLR estimator found only small differences between the single-factor model and the three-factor model. The comparison of the single-factor and respecified bifactor model also revealed slight differences between the two models. The inconsistent differences between the AIC and BIC values highlight the impact of model complexity on the calculation of these relative fit statistics. For example, when comparing the three-factor model to the respecified bifactor model, the AIC favors the bifactor model while the BIC favors the three-factor model. These discrepancies are due to the way in which the BIC imposes a greater penalty for model complexity. The unknown distributional properties of the AIC and BIC makes it impossible to define meaningful cutoffs to clearly indicate which model is be than another. The similar relative fit statistics for the single-factor, three-factor, and bifactor models do not produce compelling evidence for a clearly superior model.

Table 20

Differences in Relative Fit Statistics

Model	Respecified bifactor			Three factor		
	ΔAIC	ΔBIC	ΔABIC	ΔAIC	ΔBIC	ΔABIC
Respecified bifactor	–	–	–			
Three factor	50	-19	22	–	–	–
Single factor	115	30	81	65	49	59

The second rival model tested was the second-order factor model. The purpose of estimating this model was to gain additional insight into the relationship between the three factors. Although the three-factor model and the second-order factor model are considered

www.manaraa.com

equivalent solutions (i.e., produce identical goodness of fit statistics; Brown, 2015, p. 179), the second-order factor model provided additional parameters and statistics which helped to understand the relationship between factors.

Modeling the second-order model permitted the inspection of the residual variances and R^2 values for the three first-order factors. It was anticipated that all three first-order factors would have large R^2 values and small residual variances, indicating that a large portion of their variance was being accounted for by the second-order factor. However, it was not anticipated that the R^2 value for the R3 factor in the second-order model would be as large as .963 and the residual variance as small as .037. These results indicate that almost all of the variance of the R3 factor was accounted for by the second-order factor. The additional information gained from the second-order model regarding the R3 factor proved useful in the specification of the respecified bifactor model, which empirically collapsed the specific R3 factor into the general reasoning ability factor.

The third rival was the bifactor model. The initial attempt to estimate a bifactor model with one general factor and three specific factors was unsuccessful. The respecified bifactor model collapsed the specific R3 factor into the general factor. The resulting respecified bifactor model hypothesized that all items loading on a general factor, 11 items loading on the R1 specific factor, and 5 items loading on the R2 specific factor (Figure 9). Fit statistics indicated that this model fit the data well (Table 19) and slightly better than the other rival models.

The bifactor model provided the advantage of partitioning the shared and unique variance for each item contributing to a specific factor. This partitioning allowed for the inspection of specific factors as a unique component of the model isolated from the shared characteristics of the general factor. Although the revised bifactor model was a suitable alternative to the three-

www.manaraa.com

factor model, the factor loadings on many of the specific factor items were low. These low factor loadings may be an indication that the specific factors do not have substantive meaning apart from the general Reasoning Ability factor.

Another notable advantage of the bifactor model was the ability to use model-based reliability estimates such as omega (ω), omega-hierarchical (omegaH or ω_H), and omega hierarchical subscale (omegaHS or ω_{HS}) to address questions about the degree to which the general factor dominated the specific factors. The omega coefficient was .841 representing the proportion of variance attributed to both general and specific factors. The omega-hierarchical coefficient was .763, which is the proportion of the variance attributed solely to the general factor. The ratio of omegaH to omega revealed that approximately 91% of the overall reliability was attributed to the general reasoning factor. The reliability of the remaining 9% of variance attributed to the specific factors R1 and R2 was low (ω_{HS} = .220 and .122 respectively). Accordingly, the specific factors in this model appear to largely represent nuisance variance.

The usefulness of the bifactor model to provide evidence for the use of subscores depended on the extent to which item variances were accounted for by the general factor as opposed to a specific factor. One indication that the specific factors may not be accounting for a large portion of item variance was the standardized factor loadings (Table 9). In all but one instance, the standardized loadings for the specific factors were lower than their corresponding loading on the general factor. These differences were, on average, .127 (R1) and .209 (R2).

The investigation into each of the rival models provided unique insights into the internal structure of the PCA pretest data. All of the models had adequate model fit with only slight differences. The use of the MLR estimator and the associated AIC, BIC, and ABIC highlighted

www.manaraa.com

how close each of these models were in terms of model fit. Where no one model clearly outperforms another, the most parsimonious model is often selected as the preferred model. In this study, the parsimonious nature of the single-factor model makes it the preferred model.

PCA Scoring and Interpretation

The investigation of the internal structure of PCA response data has specific implications for the validity of practices used to score and interpret PCA results. The initial analysis of the three-factor model revealed a strong relationship between the three reasoning ability constructs. Ignoring the high factor correlations and reporting subscores for each of the three constructs would make it difficult to interpret the meaning of these scores. Therefore, PCA subscores should not be reported for each of the three reasoning ability constructs.

The rival bifactor model results further supported reporting a single total score. The several non-significant specific factor loadings and other factor loadings well below acceptable levels indicated that the specific factors did not account for a large portion of the variance above and beyond the general factor. Reise et al. (2010) emphasized the implications of low specific factor loadings when they wrote, "[t]o the degree that the items reflect primarily the general factor and have low loadings on the group factors, subscales make little sense" (p. 555). Furthermore, the reliability analysis revealed that the unique variance attributed to the specific factors was small and unreliable. The small differences between the fit the single-factor model and other rival models supports the use of a single total score.

Due to the lack of validity evidence to support the use of subscores, the use of a single total score on the PCA may be the most psychometrically defensible method of scoring the instrument. These findings further support the assertion by Carlson et al. (2010) that "it would

www.manaraa.com

not be appropriate to draw inferences about the abilities of an individual student relative to PCA subscores" (p. 137).

Retrofitting the PCA for Diagnostic Classification Modeling

Diagnostic classification models were retrofitted to PCA data to investigate the ability of these models to provide diagnostic results for individual students. The LCDM with the constrained parameterization of the C-RUM model limited the estimation of item parameters to main effects only. Even with the reduced computational requirements of C-RUM, the model with the Q-matrix structure implied by the PCA Taxonomy (Table 6) resulted in a non-positive definite first-order derivative product matrix. Consequently, the Q-matrix was respecified (Table 11) based on the simplified item to attribute relationship found in the three-factor CFA model.

One notable difference between the Q-matrix structure implied by the PCA Taxonomy and the structure derived from the three-factor CFA model is the absence of multidimensional items and the removal of items 13, 14, and 23 from the data set. The presence of only simple structure items (i.e., items associated with only one attribute) reduced the model's complexity by eliminating the need to estimate a main effect for each of the removed item-to-attribute associations.

The research on using absolute fit statistics with DCMs is inconclusive and has limited availability in statistical software (Ravand & Baghaei, 2019). Unfortunately, chi-square, the only absolute fit statistic available when estimating DCMs in Mplus, was not successfully estimated. However, an evaluation of the model's overall performance was still conducted using relative fit statistics, attribute classification reliabilities, attribute profiles and mastery proportions, and attribute correlations.

www.manaraa.com

The high reliabilities of the R1 and R3 attributes provided evidence that students were consistently being classified as masters or non-masters of the attributes. The lower reliability of the R2 factor suggests that the model struggled more to establish consistent classifications for this attribute. Overall, 31% of students were classified as masters of R1 and R3, while 20% were classified as masters of R2. Although these findings provide evidence of the consistency of attribute classifications, they do not provide evidence that the classifications were correct.

An inspection of the attribute profiles and mastery proportions facilitates the evaluation of the model's estimated proficiencies. One of the claimed advantages of DCMs is that students may be assigned a mastery profile, which classifies students as masters or non-masters of specified attributes. A DCM with three attributes has the potential to assign students to one of eight mastery profiles. The DCM using the respecified Q-matrix only classified students into three of the eight possible profiles (Table 11). The majority of students (94%) were classified into two profiles, either mastering all or none of the three attributes. The inability of the model's estimation of attribute profiles could reflect the realities of students' true mastery of the theorized attributes or inadequacies of model estimation. In the absence of absolute fit statistics, it is difficult to pinpoint why students were assigned to such a limited number of mastery profiles.

An inspection of the item parameters (Table 14) provided additional insight into the relationship between attribute mastery and correct item response. The item parameters can be used to determine the estimated increase in the log-odds of a correct item response. This increase is calculated by taking the difference between the item intercept and the main effect. The increase in log-odds can also be represented as a probability. All main effects were statistically significant except for item 25 on the R2 attribute. These statistically significant main effects

www.manaraa.com

provide evidence that the item to attribute association in the Q-matrix was not misaligned and contributed to the estimation of attribute mastery.

The high attribute correlations for this model were concerning (Table 13). The correlation between R3 and the other two attributes were both above .90. These high correlations have potential implications for the estimation of mastery profiles in which a student is classified as a master of one but not the other. The estimated mastery profiles found in Table 12 highlighted that a very small percentage of students (6.43%) did not have the same mastery classification for all three attributes. Meaning, the majority of students were either classified as having mastered all or none of the attributes. This finding resurfaces the question of whether or not this model can successfully estimate distinct classifications mastery for each attribute. The strong relationship between the attributes appears to largely go beyond the model's ability to parse distinct attribute mastery.

The analysis of the Q-matrix structure derived from the bifactor model (Table 15) found slight differences from the respecified Q-matrix structure (Table 11). The relative fit statistics for the bifactor derived Q-matrix were consistently smaller than the respecified Q-matrix structure (Table 21). However, these differences were relatively small (ΔAIC = 566, ΔBIC = 470, ΔSample-Size Adjusted BIC = 521). The percent improvement in model fit was less than 1% on all fit statistics ranging from 0.67% to 0.81%. These slight improvements in model fit are not, on their own, reason to suppose that the bifactor Q-matrix structure was superior.

The attribute classification reliabilities for the bifactor derived Q-matrix structure were not notably different from the revised Q-matrix structure. The use of the bifactor derived Q-matrix structure did have a large impact on the attribute correlations (Table 17).

www.manaraa.com

Table 21

Fit Statistics for Rival Q-Matrix Structures

Model	AIC	BIC	Sample-Size Adjusted BIC
Respecified Q-matrix	70148.39	70455.02	70292.97
Bifactor Derived Q-matrix	69582.10	69984.92	69772.04

The attribute correlations between G1 and the other two attributes were low enough to suggest that G1 was distinct from these other attributes. The extreme correlation (.993) between R1 and R2 is very concerning and suggests that the model could not differentiate between masters of these two attributes adequately.

The bifactor derived Q-matrix structure estimated an increased number of mastery profiles (Table 16). Students were classified into six of the eight possible mastery profiles using this Q-matrix structure. As expected, based on the attribute correlations between R1 and R2, very few students did not receive the same classification for both attributes (n = 6, 0.2%). Another notable difference with this Q-matrix structure was the reduction of students classified as masters of all attributes (n = 231). Item parameterization for this Q-matrix structure was more unstable with three main effects fixed by Mplus to stabilize model estimation. The number of non-statistically significant main effects increased from one to three with the new Q-matrix structure.

Use of Diagnostic Classification Models for PCA Mastery Profiles

The attempt to retrofit PCA response data to a DCM using the Q-matrix structure implied by the PCA Taxonomy was unsuccessful. The rival respecified and bifactor derived Q-matrix structures were estimated with limited success. Although some aspects of these models appeared to function well, attribute correlations and mastery profile estimates highlighted some fundamental deficiencies of these models. These model deficiencies, combined with the lack of

www.manaraa.com

an ability to inspect an absolute model fit index led to the recommendation that the C-RUM parameterization of the LCDM not be used to generate mastery profiles.

Limitations

While evidence of good model fit was presented in the factor analysis portion of this study, these models were the result of a specification search and did not directly represent the factor and Q-matrix structures implied by the PCA Taxonomy. The revised three-factor structure reduced the complexity of the model by removing three items from the data set and removing all cross-loadings. Therefore, the results of this study are based on a close approximation of the model implied by the PCA Taxonomy.

A second limitation of this study was the exclusive use of pretest PCA scores. It is reasonable to believe that score distributions may be different when posttest PCA scores are evaluated. The results of this study should only be considered in the context of pretest PCA data.

Recommendations for Future Research

There are several areas in which the results of this study could inform future research. Future research on or using the PCA data should:

1. Consider the use of both pretest and posttest data.
2. Modify or develop items with the specific purpose of being analyzed by a DCM.
3. Attempt to replicate the results of the factor analysis portion of this study.
4. Use the more extensive techniques of structural equation modeling (SEM) for predictive research.
5. Limit the PCA measurement model in SEM to either the general factor of the bifactor model or the single-factor model.

www.manaraa.com

Future psychometric research of the PCA should include a consideration of both pretest and posttest data. It is currently unknown if the models presented in this study are invariant across pretest and posttest occasions. Several of the items in the pretest context were extremely difficult for students which may have attenuated variances and covariances between items. An investigation of posttest data might find that item difficulty diminishes due to a semester of mathematics instruction. The results from a study of measurement invariance could provide additional insight into the validity of the internal structure of PCA response data.

The results of this study found that retrofitting the PCA to a specific DCM parameterization was not successful. However, this finding does not entirely preclude the use of DCMs to provide personalized mastery profiles to students. Suppose there is a need or desire to provide mastery profiles. In that case, researchers should modify or develop items with the specific purpose of being analyzed by a DCM (e.g., Bradshaw et al., 2014). This process would include many of the same quality test development procedures used in developing the PCA and specific design considerations uniquely related to DCMs, including Q-matrix design and DCM parameterization (Bradshaw, 2017; Madison & Bradshaw, 2015; Sessoms & Henson, 2018).

Future research into the internal structure of PCA data should attempt to replicate the findings of the factor analysis portion of this study. Conducting a replication study would provide an opportunity to verify the generalizability of the findings in this study. A replication study would be particularly valuable in the context of the more complex bifactor which, from time to time, can be difficult to replicate.

www.manaraa.com

In their description of the PCA, Carlson et al. wrote that the relationship between PCA scores and student performance in calculus had been investigated (2010, pp. 140–141). This study's results have implications for future research into the relationship between PCA scores and other outcome variables such as performance in a calculus course. Future research should use the more extensive techniques of structural equation model (SEM) in conjunction with the CFA models investigated in this study.

The factor loadings for the General Reasoning Ability factor (Table 9) highlighted that each item of the PCA does not contribute equally to the measurement of student reasoning abilities. The respecified bifactor model (Figure 9) notably has only two specific factors (R1, R2) and the absence of items 13, 14, and 23. The respecified bifactor model analysis highlighted that the two specific R1 and R2 factors accounted for a limited amount of unique item variance (ECV = .77, PUC = .72). These results showed that although there is some degree of multidimensionality to the PCA, it should be considered *essentially unidimensional.* Therefore, researchers should only use the General Reasoning Ability factor from the respecified bifactor model (Figure 9) or the single-factor model as an outcome predictor in a structural equation model.

Recommendations for Practice

The recommended practice for scoring the PCA by the authors of the instrument was to report a single sum score (i.e., unit-weighted composite score) for the PCA as a "broad indicator of reasoning abilities and understandings relative to the PCA Taxonomy" (Carlson et al., 2010, p. 137). This research's findings provided evidence to support the recommendation of reporting a single total score for the PCA. From the perspective of the instrument's internal structure, this study found limited evidence to

www.manaraa.com

support the validity of reporting subscores of student reasoning abilities. To the best of the author's knowledge, reporting a single total score for the PCA is currently the most common scoring method. As such, practitioners should continue with this scoring approach.

It is important to note that the analyses reported in this study were conducted in the absence of items 13, 14, and 23. However, it is recommended that in practice, these items be retained for content validity purposes while computing a total score for the PCA.

Conclusion

This study explored the PCA's internal structure in relation to the structure implied by the PCA Taxonomy. CFA was used to investigate the extent to which PCA pretest data supports the three-factor structure theorized by the PCA Taxonomy. Results found that overall the model fit the data well, but high factor correlations brought into question the distinct nature of each factor. A rival bifactor model sought to illuminate the interrelationships among the three factors by allowing all items to load on a general factor and two specific factors. The fit statistics of the bifactor model were only slightly more favorable than the fit statistics of the three-factor model. However, low specific factor loadings and low reliability lead to the conclusion that these factors were not substantively modeling the constructs. Although there is some level of multidimensionality, the single-factor model appears to be the most parsimonious approach to modeling the internal structure of PCA pretest data. In summary, these results suggest that a single composite total score based on all 25 items be reported when the PCA is administered.

An additional analysis was conducted to examine the extent to which PCA response data could be retrofitted to a DCM. The purpose of this research was to explore the potential of using DCMs to provide students with individual mastery profiles for the three reasoning ability

www.manaraa.com

constructs found in the PCA Taxonomy. Several different Q-matrix structures were examined, but all were unsuccessful in providing adequate evidence to support the use of student mastery profiles generated from a DCM retrofitted with PCA data.

www.manaraa.com

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723. https://doi.org/10.1109/TAC.1974.1100705

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*(3), 317-332. https://doi.org/10.1007/978-1-4612-1694-0_29

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Anderson, D. L., Fisher, K. M., & Norman, G. J. (2002). Development and evaluation of the Conceptual Inventory of Natural Selection. *Journal of Research in Science Teaching*, *39*(10), 952–978. https://doi.org/10.1002/tea.10053

Bailey, J. M., Johnson, B., Prather, E. E., & Slater, T. F. (2012). Development and validation of the Star Properties Concept Inventory. *International Journal of Science Education*, *34*(14), 2257–2286. https://doi.org/10.1080/09500693.2011.589869

Bannerjee, P. (2017). Students' understanding of the concepts of rates of change and functions. In T. A. Olson & L. Venenciano (Eds.), Proceedings of the 44th Annual Meeting of the Research Council on Mathematics Learning, Fort Worth, TX.

Beavers, A. S., Lounsbury, J. W., Richards, J. K., Huck, S. W., Skolits, G. J., & Esquivel, S. L. (2013). Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research, & Evaluation*, *18*, 6. https://doi.org/10.7275/qv2q-rk76

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

www.manaraa.com

Bentler, P. M. (2007). On tests and indices for evaluating structural models. *Personality and Individual Differences*, *42*(5), 825–829. https://doi.org/10.1016/j.paid.2006.09.024

Borsboom, D., Romeijn, J.-W., & Wicherts, J. M. (2008). Measurement invariance versus selection invariance: Is fair selection possible? *Psychological Methods*, *13*(2), 75–98. https://doi.org/10.1037/1082-989X.13.2.75

Bowling, B. V., Acra, E. E., Wang, L., Myers, M. F., Dean, G. E., Markle, G. C., Moskalik, C. L., & Huether, C. A. (2008). Development and evaluation of a genetics literacy assessment instrument for undergraduates. *Genetics*, *178*(1), 15–22. https://doi.org/10.1534/genetics.107.079533

Bradshaw, L. (2017). Diagnostic classification models. In A. A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 297–327). John Wiley & Sons. https://doi.org/10.1002/9781118956588

Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues & Practice*, *33*(1), 2–14. https://doi.org/10.1111/emip.12020

Bretz, S. L., & Linenberger, K. J. (2012). Development of the Enzyme-Substrate Interactions Concept Inventory. *Biochemistry and Molecular Biology Education*, *40*(4), 229–233. https://doi.org/10.1002/bmb.20622

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

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. Bollen & J. Long (Eds.), *Testing structural equation models* (pp. 136–162). Sage.

www.manaraa.com

Carlson, M. (1995). *A cross-sectional investigation of the development of the function concept.* ProQuest LLC, Ann Arbor, MI.

Carlson, M., Oehrtman, M., & Engelke, N. (2010). The precalculus concept assessment: A tool for assessing students' reasoning abilities and understandings. *Cognition & Instruction*, *28*(2), 113–145. https://doi.org/10.1080/07370001003676587

Cetin-Dindar, A., & Geban, O. (2011). Development of a three-tier test to assess high school students' understanding of acids and bases. *Procedia-Social and Behavioral Sciences*, *15*, 600–604. https://doi.org/10.1016/j.sbspro.2011.03.147

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research*, *41*(2), 189–225.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. https://doi.org/10.3102/10769986022003265

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of q-matrix based diagnostic classification models. *Journal of the American Statistical Association*, *110*(510), 850–866. https://doi.org/10.1080/01621459.2014.934827

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

www.manaraa.com

Chiu, C.-Y. (2013). Statistical refinement of the q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, *37*(8), 598–618. https://doi.org/10.1177/0146621613488436

Cho, P., & Nagle, C. (2017). Procedural and conceptual difficulties with slope: An analysis of students' mistakes on routine tasks. *International Journal of Research in Education and Science*, *3*(1), 135–150.

Cloonan, C. A., & Hutchinson, J. S. (2011). A chemistry concept reasoning test. *Chemistry Education Research and Practice*, *12*(2), 205–209. https://doi.org/10.1039/C1RP90025K

Cousino, A. (2013). *Using bayesian learning to classify college algebra students by understanding in real-time* [Doctoral dissertation, Kansas State University]. K-State Research Exchange. http://hdl.handle.net/2097/15630

Cromley, J. G., Booth, J. L., Wills, T. W., Chang, B. L., Tran, N., Madeja, M., Shipley, T. F., & Zahner, W. (2017). Relation of spatial skills to calculus proficiency: A brief report. *Mathematical Thinking and Learning*, *19*(1), 55–68. https://doi.org/10.1080/10986065.2017.1258614

de la Torre, J. (2008). An empirically based method of q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*(4), 343–362. https://doi.org/10.1111/j.1745-3984.2008.00069.x

de la Torre, J. (2011). The Generalized DINA Model Framework. *Psychometrika*, *76*(2), 179–199. https://doi.org/10.1007/s11336-011-9207-7

DeCarlo, L. T. (2012). Recognizing uncertainty in the q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement*, *36*(6), 447–468. https://doi.org/10.1177/0146621612449069

DeVellis, R. F. (2012). *Scale development: Theory and applications* (3rd ed.). SAGE.

www.manaraa.com

DiBello, L. V., Roussos, L. A., & Stout, W. F. (2007). 31A Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics (Vol. 26): Psychometrics* (pp. 979–1030). Elsevier. https://doi.org/10.1016/S0169-7161(06)26031-0

DiStefano, C., & Hess, B. (2005). Using confirmatory factor analysis for construct validation: An empirical review. *Journal of Psychoeducational Assessment*, *23*(3), 225–241. https://doi.org/10.1177/073428290502300303

Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. Oxford University Press.

Fager, M., Pace, J., & Templin, J. L. (2019). Using Mplus to estimate the log-linear cognitive diagnosis model. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models: Models and model extensions, applications, software packages* (pp. 581–591). Springer International Publishing. https://doi.org/10.1007/978-3-030-05584-4_28

Fisher, K. M., Williams, K. S., & Lineback, J. E. (2011). Osmosis and diffusion conceptual assessment. *CBE Life Sciences Education*, *10*(4), 418–429. https://doi.org/10.1187/cbe.11-04-0038

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*(4), 466–491. https://doi.org/10.1037/1082-989X.9.4.466

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, *7*(3), 286–299. https://doi.org/10.1037/1040-3590.7.3.286

www.manaraa.com

Gaskin, C. J., & Happell, B. (2014). On exploratory factor analysis: A review of recent evidence, an assessment of current practice, and recommendations for future use. *International Journal of Nursing Studies*, *51*(3), 511–521. https://doi.org/10.1016/j.ijnurstu.2013.10.005

Gierl, M. J., & Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment. *Measurement*, *6*(4), 263–268. https://doi.org/10.1080/15366360802497762

Gorin, J. S. (2009). Diagnostic classification models: Are they necessary? Commentary on Rupp and Templin (2008). *Measurement: Interdisciplinary Research and Perspectives*, *7*(1), 30–33. https://doi.org/10.1080/15366360802715387

Guenole, N., & Brown, A. (2014). The consequences of ignoring measurement invariance for path coefficients in structural equation models. *Frontiers in Psychology*, *5*, 980. https://doi.org/10.3389/fpsyg.2014.00980

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*(4), 301–321. https://doi.org/10.1111/j.1745-3984.1989.tb00336.x

Haig, B. D. (2005). Exploratory factor analysis, theory generation, and scientific method. *Multivariate Behavioral Research*, *40*(3), 303–329. https://doi.org/10.1207/s15327906mbr4003_2

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Lawrence Erlbaum Associates Publishers.

www.manaraa.com

Hansen, M., Cai, L., Monroe, S., & Li, Z. (2016). Limited-information goodness-of-fit testing of diagnostic classification item response models. *The British Journal of Mathematical and Statistical Psychology*, *69*(3), 225–252. https://doi.org/10.1111/bmsp.12074

Harrington, D. (2009). *Confirmatory factor analysis*. Oxford University Press.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* [Doctoral dissertation, University of Illinois Urbana-Champaign]. APA PsycInfo. https://psycnet.apa.org/record/2002-95016-234

Henson, R. A. (2009). Diagnostic classification models: Thoughts and future directions. *Measurement: Interdisciplinary Research and Perspectives*, *7*(1), 34–36. https://doi.org/10.1080/15366360802715395

Henson, R. A., Templin, J., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191–210. https://doi.org/10.1007/s11336-008-9089-5

Herman, G. L. (2011). *The development of a digital logic concept inventory* [Doctoral dissertation, University of Illinois at Urbana-Champaign]. Illinois Digital Environment for Access to Learning and Scholarship. http://hdl.handle.net/2142/24134

Hestenes, D., & Halloun, I. (1995). Interpreting the Force Concept Inventory: A response to March 1995 critique by Huffman and Heller. *Physics Teacher*, *33*(8), 502-506. https://doi.org/10.1119/1.2344278

Hestenes, D., & Wells, M. (1992). A mechanics baseline test. *Physics Teacher*, *30*(3), 159–166. https://doi.org/10.1119/1.2343498

www.manaraa.com

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *Data Provides Commonsense Misconceptions in Introductory Physics*, *30*(3), 141–158. https://doi.org/10.1119/1.2343497

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424–453. https://doi.org/10.1037/1082-989X.3.4.424

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 19–60). Cambridge University Press. https://doi.org/10.1017/CBO9780511611186.002

Jackson, D. L., Gillaspy, J. A. J., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, *14*(1), 6–23. https://doi.org/10.1037/a0014694

Joreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*(2, Pt.1), 183–202. https://doi.org/10.1007/BF02289343

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258-272. https://doi.org/10.1177/01466210122032064

Jurich, D. P. (2015). *Assessing model fit of multidimensional item response theory and diagnostic classification models using limited-information statistics* (Publication No.

www.manaraa.com

3620468) [Doctoral dissertation, James Madison University]. ProQuest Dissertations and Theses. https://search.proquest.com/openview/c50b6a1acb692cecab258affaf97a927/1?pq-origsite=gscholar&cbl=18750&diss=y

Kline, R. B. (2011). *Principles and practice of structural equation modeling*. Guilford Press.

Krishnan, S. R., & Howe, A. C. (1994). The mole concept: Developing an instrument to assess conceptual understanding. *Journal of Chemical Education*, *71*(8), 653–655. https://doi.org/10.1021/ed071p653

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation*, *35*(2–3), 64–70. https://doi.org/10.1016/j.stueduc.2009.10.003

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*(1), 59–81. https://doi.org/10.1111/j.1745-3984.2011.00160.x

Lane, S., Raymond, M. R., & Haladyna, T. M. (Eds.). (2016). *Handbook of test development* (2nd Edition). Routledge.

Lane, S., Raymond, M. R., Haladyna, T. M., & Downing, S. M. (2016). Test development process. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 3–18). Routledge.

www.manaraa.com

LaRue, R. (2017). *An analysis of student approaches to solving optimization problems in first semester calculus* [Doctoral dissertation, Eberly College of Arts and Sciences]. The Research Repository @ WVU. https://doi.org/10.33915/etd.6040

Lei, P.-W., & Li, H. (2016). Performance of fit indices in choosing correct cognitive diagnostic models and q-matrices. *Applied Psychological Measurement*, *40*(6), 405–417. https://doi.org/10.1177/0146621616647954

Leighton, J. P. (2008). Where's the psychology? a commentary on unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, *6*(4), 272–275. https://doi.org/10.1080/15366360802497804

Leontyev, A. (2016). *Development of a stereochemistry concept inventory* [Doctoral dissertation, University of Northern Colorado]. Scholarship & Creative Works @ Digital UNC. http://digscholarship.unco.edu/dissertations/33

Lindell, R. S., & Olsen, J. P. (2002, August 7-8). *Developing the lunar phases concept inventory* [Paper presentation]. Pysics Education Research Conference, Boise, ID. https://www.compadre.org/per/items/detail.cfm?ID=4323&Relations=1

Liu, J., Xu, G., & Ying, Z. (2013). Theory of self-learning q-matrix. *Bernoulli*, *19*(5A), 1790–1817. https://doi.org/10.3150/12-BEJ430

Liu, R., Huggins-Manley, A. C., & Bulut, O. (2017). Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educational and Psychological Measurement*, *78*(3), 357–383. https://doi.org/10.1177/0013164416685599

Liu, Y., Tian, W., & Xin, T. (2016). An application of M2 statistic to evaluate the fit of cognitive diagnostic models. *Journal of Educational and Behavioral Statistics*, *41*(1), 3–26. https://doi.org/10.3102/1076998615621293

www.manaraa.com

Lubke, G. H., Dolan, C. V, Kelderman, H., & Mellenbergh, G. J. (2003). Weak measurement invariance with respect to unmeasured variables: An implication of strict factorial invariance. *The British Journal of Mathematical and Statistical Psychology*, *56*(2), 231-248. https://doi.org/10.1348/000711003770480020

MacCallum, R. C. (2003). 2001 presidential address: Working with imperfect models. *Multivariate Behavioral Research*, *38*(1), 113–139. https://doi.org/10.1207/S15327906MBR3801_5

MacCallum, R. C., Roznowski, M., & Necowitz, L. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*(3), 490–504.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84–99. https://doi.org/10.1037/1082-989X.4.1.84

Madison, M. J., & Bradshaw, L. P. (2015). The effects of q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement*, *75*(3), 491–511. https://doi.org/10.1177/0013164414539162

Marbach-Ad, G., McAdams, K. C., Benson, S., Briken, V., Cathcart, L., Chase, M., El-Sayed, N. M., Frauwirth, K., Fredericksen, B., Joseph, S. W., Lee, V., McIver, K. S., Mosser, D., Quimby, B. B., Shields, P., Song, W., Stein, D. C., Stewart, R., Thompson, K. V, & Smith, A. C. (2010). A model for using a concept inventory as a tool for students' assessment and faculty professional development. *CBE - Life Sciences Education*, *9*(4), 408–416. https://doi.org/10.1187/cbe.10-05-0069

Marfai, F. S. (2016). *Characterizing teacher change through the perturbation of pedagogical goals* [Doctoral dissertation, Arizona State University]. ASU Electronic Theses and

www.manaraa.com

Dissertations. https://repository.asu.edu/attachments/170440/content/Marfai_asu_0010E_15940.pdf

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement*, *11*(3), 71–101. https://doi.org/10.1080/15366367.2013.831680

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713–732. https://doi.org/10.1007/s11336-005-1295-9

McClary, L. M., & Bretz, S. L. (2012). Development and assessment of a diagnostic tool to identify organic chemistry students' alternative conceptions related to acid strength. *International Journal of Science Education*, *34*(15), 2317–2341. https://doi.org/10.1080/09500693.2012.684433

McDonald, R. P., & Mok, M. M.-C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, *30*(1), 23–40. https://doi.org/10.1207/s15327906mbr3001_2

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*(3), 568–592. https://doi.org/10.1037/0021-9010.93.3.568

www.manaraa.com

Mejia-Ramos, J. P., Lew, K., de la Torre, J., & Weber, K. (2017). Developing and validating proof comprehension tests in undergraduate mathematics. *Research in Mathematics Education*, *19*(2, SI), 130–146. https://doi.org/10.1080/14794802.2017.1325776

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2013). *Measurement and assessment in teaching* (11th ed.). Pearson.

Mulford, D. R., & Robinson, W. R. (2002). An inventory for alternate conceptions among first-semester general chemistry students. *Journal of Chemical Education*, *79*(6), 739-744. https://doi.org/10.1021/ed079p739

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide: Statistical analysis with latent variables, user's guide* (8th ed.). Muthén & Muthén.

Ravand, H., & Baghaei, P. (2019). Diagnostic classification models: Recent developments, practical issues, and prospects. *International Journal of Testing*, *20*(1), 24–56. https://doi.org/10.1080/15305058.2019.1588278

Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modeling*. Routledge.

Reise, S., Moore, T., & Haviland, M. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, *92*(6), 544–559. https://doi.org/10.1080/00223891.2010.496477

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, *95*(2), 129–140. https://doi.org/10.1080/00223891.2012.725437

www.manaraa.com

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016a). Applying difactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment*, *98*(3), 223–237. https://doi.org/10.1080/00223891.2015.1089249

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016b). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(2), 137–150. https://doi.org/10.1037/met0000045

Rodriguez, M. C. (2016). Selected-response item development. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 259–273). Routledge.

Rupp, A. A., & Templin, J. (2008a). The effects of q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational & Psychological Measurement*, *68*(1), 78–96. https://doi.org/10.1177/0013164407301545

Rupp, A. A., & Templin, J. (2008b). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives*, *6*(4), 219–262. https://doi.org/10.1080/15366360802490866

Rupp, A. A., & Templin, J. (2009). The (un)usual suspects? A measurement community in search of its identity. *Measurement: Interdisciplinary Research and Perspectives*, *7*(2), 115–121. https://doi.org/10.1080/15366360903187700

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications.* Guilford Press.

Sadler, P. M., Coyle, H., Miller, J. L., Cook-Smith, N., Dussault, M., & Gould, R. R. (2009). The astronomy and space science concept inventory: Development and validation of assessment instruments aligned with the k-12 national science standards. *Astronomy Education Review*, *8*(1), 1–26. https://doi.org/10.3847/AER2009024

www.manaraa.com

Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling*, *21*(2), 167–180. https://doi.org/10.1080/10705511.2014.882658

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment*, *29*(4), 304–321. https://doi.org/10.1177/0734282911406653

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461-464. https://doi.org/10.1214/aos/1176344136

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*(3), 333–343. https://doi.org/10.1007/BF02294360

Sen, S., & Bradshaw, L. (2017). Comparison of relative fit indices for diagnostic model selection. *Applied Psychological Measurement*, *41*(6), 422–438. https://doi.org/10.1177/0146621617695521

Sessoms, J., & Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary. *Measurement: Interdisciplinary Research and Perspectives*, *16*(1), 1–17. https://doi.org/10.1080/15366367.2018.1435104

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic models a case study. *Educational and Psychological Measurement*, *67*(2), 239–257. https://doi.org/10.1177/0013164406292025

Sinharay, S., & Haberman, S. J. (2009). How much can we reliably know about what examinees know? *Measurement: Interdisciplinary Research and Perspectives*, *7*(1), 46–49. https://doi.org/10.1080/15366360802715486

www.manaraa.com

Smith, M. K., Wood, W. B., & Knight, J. K. (2008). The genetics concept assessment: A new concept inventory for gauging student understanding of genetics. *CBE - Life Sciences Education*, *7*(4), 422–430. https://doi.org/10.1187/cbe.08-08-0045

Stanhope, L., Ziegler, L., Haque, T., Le, L., Vinces, M., Davis, G. K., Zieffler, A., Brodfuehrer, P., Preest, M., & Belitsky, J. M. (2017). Development of a biological science quantitative reasoning exam (BioSQuaRE). *CBE-Life Sciences Education*, *16*(4), ar66. https://doi.org/10.1187/cbe.16-10-0301

Steif, P. S., & Dantzler, J. A. (2005). A statics concept inventory: Development and psychometric analysis. *Journal of Engineering Education*, *94*(4), 363–371. https://doi.org/10.1002/j.2168-9830.2005.tb00864.x

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*(4), 345–354.

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, *30*(2), 251–275. https://doi.org/10.1007/s00357-013-9129-4

Templin, J., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287–305. https://doi.org/10.1037/1082-989X.11.3.287

Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice*, *32*(2), 37–50. https://doi.org/10.1111/emip.12010

Thomas, M., & Lozano, G. (2012). *Analyzing calculus concept inventory gains in introductory calculus*. Research in Undergraduate Mathematics Education.

www.manaraa.com

http://pzacad.pitzer.edu/~dbachman/rume_xvi_linked_schedule/rume16_submission_95.pdf

Tinsley, H. E., & Tinsley, D. J. (1987). Uses of factor analysis in counseling psychology research. *Journal of Counseling Psychology*, *34*(4), 414-424. https://doi.org/10.1037/0022-0167.34.4.414

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10. https://doi.org/10.1007/BF02291170

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*(Pt 2), 287–307. https://doi.org/10.1348/000711007X193957

von Davier, M. (2014). The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM). *ETS Research Report Series*, *2014*(2), 1–13. https://doi.org/10.1002/ets2.12043

Voska, K. W., & Heikkinen, H. W. (2000). Identification and analysis of student conceptions used to solve chemical equilibrium problems. *Journal of Research in Science Teaching*, *37*(2), 160–176. https://doi.org/10.1002/(SICI)1098-2736(200002)37:2<160::AID-TEA5>3.0.CO;2-M

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus*. John Wiley & Sons.

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231). The Guilford Press.

www.manaraa.com

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, *34*(6), 806–838. https://doi.org/10.1177/0011000006288127

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125–145. https://doi.org/10.1177/014662168400800201

Yu, C.-Y. (2002). *Evaluation of model fit indices for latent variable models with categorical and continuous outcomes*. https://www.statmodel.com/download/Yudissertation.pdf

Zahner, W., Dai, T., Cromley, J. G., Wills, T. W., Booth, J. L., Shipley, T. F., & Stepnowski, W. (2017). Coordinating multiple representations of polynomials: What do patterns in students' solution strategies reveal? *Learning and Instruction*, *49*, 131–141. https://doi.org/10.1016/j.learninstruc.2017.01.007

www.manaraa.com

APPENDIX A

PCA Taxonomy of Foundational Knowledge for Beginning Calculus

Adapted From Carlson et al. (2010)

Reasoning Abilities

- R1 *Process view of function* (items 1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 17, 20, 22, 23)
    - View a function as a generalized process that accepts input and produces output. Appropriate coordination of multiple function processes
- R2 *Covariational reasoning* (items 15, 18, 19, 24, 25)
    - Coordinate two varying quantities that change in tandem while attending to how the quantities change in relation to each other
- R3 *Computational abilities* (items 1, 3, 4, 10, 11, 14, 16, 17, 21)
    - Identify and apply appropriate algebraic manipulations and procedures to support creating and reasoning about function models

Understandings

- Understand meaning of function concepts
    - ME Function evaluation (items 1, 5, 6, 11, 12, 16, 20)
    - MR Rate of change (items 8, 10, 11, 15, 19, 22)
    - MC Function composition (items 4, 5, 12, 16, 17, 20, 23)
    - MI Function inverse (items 2, 4, 9, 10, 13, 14, 23)
- Understand growth rate of function types
    - GL Linear (items 3, 10, 22)
    - GE Exponential (item 7)
    - GR Rational (items 18, 25)
    - GN General non-linear (items 15, 19, 24)
- Understand function representations (interpret, use, construct, connect)
    - RG Graphical (items 2, 5, 6, 8, 9, 10, 15, 19, 24)
    - RA Algebraic (items 1, 4, 7, 10, 11, 14, 16, 17, 18, 21, 22, 23, 25)
    - RN Numerical (items 3, 12, 13)
    - RC Contextual (items 3, 4, 7, 8, 10, 11, 15, 17, 18, 20, 22)

www.manaraa.com

APPENDIX B

Analysis of Articles Citing Carlson et al. (2010)

Table B1

Carlson et al. (2010) Citation Matrix

Citation List	Reference to Theory	Instrument Reference	Type of Test	PCA Data	Psychometric
(Aguilar et al., 2017; Ayalon et al., 2016; Bannerjee, 2017; Breen et al., 2015; Cho & Nagle, 2017; Dawkins & Epperson, 2014; de Beer, 2011, 2016; de Beer et al., 2018; Engelke et al., 2018; Ferguson, 2012; Flynn et al., 2015; Fowler, 2014; Hitt & González-Martín, 2015; Huang et al., 2012; Koştur & Yılmaz, 2017; LaRue, 2017; Leshota, 2015; Mielicki & Wiley, 2016; Nagle et al., 2013; Nagle et al., 2017; Nagle & Moore-Russo, 2014; Nagle, 2013; Özdil, 2012; Phifer, 2014; Rostorfer, 2014; Savic et al., 2017; Sevim & Cifarelli, 2014; Sutton, 2015; Tallman, 2015; Tang, 2012; R. Thomas, 2015; Thompson, 2013; Yemen-Karpuzcu et al., 2017)	1				
(Marfai, 2016; McCrory et al., 2012; Mejia-Ramos et al., 2017; Musgrave & Carlson, 2017; Speer & Kung, 2018; Thompson et al., 2013)		1			
(Bagley et al., 2015; Bagley et al., 2016; Froyd et al., 2012; Giovanniello, 2017; Haider et al., 2016; Stanhope et al., 2017; M. Thomas, 2013; M. Thomas & Lozano, 2012; Thompson, 2014)			1		
(Karakok et al., 2013)				1	
(Horvath, 2012; K. Moore, 2013; Perez, 2013; Watson, 2015)	1	1			
(Bain & Towns, 2016; Byerley, 2016; Gleason et al., 2015; Melhuish, 2015; Thompson, 2015)		1	1		
(Avila, 2013)	1		1		
(Cousino, 2013; Cromley et al., 2017; Kassaee, 2016; Kim, 2017; K. C. Moore et al., 2014; Thompson & Carlson, 2017; Weber et al., 2015)		1		1	
(Byerley, 2016)	1			1	
(O'Shea et al., 2016)	1	1	1		
(Doerr et al., 2014; Drlik, 2015; Meylani & Teuscher, 2011a, 2011b; D. Miller et al., 2015; Palha & Koopman, 2016; Silverman, 2017; Teuscher & Reys, 2012; R. V. Thomas, 2016; Vrabel, 2014; Wills et al., 2014)	1	1		1	
(Zahner et al., 2017)		1	1	1	1
(Williams, 2017)	1	1	1	1	

www.manaraa.com

APPENDIX B REFERENCES

Aguilar, M. S., Castañeda, A., & González-Polo, R. I. (2017). Research findings associated with the concept of function and their implementation in the design of mathematics textbooks tasks. *CERME*, *10*, 3776–3783.

Avila, C. (2013). *Secondary and postsecondary calculus instructors' expectations of student knowledge of functions: A multiple-case study* (Publication No. CFE0004809)[Doctoral dissertation, University of Central Florida]. University of Central Florida Showcase of Text, Archives, Research, and Scholarship. http://purl.fcla.edu/fcla/etd/CFE0004809

Ayalon, M., Watson, A., & Lerman, S. (2016). Reasoning about variables in 11 to 18 year olds: Informal, schooled and formal expression in learning about functions. *Mathematics Education Research Journal*, *28*(3), 379–404. https://doi.org/10.1007/s13394-016-0171-5

Bagley, S., Gleason, J., Rice, L., Thomas, M., & White, D. (2016, July). *Does the Calculus Concept Inventory really measure conceptual understanding of calculus?* American Mathematical Society. https://blogs.ams.org/matheducation/2016/07/25/does-the-calculus-concept-inventory-really-measure-conceptual-understanding-of-calculus/

Bagley, S., Rasmussen, C., & Zandieh, M. (2015). Inverse, composition, and identity: The case of function and linear transformation. *The Journal of Mathematical Behavior*, *37*, 36–47. https://doi.org/10.1016/j.jmathb.2014.11.003

Bain, K., & Towns, M. H. (2016). A review of research on the teaching and learning of chemical kinetics. *Chemistry Education Research and Practice*, *17*(2), 246–262. https://doi.org/10.1039/C5RP00176E

www.manaraa.com

Bannerjee, P. (2017). Students' understanding of the concepts of rates of change and functions. In T. A. Olson & L. Venenciano (Eds.), Proceedings of the 44th Annual Meeting of the Research Council on Mathematics Learning, Fort Worth, TX.

Breen, S., Larson, N., O'Shea, A., & Pettersson, K. (2015). Students' concept images of inverse functions. In *Proceedings of the Ninth Congress of the European Society for Research in Mathematics Education* (CERME 9; pp. 2228–2234). Charles University in Prague, Faculty of Education.

Byerley, C. (2016). *Secondary teachers' and calculus students' meanings for fraction, measure and rate of change* [Doctoral dissertation, Arizona State University]. ASU Electronic Theses and Dissertations. https://repository.asu.edu/attachments/175144/content/Byerley_asu_0010E_16389.pdf

Cho, P., & Nagle, C. (2017). Procedural and conceptual difficulties with slope: An analysis of students' mistakes on routine tasks. *International Journal of Research in Education and Science*, *3*(1), 135–150.

Cousino, A. (2013). *Using bayesian learning to classify college algebra students by understanding in real-time* [Doctoral dissertation, Kansas State University]. K-State Research Exchange. http://hdl.handle.net/2097/15630

Cromley, J. G., Booth, J. L., Wills, T. W., Chang, B. L., Tran, N., Madeja, M., Shipley, T. F., & Zahner, W. (2017). Relation of spatial skills to calculus proficiency: A brief report. *Mathematical Thinking and Learning*, *19*(1), 55–68. https://doi.org/10.1080/10986065.2017.1258614

Dawkins, P. C., & Epperson, J. A. M. (2014). The development and nature of problem-solving among first-semester calculus students. *International Journal of Mathematical Education*

www.manaraa.com

in Science and Technology, *45*(6), 839–862. https://doi.org/10.1080/0020739X.2014.884645

de Beer, H. (2011). *Research proposal*. https://www.heerdebeer.org/DR/publications/ICO_research_plan.pdf

de Beer, H. (2016). *Exploring instantaneous speed in grade five: A design research* [Doctoral dissertation, Eindhoven University of Technology]. ESoE. https://research.tue.nl/nl/publications/exploring-instantaneous-speed-in-grade-five-a-design-research

de Beer, H., van Eijck, M., & Gravemeijer, K. (2018). *Design principles for teaching primary calculus*. https://heerdebeer.org/DR/publications/de_Beer_paper_Hamburg2011.pdf

Doerr, H. M., Ärlebäck, J. B., & Costello Staniec, A. (2014). Design and effectiveness of modeling-based mathematics in a summer bridge program. *Journal of Engineering Education*, *103*(1), 92–114. https://doi.org/10.1002/jee.20037

Drlik, D. I. (2015). *Student understanding of function and success in calculus* [Master's thesis, Boise State University]. ScholarWorks. https://scholarworks.boisestate.edu/td/893

Engelke, N., Karakok, G., & Wangberg, A. (2018). *An annotation tool designed to interface with webwork: Interpreting students'written work*. http://pzacad.pitzer.edu/~dbachman/RUME_XVI_Linked_Schedule/rume16_submission_134.pdf

Ferguson, L. (2012). Two parallel calculuses: The one taught and the one used. In *Proceedings of the 15th Annual Conference On Research in Undergraduate Mathematics Education*, 45. http://sigmaa.maa.org/rume/crume2012/RUME_Home/For_Authors_files/RUME 2012 Schedule_Author.docx

www.manaraa.com

Flynn, C. D., Davidson, C. I., Dotger, S., & Sullivan, M. (2015). Development and pilot test of the rate and accumulation concept inventory.. *2015 122nd ASEE Annual Conference and Exposition*. https://experts.syr.edu/en/publications/development-and-pilot-test-of-the-rate-and-accumulation-concept-i-2

Fowler, B. (2014). *An investigation of the teaching and learning of function inverse* [Master's thesis, Arizona State University]. ASU Electronic Theses and Dissertations. https://repository.asu.edu/attachments/135155/content/Fowler_asu_0010N_13982.pdf

Froyd, J. E., Hurtado, D., Lagoudas, M. Z., Nite, S., Hobson, M., Hodge, J., & Monroe, J. (2012). Increasing access to engineering. In *Frontiers in Education Conference (FIE) proceedings, 2012* (pp. 1–6). https://doi.org/10.1109/FIE.2012.6462265

Giovanniello, S. (2017). *What algebra do calculus students need to know?* (Publication No. 982929) [Master's thesis, Concordia University]. Spectrum Research Repository. https://spectrum.library.concordia.ca/982929/

Gleason, J., White, D., Thomas, M., Bagley, S., & Rice, L. (2015). The calculus concept inventory: A psychometric analysis and framework for a new instrument. In *Proceedings of the 18 Th Annual Conference on Research in Undergraduate Mathematics Education*, (pp. 135–149). https://www.researchgate.net/publication/318489141_The_calculus_concept_inventory_A_psychometric_analysis_and_framework_for_a_new_instrument

Haider, M., Bouhjar, K., Findley, K., Quea, R., Keegan, B., & Andrews-Larson, C. (2016). Using student reasoning to inform assessment development in linear algebra. T. Fukawa-Connelly, N. Infante, M. Wawro, & S. Brown (Eds.), *Proceedings of the19th Annual Conference on Research in Undergraduate Mathematics Education* (pp. 163–177).

www.manaraa.com

Hitt, F., & González-Martín, A. S. (2015). Covariation between variables in a modelling process: The ACODESA (collaborative learning, scientific debate and self-reflection) method. *Educational Studies in Mathematics*, *88*(2), 201–219. https://doi.org/10.1007/s10649-014-9578-7

Horvath, A. K. (2012). *The treatment of composition in secondary and early collegiate mathematics curricula* [Doctoral dissertation, Michigan State University]. MSU Libraries Digital Repository. https://d.lib.msu.edu/etd/1132

Huang, X., Li, S., & An, S. (2012). Understanding of teaching strategies on quadratic functions in chinese mathematics classrooms. *Research in Mathematical Education*, *16*(3), 177–194. https://doi.org/10.7468/jksmed.2012.16.3.177

Karakok, G., Engelke, N., & Wangberg, A. (2013). WeBWorK CLASS: Fostering design experiment research on concept development. *Lighthouse Delta 2013: The 9th Delta Conference on Teaching and Learning of Undergraduate Mathematics and Statistics*. https://www.semanticscholar.org/paper/WeBWorK-CLASS-%3A-Fostering-design-experiment-on-Karakok-Engelke/5f0376f79cfdb2f6ca0d1467522738c2dce15f43

Kassaee, A. M. (2016). *Examining the role of motivation and mindset in the performance of college students majoring in STEM fields* [Doctoral dissertation, Middle Tennessee State University]. JEWLScholar@MTSU. http://jewlscholar.mtsu.edu/handle/mtsu/4991

Kim, A. (2017). Solutions to selected problems from Mathematical Gazette. *The Journal of Undergraduate Research in Natural Sciences and Mathematics*, *19*, 141. https://www.fullerton.edu/nsm/_resources/pdfs/dimensions/Dimensions2017_web.pdf

Koştur, M., & Yılmaz, A. (2017). Technology support for learning exponential and logarithmic functions. *Ihlara Eğitim Araştırmaları Dergisi*, *2*(2), 50–68.

www.manaraa.com

LaRue, R. (2017). *An analysis of student approaches to solving optimization problems in first semester calculus* [Doctoral dissertation, Eberly College of Arts and Sciences]. The Research Repository @ WVU. https://doi.org/10.33915/etd.6040

Leshota, M. (2015). *The relationship between textbook affordances and mathematics' teachers' pedagogical design capacity (PDC)* [Doctoral dissertation, University of the Witwatersrand]. CORE. http://wiredspace.wits.ac.za/bitstream/handle/10539/18211/PhD%20thesis_Leshota.pdf?sequence=2&isAllowed=y

Marfai, F. S. (2016). *Characterizing teacher change through the perturbation of pedagogical goals* [Doctoral dissertation, Arizona State University]. ASU Electronic Theses and Dissertations. https://repository.asu.edu/attachments/170440/content/Marfai_asu_0010E_15940.pdf

McCrory, R., Floden, R., Ferrini-Mundy, J., Reckase, M. D., & Senk, S. L. (2012). Knowledge of algebra for teaching: A framework of knowledge and practices. *Journal for Research in Mathematics Education*, *43*(5), 584–615. https://doi.org/10.5951/jresematheduc.43.5.0584

Mejia-Ramos, J. P., Lew, K., de la Torre, J., & Weber, K. (2017). Developing and validating proof comprehension tests in undergraduate mathematics. *Research in Mathematics Education*, *19*(2, SI), 130–146. https://doi.org/10.1080/14794802.2017.1325776

Melhuish, K. M. (2015). The design and validation of a group theory concept inventory [Doctoral dissertation, Portland State University]. PDXScholar. https://doi.org/10.15760/etd.2487

www.manaraa.com

Meylani, R., & Teuscher, D. (2011a). Precalculus concept assessment: A predictor of AP calculus AB and BC scores. In L. R. Wiest & T. Lamberg (Eds.), *Proceedings of the Thirty-Third Annual Conference of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 787–794).

Meylani, R., & Teuscher, D. (2011b). Calculus readiness: Comparing student outcomes from traditional precalculus and AP calculus AB with a novel precalculus program. In L. R. Wiest & T. Lamberg (Eds.), *Proceedings of the Thirty-Third Annual Conference of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 778–786).

Mielicki, M. K., & Wiley, J. (2016). College students' understanding of linear functions: Slope is slippery. *Cognitive Science*, 1811–1816.

Miller, D., Infante, N., & Adu, S. (2015). *Students' understanding of composition of functions using model analysis* [Paper presentation]. 18th Annual Conference on Research in Undergraduate Mathematics Education, Pittsburgh, PA. https://www.researchgate.net/publication/285421281_Students'_Understanding_Of_Composition_Of_Functions_Using_Model_Analysis

Moore, K. (2013). Making sense by measuring arcs: A teaching experiment in angle measure. *Educational Studies in Mathematics*, *83*(2), 225–245. https://doi.org/10.1007/s10649-012-9450-6

Moore, K. C., Paoletti, T., & Musgrave, S. (2014). Complexities in students' construction of the polar coordinate system. *Journal of Mathematical Behavior*, *36*, 135–149. https://doi.org/10.1016/j.jmathb.2014.10.001

www.manaraa.com

Musgrave, S., & Carlson, M. P. (2017). Understanding and advancing graduate teaching assistants' mathematical knowledge for teaching. *Journal of Mathematical Behavior*, *45*, 137–149. https://doi.org/10.1016/j.jmathb.2016.12.011

Nagle, C., Casey, S., & Moore-Russo, D. (2017). Slope and line of best fit: A transfer of knowledge case study. *School Science and Mathematics*, *117*(1–2), 13–26. https://doi.org/10.1111/ssm.12203

Nagle, C., & Moore-Russo, D. (2014). Slope across the curriculum: Principles and standards for school mathematics and common core state standards. *The Mathematics Educator*, *23*(2), 40-59.

Nagle, C., Moore-Russo, D., Viglietti, J., & Martin, K. (2013). Calculus students' and instructors' conceptualizations of slope: A comparison across academic levels. *International Journal of Science and Mathematics Education*, *11*(6), 1491–1515. https://doi.org/10.1007/s10763-013-9411-2

Nagle, C. R. (2013). *The development of prerequisite notions for an introductory conception of a functional limit* (Publication No. 3516433) [Doctoral dissertation, University of New York]. ProQuest Dissertations and Theses. https://search.proquest.com/openview/5d5f2f2112c587149786dd1dbcde0250/1?pq-origsite=gscholar&cbl=18750&diss=y

O'Shea, A., Breen, S., & Jaworski, B. (2016). The development of a function concept inventory. *International Journal of Research in Undergraduate Mathematics Education*, *2*(3), 279–296. https://doi.org/10.1007/s40753-016-0030-5

www.manaraa.com

Özdil, U. (2012). *A multilevel structural model of mathematical thinking in derivative concept* [Doctoral dissertation, Middle East Technical University]. OpenMETU. https://hdl.handle.net/11511/21252

Palha, S., & Koopman, S. (2016). Interactive virtual math: A tool to support self-construction graphs by dynamical relations. *Proceedings of the 10th Congress of European Research on Mathematics Education (CERME 10).* https://hal.archives-ouvertes.fr/hal-01946333

Perez, B. (2013). *Teacher quality and teaching quality of 7th-grade algebra I honors teachers* [Doctoral dissertation, Florida Atlantic University]. Florida Atlantic University Digital Library. http://purl.flvc.org/fcla/dt/3360970

Phifer, C. R. (2014). *The cycle intersection matrix and applications to planar graphs and data analysis for postsecondary mathematics education* [Doctoral dissertation, University of Rhode Island]. Open Access Dissertations. http://digitalcommons.uri.edu/oa_diss/210

Rostorfer, R. L. (2014). *A cluster analysis of precalculus student performance on function translation fluency test items* [Master's thesis, Middle Tennessee State University]. JEWLScholar@MTSU. http://jewlscholar.mtsu.edu/handle/mtsu/4336

Savic, M., Karakok, G., Tang, G., El Turkey, H., & Naccarato, E. (2017). Formative assessment of creativity in undergraduate mathematics: Using a creativity-in-progress rubric (CPR) on proving. In R. Leikin & B. Sriraman (Eds.), *Creativity and Giftedness* (pp. 23–46). Springer.

Sevim, V., & Cifarelli, V. V. (2014). Authors' response: Radical constructivist conceptual analyses in mathematical problem solving and their implications for teaching. *Constructivist Foundations*, *9*(3, SI), 386–392.

www.manaraa.com

Silverman, J. (2017). Supporting teachers' understandings of function through online professional development. *Journal of Computers in Mathematics and Science Teaching*, *36*(1), 17–39.

Speer, N., & Kung, D. (2018). *Research on teaching and how it relates to graduate student professional development*. https://services.math.duke.edu/~bookman/TAconfSpeerKung.pdf

Stanhope, L., Ziegler, L., Haque, T., Le, L., Vinces, M., Davis, G. K., Zieffler, A., Brodfuehrer, P., Preest, M., & Belitsky, J. M. (2017). Development of a biological science quantitative reasoning exam (BioSQuaRE). *CBE-Life Sciences Education*, *16*(4), ar66. https://doi.org/10.1187/cbe.16-10-0301

Sutton, J. M. S. (2015). *The influence of dynamic visualization on undergraduate calculus learning* [Doctoral dissertation, The University of Texas at Arlington]. UTA Libraries. http://hdl.handle.net/10106/25346

Tallman, M. A. (2015). *An examination of the effect of a secondary teacher's image of instructional constraints on his enacted subject matter knowledge* [Doctoral dissertation, Arizona State University]. ASU Electronic Theses and Dissertations. https://repository.asu.edu/attachments/157960/content/Tallman_asu_0010E_15189.pdf

Tang, G. (2012, Februrary 23-25). *Student thinking of function composition and its impact on their ability to set up the difference quotients of the derivative* [Paper presentation]. Proceedings of the 15th Annual Conference on Research in Undergraduate Mathematics Education, Portland, OR. http://sigmaa.maa.org/rume/crume2012/RUME_Home/RUME_Conference_Papers_files/RUME_XV_Conference_Papers.pdf

www.manaraa.com

Teuscher, D., & Reys, R. E. (2012). Rate of change: AP calculus students' understandings and misconceptions after completing different curricular paths. *School Science and Mathematics*, *112*(6), 359–376. https://doi.org/10.1111/j.1949-8594.2012.00150.x

Thomas, M. (2013). *Analyzing conceptual gains in introductory calculus with interactively-engaged teaching styles* [Doctoral dissertation, University of Arizona]. UA Theses and Dissertations. http://hdl.handle.net/10150/299075

Thomas, M., & Lozano, G. (2012). *Analyzing calculus concept inventory gains in introductory calculus*. Research in Undergraduate Mathematics Education. http://pzacad.pitzer.edu/~dbachman/rume_xvi_linked_schedule/rume16_submission_95.pdf

Thomas, R. (2015, March 12-15). *A graphing approach to algebra using Desmos* [Paper presentation]. Proceedings of the 27th International Conference on Technology in Collegiate Mathematics, Las Vegas, NV. http://archives.math.utk.edu/ICTCM/VOL27/A026/paper.pdf

Thomas, R. V. (2016). *The effects of dynamic graphing utilities on student attitudes and conceptual understanding in college algebra* [Doctoral dissertation, University of Arkansas, Fayetteville]. ScholarWorks@UARK. https://scholarworks.uark.edu/etd/1569

Thompson, P. W. (2013). In the absence of meaning…. In K. Leatham (Ed.), *Vital directions for mathematics education research* (pp. 57–93). Springer. https://doi.org/10.1007/978-1-4614-6977-3_4

Thompson, P. W. (2014). Constructivism in mathematics education. In S. Lerman (Ed.), *Encyclopedia of mathematics education* (pp. 96–102). Springer.

www.manaraa.com

Thompson, P. W. (2015). Researching mathematical meanings for teaching. In L. English & D. Kirshner (Eds.), *Handbook of international research in mathematics education* (3rd ed., pp. 435–461). Taylor and Francis.

Thompson, P. W., Byerley, C., & Hatfield, N. (2013). A conceptual approach to calculus made possible by technology. *Computers in the Schools*, *30*(1–2), 124–147. https://doi.org/10.1080/07380569.2013.768941

Thompson, P. W., & Carlson, M. P. (2017). Variation, covariation, and functions: Foundational ways of thinking mathematically. In J. Cai (Ed.), *Compendium for research in mathematics education* (pp. 421–456). National Council of Teachers of Mathematics.

Vrabel, A. R. (2014). *Function conceptions of AP calculus students* [Doctoral dissertation, University of Pittsburgh]. D-Scholarship@Pit*t*. http://d-scholarship.pitt.edu/id/eprint/22820

Watson, K. L. (2015). *Examining the effects of college algebra on students' mathematical dispositions* [Master's thesis, Brigham Young University]. BYU ScholarsArchive. https://scholarsarchive.byu.edu/etd/5601

Weber, E., Tallman, M. A., & Middleton, J. A. (2015). Developing elementary teachers' knowledge about functions and rate of change through modeling. *Mathematical Thinking and Learning*, *17*(1), 1–33. https://doi.org/10.1080/10986065.2015.981940

Williams, D. A. (2017). *Student experiences in community college precalculus: A mixed methods study of student engagement and understanding* [Doctoral dissertation, North Carolina State University]. NC State Repository. http://www.lib.ncsu.edu/resolver/1840.20/33658

www.manaraa.com

Wills, T., Shipley, T., Chang, B., Cromley, J., & Booth, J. (2014). What gaze data reveal about coordinating multiple mathematical representations. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *36*. https://escholarship.org/uc/item/8hw5s599

Yemen-Karpuzcu, S., Ulusoy, F., & Işıksal-Bostan, M. (2017). Prospective middle school mathematics teachers' covariational reasoning for interpreting dynamic events during peer interactions. *International Journal of Science and Mathematics Education*, *15*(1), 89–108. https://doi.org/10.1007/s10763-015-9668-8

Zahner, W., Dai, T., Cromley, J. G., Wills, T. W., Booth, J. L., Shipley, T. F., & Stepnowski, W. (2017). Coordinating multiple representations of polynomials: What do patterns in students' solution strategies reveal? *Learning and Instruction*, *49*, 131–141. https://doi.org/10.1016/j.learninstruc.2017.01.007

www.manaraa.com