

Diagnostic Classification Modeling of Rubric-Scored Constructed-
Response Items

Eric William Muller

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2018

© 2018
Eric William Muller
All rights reserved

ABSTRACT

Diagnostic Classification Modeling of Rubric-Scored Constructed-Response Items

Eric William Muller

The need for formative assessments has led to the development of a psychometric framework known as diagnostic classification models (DCMs), which are mathematical measurement models designed to estimate the possession or mastery of a designated set of skills or attributes within a chosen construct. Furthermore, much research has gone into the practice of “retrofitting” diagnostic measurement models to existing assessments in order to improve their diagnostic capability. Although retrofitting DCMs to existing assessments can theoretically improve diagnostic potential, it is also prone to challenges including identifying multidimensional traits from largely unidimensional assessments, a lack of assessments that are suitable for the DCM framework, and statistical quality, specifically highly correlated attributes and poor model fit. Another recent trend in assessment has been a move towards creating more authentic constructed-response assessments. For such assessments, rubric-based scoring is often seen as method of providing reliable scoring and interpretive formative feedback. However, rubric-scored tests are limited in their diagnostic potential in that they are usually used to assign unidimensional numeric scores.

It is the purpose of this thesis to propose general methods for retrofitting DCMs to rubric-scored assessments. Two methods will be proposed and compared: (1) automatic construction of an attribute hierarchy to represent all possible numeric score levels from a

rubric-scored assessment and (2) using rubric criterion score level descriptions to imply an attribute hierarchy. This dissertation will describe these methods, discuss the technical and mathematical issues that arise in using them, and apply and compare both methods to a prominent rubric-scored test of critical thinking skills, the Collegiate Learning Assessment+ (CLA+). Finally, the utility of the proposed methods will be compared to a reasonable alternative methodology: the use of polytomous IRT models, including the Graded Response Model (GRM), the Partial Credit Model (PCM), and the Generalized-Partial Credit Model (G-PCM), for this type of test score data.

Table of Contents

LIST OF TABLES	iii
LIST OF FIGURES	v
ACKNOWLEDGMENTS	vi
1 INTRODUCTION	1
1.1 THE NEED FOR MORE FORMATIVE ASSESSMENT	2
1.2 IMPROVING THE DIAGNOSTIC ABILITY OF ASSESSMENTS	4
1.2.1 RETROFITTING DCMs TO EXISTING ASSESSMENTS	6
2 LITERATURE REVIEW	11
2.1 ATTRIBUTE SPECIFICATION	11
2.2 THE DCM FRAMEWORK AND THE STATISTICAL NATURE OF DCMs	14
2.2.1 COMPENSATORY VS. NON-COMPENSATORY	16
2.2.2 CONJUNCTIVE VS. DISJUNCTIVE	17
2.2.3 GUESS VS. SLIP PARAMETERS	17
2.2.4 THEORETICAL FRAMEWORKS FOR DCA DESIGN	18
2.3 COMMON DIAGNOSTIC CLASSIFICATION MODELS	22
2.3.1 THE DINA AND DINO MODELS	23
2.3.2 THE G-DINA MODEL	26
2.3.3 HIERARCHICAL DIAGNOSTIC CLASSIFICATION	28
2.4 ASSESSING MODEL FIT	33
2.5 RUBRICS AS DIAGNOSTIC GUIDES	36
2.5.1 HIERARCHICAL ATTRIBUTE STRUCTURES IN DCMs	39
2.5.2 ALTERNATIVE MODELS FOR RUBRIC-SCORED TESTS	46
2.6 ITEM RESPONSE THEORY MODELS FOR POLYTOMOUS DATA	47
2.6.1 THE GRADED RESPONSE MODEL	49
2.6.2 THE PARTIAL CREDIT MODEL	53
2.6.3 THE GENERALIZED PARTIAL CREDIT MODEL	56
2.7 SUMMARY OF THE LITERATURE	57
2.8 GOALS OF THE PRESENT STUDY / THEORETICAL CONTRIBUTIONS	59
3 METHOD	61
3.1 OVERVIEW	61
3.2 AN APPLICATION – THE COLLEGIATE LEARNING ASSESSMENT +	63
3.2.1 THE RUBRIC-SCORED CONSTRUCTED-RESPONSE ITEM SECTION	64
3.2.2 PARTICIPANTS	66
3.2.3 SAMPLE A DEMOGRAPHICS	67
3.2.4 SAMPLE B DEMOGRAPHICS	67
3.3 RETROFITTING THE DINA MODEL TO CONSTRUCTED-RESPONSE ITEMS	68
3.3.1 POLYTOMOUS SCORE TO DICHOTOMOUS VECTOR CONVERSION	68
3.3.2 RUBRIC TO Q-MATRIX CONVERSION (RUBRIC CODING)	71
3.3.3 AN AUTOMATIC M-ATTRIBUTE METHOD (FULL-SCORE CODING)	82
3.4 SOFTWARE FOR FITTING THE PROPOSED MODELS	84

3.5 DIAGNOSTIC CLASSIFICATION MODEL ANALYSIS	84
3.5.1 CONDITIONS FOR DINA MODEL ANALYSIS	85
3.5.2 MODEL FIT STATISTICS	88
3.5.3 ITEM FIT/PARAMETER ESTIMATES	90
3.5.4 ATTRIBUTE TETRACHORIC CORRELATION	91
3.5.5 INFORMATION-BASED ITEM DISCRIMINATION INDICES FOR DCMs	92
3.6 GENERALIZED PARTIAL CREDIT MODEL (G-PCM) ANALYSIS	95
4 EMPIRICAL RESULTS/APPLICATION TO THE CLA+	96
4.1 DESCRIPTIVE STATISTICS	96
4.2 CONSTRUCTED-RESPONSE SECTION RETROFIT RESULTS	99
4.2.1 MODEL FIT	99
4.2.2 ITEM PARAMETER ESTIMATES	103
4.2.3 COGNITIVE DIAGNOSTIC INDICES (CDI)	107
4.2.4 TETRACHORIC CORRELATIONS	110
4.2.5 SKILL CLASSIFICATION ESTIMATES	111
4.3 G-PCM ANALYSIS: CONSTRUCTED-RESPONSE SECTION	115
4.3.1 G-PCM ANALYSIS	115
4.3.2 ITEM AND TEST INFORMATION	122
4.3.3 COMPARING G-PCM AND PCM MODEL FIT	125
4.3.4 G-PCM EXAMINEE ABILITY ESTIMATES	126
5 DISCUSSION AND CONCLUSIONS	127
5.1 DISCUSSION OF THE ANALYSES	127
5.1.1 MODEL FIT RESULTS	128
5.1.2 ITEM FIT RESULTS	132
5.1.3 COGNITIVE DIAGNOSTIC INDICES	136
5.1.3 SKILL CLASSIFICATION ESTIMATES	137
5.2 CONCLUSIONS	139
5.3 LIMITATIONS AND FUTURE RESEARCH	148
REFERENCES	150
APPENDIX I. THE CLA+ SCORING RUBRIC	175
APPENDIX II. SAMPLE CLA+ ASSESSMENT	176

List of Tables

TABLE 1. EXAMPLE OF A Q-MATRIX	13
TABLE 2. EXAMPLE OF A Q-MATRIX AS A PERFECT GUTTMAN SCALE	42
TABLE 3. EXAMPLE OF AN INDEPENDENT APPROACH Q-MATRIX	43
TABLE 4. EXAMPLE OF AN ADJACENT APPROACH Q-MATRIX	44
TABLE 5. EXAMPLE OF A REACHABLE APPROACH Q-MATRIX	45
TABLE 6. BETWEEN CATEGORY THRESHOLD PARAMETERS FOR FIGURE 2	51
TABLE 7. BETWEEN CATEGORY THRESHOLD PARAMETERS FOR FIGURE 3	55
TABLE 8. TABLE FOR CONVERTING RUBRIC SUB-AREA SCORES INTO PSEUDO-ITEM RESPONSE PATTERNS	70
TABLE 9. ANALYSIS AND PROBLEM SOLVING (APS) SUB-AREA ATTRIBUTE SPECIFICATIONS	73
TABLE 10. Q-MATRIX DESIGN FOR ANALYSIS AND PROBLEM SOLVING (APS) SUB-AREA	75
TABLE 11. WRITING EFFECTIVENESS (WE) SUB-AREA ATTRIBUTE SPECIFICATIONS	75
TABLE 12. Q-MATRIX DESIGN FOR WRITING EFFECTIVENESS (WE) SUB-AREA	77
TABLE 13. WRITING MECHANICS (WM) SUB-AREA ATTRIBUTE SPECIFICATIONS	78
TABLE 14. Q-MATRIX DESIGN FOR WRITING MECHANICS (WM) SUB-AREA	79
TABLE 15. EXAMPLE OF Q-MATRIX DESIGN FOR ALL SUB-AREAS COMBINED	80
TABLE 16. EXAMPLE OF A SUB-AREA Q-MATRIX FOR A RUBRIC WITH SEVEN LEVELS, DESIGNED USING THE AUTOMATIC M-METHOD	83
TABLE 17. CLA+ RUBRIC SUB-AREA Q-MATRIX DESIGNED USING THE AUTOMATIC M-METHOD	83
TABLE 18. CONDITIONS IN DINA MODEL ANALYSIS	86
TABLE 19. TEST FORM A SUB-AREA SCORE PROPORTIONS	97
TABLE 20. TEST FORM B SUB-AREA SCORE PROPORTIONS	97
TABLE 21. CRONBACH'S ALPHA FOR OVERALL TEST RELIABILITY	97
TABLE 22. FORM A: ITEM (SUB-AREA) CORRELATIONS	98
TABLE 23. FORM B: ITEM (SUB-AREA) CORRELATIONS	98
TABLE 24. FORM A RUBRIC CODED Q-MATRIX DESIGN MODEL FIT STATISTICS	100
TABLE 25. FORM B RUBRIC CODED Q-MATRIX DESIGN MODEL FIT STATISTICS	101
TABLE 26. FORM A FULL-SCORE CODED Q-MATRIX DESIGN MODEL FIT STATISTICS	102
TABLE 27. FORM B FULL-SCORE CODED Q-MATRIX DESIGN MODEL FIT STATISTICS	102
TABLE 28. FORM A RUBRIC-CODED Q-MATRIX DESIGN ITEM PARAMETER ESTIMATES WITHOUT GUESS OR SLIP PARAMETER CONSTRAINTS	104
TABLE 29. FORM B RUBRIC CODED Q-MATRIX DESIGN ITEM PARAMETER ESTIMATES WITHOUT GUESS OR SLIP PARAMETER CONSTRAINTS	104

TABLE 30. FORM A FULL-SCORE CODED Q-MATRIX DESIGN ITEM PARAMETER ESTIMATES WITHOUT GUESS OR SLIP PARAMETER CONSTRAINTS	106
TABLE 31. FORM B FULL-SCORE CODED Q-MATRIX DESIGN ITEM PARAMETER ESTIMATES WITHOUT GUESS OR SLIP PARAMETER CONSTRAINTS	106
TABLE 32. ITEM ATTRIBUTE DISCRIMINATION VALUES FOR RUBRIC CODED COMBINED SUB-AREA Q-MATRIX DESIGNS WITH NO PARAMETER CONSTRAINTS	110
TABLE 33. FORM A RUBRIC CODED COMBINED SUB-AREA Q-MATRIX DESIGN ATTRIBUTE TETRACHORIC CORRELATIONS	111
TABLE 34. FORM B RUBRIC CODED COMBINED SUB-AREA Q-MATRIX DESIGN ATTRIBUTE TETRACHORIC CORRELATIONS	111
TABLE 35. FORM A UNCONSTRAINED DINA MODEL RUBRIC CODED COMBINED SUB-AREA SAMPLE ATTRIBUTE MASTERY PROBABILITY PROFILES	114
TABLE 36. FORM A: ESTIMATES OF COEFFICIENTS FROM FITTING A G-PCM	116
TABLE 37. FORM A: LATENT TRAIT ABILITY DISTANCES BETWEEN G-PCM DIFFICULTY STEP PARAMETERS	116
TABLE 38. FORM B: ESTIMATES OF COEFFICIENTS FROM FITTING A G-PCM	117
TABLE 39. FORM B: LATENT TRAIT ABILITY DISTANCES BETWEEN G-PCM DIFFICULTY STEP PARAMETERS	118
TABLE 40. ITEM (SUB-AREA) INFORMATION SHARES	124
TABLE 41. FORM A: POLYTOMOUS-IRT MODEL FIT STATISTICS AND ANOVA RESULTS	125
TABLE 42. FORM B: POLYTOMOUS-IRT MODEL FIT STATISTICS AND ANOVA RESULTS	125
TABLE 43. EXAMPLE OF A REVISED Q-MATRIX FOR ANALYSIS AND PROBLEM SOLVING (APS)	143

List of Figures

FIGURE 1. FOUR FORMS OF HIERARCHICAL ATTRIBUTE STRUCTURES.	40
FIGURE 2. EXAMPLE OF BOUNDARY CHARACTERISTIC CURVES.	51
FIGURE 3. EXAMPLE OF ITEM CATEGORY RESPONSE CURVES.	55
FIGURE 4. CLA+ SCORING RUBRIC SUB-AREA DEFINITIONS.	71
FIGURE 5. ANALYSIS AND PROBLEM SOLVING (APS) SUB-AREA ATTRIBUTE DEFINITIONS.	72
FIGURE 6. WRITING EFFECTIVENESS (WE) SUB-AREA ATTRIBUTE DEFINITIONS.	75
FIGURE 7. WRITING MECHANICS (WM) SUB-AREA DEFINITIONS.	78
FIGURE 8. FORM A TEST DIAGNOSTIC INFORMATION	108
FIGURE 9. FORM B TEST DIAGNOSTIC INFORMATION	109
FIGURE 10. FORM A RUBRIC CODED SINGLE SUB-AREA LATENT CLASS PROFILE POPULATION MEMBERSHIP PROBABILITIES	112
FIGURE 11. FORM B RUBRIC CODED SINGLE SUB-AREA LATENT CLASS PROFILE POPULATION MEMBERSHIP PROBABILITIES	113
FIGURE 12. FORM A: ITEM (SUB-AREA) CATEGORY RESPONSE CURVES FOR ANALYSIS AND PROBLEM SOLVING (APS).	119
FIGURE 13. FORM A: ITEM (SUB-AREA) CATEGORY RESPONSE CURVES FOR WRITING EFFECTIVENESS (WE).	119
FIGURE 14. FORM A: ITEM (SUB-AREA) CATEGORY RESPONSE CURVES FOR WRITING MECHANICS (WM).	120
FIGURE 15. FORM B: ITEM (SUB-AREA) CATEGORY RESPONSE CURVES FOR ANALYSIS AND PROBLEM SOLVING (APS).	121
FIGURE 16. FORM B: ITEM (SUB-AREA) CATEGORY RESPONSE CURVES FOR WRITING EFFECTIVENESS (WE).	121
FIGURE 17. FORM B: ITEM (SUB-AREA) CATEGORY RESPONSE CURVES FOR WRITING MECHANICS (WM).	122
FIGURE 18. FORM A: ITEM (SUB-AREA) INFORMATION CURVES.	123
FIGURE 19. FORM B: ITEM (SUB-AREA) INFORMATION CURVES.	124
FIGURE 20. FORM A DISTRIBUTION OF ABILITY ESTIMATES.	126

Acknowledgments

While this dissertation represents the culmination of my own doctoral work, it could not have been accomplished without the aid of many great sources. First, I would like to thank my advisor, Dr. James Corter, for his constant support, guidance, expertise, and mentorship in my academic endeavors, for which I am truly grateful. Dr. Corter's counsel and direction throughout my academic career were critical in both reaching this pinnacle of achievement and its completion.

I would next like to thank the other members of my defense committee, Dr. Charles Lang, Dr. Young-Sun Lee, Dr. Stephen Peverly, and Dr. Doris Zahner for their time, commitment, and suggestions towards the revision and completion of this thesis. Each of you provided me with feedback that was integral to producing a doctoral thesis of which I can take great pride in calling my own. I would especially like to thank Dr. Zahner and the Council for Aid to Education for providing me with the data used in this study, as well as Dr. Zahner's own support, encouragement, and mentorship during my internship at the Council for Aid to Education.

Lastly, I would like to thank my friends and family. To my loyal friends, thank you for always being there to make things easier when times were hard. To my family, especially my parents, Junie and Eugene, thank you for your unwavering support, guidance, and encouragement, all of my accomplishments in life are, largely in part, a result of your love.

1 Introduction

The need for more formative assessment, assessment that aids and facilitates student learning by pinpointing specific strengths and weaknesses (Garrison & Ehringhaus, 2007), has led to the development of a psychometric framework known as diagnostic classification models (DCMs). DCMs are mathematical measurement models designed to estimate the possession or “mastery” of a designated set of skills or attributes within a chosen construct (Rupp, Templin, & Henson, 2010). For DCMs to be most effective, they must be applied to assessments that are designed underneath the DCM framework, which are known as Diagnostic Classification Assessments (DCAs). However, due to the significant amount of time and resources required to design, produce, and analyze a DCA, developments in DCMs far outpace developments in DCAs (de la Torre & Minchen, 2014; Henson, 2009).

In an attempt overcome this dilemma, much research has gone into the practice of retrofitting DCMs to existing assessments in order to improve their diagnostic capability and gain meaning insights and knowledge. However, although retrofitting DCMs to existing assessments has promise in terms of diagnostic potential, it is also prone to challenges including assumptions including identifying multidimensional traits from unidimensional assessments, a lack of assessments that are suitable for the DCM framework, and statistical quality, specifically highly correlated attributes and poor model fit (Huff & Goodman, 2007; Liu, Huggins-Manley, & Bulut, 2017). Similarly, the recent trend in assessments towards more formative assessments has paralleled a move to

more authentic constructed-response assessments. In order to do this, rubric-based scoring is often seen as method of providing reliable diagnostic and formative feedback (Banerjee, Yan, Chapman, & Elliot, 2015). However, often rubric-scored tests are limited in that they are only able to produce unidimensional results since they are designed to measure aptitude or proficiency of a unidimensional construct.

To serve the objective of retrofitting DCMs to traditional assessments in order to produce more diagnostic feedback, this study will propose general methods for retrofitting DCMs to rubric-scored assessments. Two methods will be proposed and compared: (1) an automatic construction of an attribute hierarchy to represent possible numeric score levels from a rubric-scored assessment and (2) using rubric criterion descriptions as a proposed content-based attribute hierarchy. This dissertation will describe these methods in detail, discuss the technical and mathematical issues that arise in using them, and apply and compare both methods to a prominent rubric-scored test of critical thinking skills, the Collegiate Learning Assessment+ (CLA+). Finally, the utility of the proposed methods will be compared to a reasonable alternative methodology: the use of polytomous IRT models, the Graded Response Model (GRM), the Partial Credit Model (PCM), and the Generalized-Partial Credit Model (G-PCM), for rubric-scored constructed-response data.

1.1 The Need for more Formative Assessment

One of the greatest challenges that faces educators is to develop a method of educational assessment that both gauges and facilitates student learning (Chappuis & Chappuis, 2008;

de la Torre & Minchen, 2014; Rupp, Templin, & Henson, 2010). Often a test score only serves to inform either the student or a second party as to how well (measured by a single numeric index) that student performs on an exam in comparison to their peers or to a designated standard. While this kind of test result may be informative to some capacity, it may also fail to address specific areas where a student exceeds or falls below average expectations, thus failing to provide the student with feedback they could use to improve their performance. These kinds of assessments that provide these unidimensional results are classified as summative assessments.

Summative assessments evaluate the degree to which an individual meets a set criterion (Chappuis & Chappuis, 2008). For the majority of high-stakes testing, and testing within the educational system, assessments are largely summative (Linn, 2000). As a result, there is currently a large amount of effort within the education community to create assessments that can aide learning while being able to evaluate its progress as well by identifying respondent's strengths and weaknesses (Lee & Corter, 2011; Leighton & Gierl, 2007b; Leighton, Gokiart, Cor, & Heffernan, 2010; Nichols, 1994; Tatsuoka, Corter, & Tatsuoka, 2004). The assessments that provide this kind of feedback are known as formative assessments.

Formative assessments are assessments that improve and promote student learning by providing information to the student that is both constructive for learning and meaningful to the student by informing them of their strengths and weaknesses within a particular domain (Garrison & Ehringhaus, 2007). Any assessment can be used for both summative and/or formative purposes, although some are better suited for one than the other.

For example, consider an assessment that evaluates a constructed-response for a given prompt. If the assessment were to be used formatively, an instructor could point out specific areas where the student performed well and needed improvement, thus informing the student of their specific strengths and weaknesses relative to the task at hand. If the assessment were to be used for summative purposes, a grade would be assigned that merely informed the student how well they performed according to that teacher's criterion. However, although formative assessments are more beneficial for student learning than merely summative assessments, formative assessments are difficult to administer to larger populations since they require more time and resources to design, develop, and score (de la Torre & Minchen, 2014). It is primarily for these reasons that assessments such as the SAT, which was administered to over one and a half million students in 2015 and 2014 (Klein, 2015), cannot be scored with as much attention to detail as assessments designed for smaller class-room sized populations. As a result, the desire for educators to shift the focus of assessment from summative to formative purposes has generated an impetus to develop methods of formative assessment that can be administered to large-scale populations (Bennett, 2015; Huff & Goodman, 2007; Organization for Economic Co-operation and Development [OECD], 2004; U.S. Department of Education, 2004).

1.2 Improving the Diagnostic Ability of Assessments

As a response to the demand for formative large-scale assessment, diagnostic measurement has emerged as a potential solution. Diagnostic measurement is a

framework for developing assessments that can provide students with rich diagnostic, multidimensional, classification-based feedback in order to facilitate learning development (Dibello & Stout, 2007; Liu, Huggins-Manley, & Bulut, 2017; Rupp & Templin, 2008b). The diagnostic measurement framework is able to provide this kind of feedback via the application of diagnostic measurement models, known as diagnostic classification models (DCMs), to test data. DCMs are a group of probabilistic, confirmatory, multidimensional latent class models that attempt to determine the presence or absence of mastery in a set of designated skills or test “attributes” for a single or group of examinee(s) (Rupp, Templin, & Henson, 2010). Ideally, assessments are designed with the diagnostic measurement framework in mind so that the application of DCMs to test data can provide the kind of formative feedback that is desired from this approach. In contrast, assessments that are not designed within the DCM framework are not designed in a way that produces data suitable for DCMs, thus requiring that the DCMS be retrofit to that data.

These kinds of assessments that are designed within the DCM framework and measure the skills/attributes deemed necessary for proficiency within a determined domain are referred to as diagnostic classification assessments (DCAs) (de la Torre & Minchen, 2014). However, despite the apparent utility of using DCAs to create more diagnostic and formative tools of measurement, the development of DCMs has far outpaced the development of DCAs (de la Torre & Minchen, 2014; Henson, 2009; Lee, de la Torre, & Park, 2012; Lee, Park, & Taylan, 2011). The major reason for the lack of DCA development is that developing DCAs requires collaboration between experts from multiple fields such as subject matter experts, psychometricians, and educators, as well as

significant amounts of time and resources (de la Torre & Minchen, 2014; Huff & Goodman, 2007).

1.2.1 Retrofitting DCMs to Existing Assessments

As a potential solution for this complication, retrofitting DCMs to existing assessments has emerged as a possible method for obtaining diagnostic feedback from assessments that were not designed within the DCM framework (Chen & Chen, 2016; Jang, Dunlop, Wagner, Kim, & Gu, 2013; Jurich & Bradshaw, 2014; Kim, 2014; Li, Hunter, & Lei, 2015). The definition of “retrofitting” used in this study will be the same used by Liu, Huggins-Manley, & Bulut (2017) which is “the practice of fitting DCMs to responses obtained from assessments that are not designed under diagnostic measurement frameworks that typically fall under Classical Test Theory (CTT) or Item Response Theory (IRT). Diagnostic measurement frameworks essentially itemize categorical traits in order to classify each trait as being either present or not present (i.e. a skill that is mastered or not mastered) by an examinee. In contrast, traditional measurement frameworks delineate one or more continuous traits and essentially places examinees on those latent trait continuum or continua (Liu, Huggins-Manley, & Bulut, 2017).

Take, for example, an assessment that measures reading ability. In a non-diagnostic measurement framework, reading ability could be measured using a single latent trait (e.g. reading ability) or multiple continuous traits (e.g. spelling, grammar, and conceptual understanding). In either case, respondents would receive a single score for each latent trait, thus placing the respondent on a location on each respective trait’s latent

trait continuum. In a diagnostic measurement framework, reading ability would be depicted as multiple categorical traits (e.g. spelling, grammar, and conceptual understanding) and respondents would be classified as either masters or non-masters of each latent trait, based on the estimated probability of mastery for each latent trait.

Using the same example, but in a case wherein the diagnostic measurement framework was retrofitted to an existing assessment (not designed using diagnostic measurement framework) that only provided a single score for each latent trait, the process of retrofitting would begin by determining what subskills/attributes are necessary to master in order to be proficient in each latent trait. As a result, the end user would be provided with diagnostic information on what skills each respondent had either mastered or not mastered, within each latent trait, thus providing the end user with specific feedback on respondent's potential strengths and weaknesses. In this way, diagnostic measurement has been shown to be an effective framework for providing formative feedback for test takers, administrators, and developers (Rupp, Templin, & Henson, 2010) and retrofitting is considered a possible method for extracting such information (Liu, Huggins-Manley, & Bulut, 2017).

However, retrofitting DCMs to assessments is not without its challenges. Liu, Huggins-Manley, & Bulut (2017) identify several challenges that can arise as a result of either assessment design or statistical quality. In terms of assessment design, the authors call attention to three issues. The first is that diagnostic measurement is only effective if the assessment is designed with underlying cognitive or educational theories as its guiding principles. Otherwise, it is difficult to support any theoretical or empirical claims made regarding attribute specifications (Nichols, Kobrin, Lai, & Koepfler, 2016).

The second issue is that there is a conflict of goals between the dimensionality assumptions of DCMs and unidimensional assessments (which includes most assessments not designed using the diagnostic measurement framework) (de la Torre & Karelitz, 2009; Gierl & Cui, 2008). Often a major objective of CTT- and IRT-based assessment development is to support the assumption that an assessment measures a single construct and only one construct (de Ayala, 2009). For example, in an assessment that is designed to only measure skills in investment banking, it would be problematic to discover that an item were measuring a separate construct such as financial accounting, or that the assessment on the whole measures skills both in investment banking and in financial accounting. In such a case the construct validity of the exam would be compromised and revisions would be made to either the problematic item or the assessment as a whole in order to ensure unidimensionality. Therefore, retrofitting DCM's objective to institute multidimensionality into a unidimensional assessment can be complicated and not straightforward. In some cases, subject matters experts may even be necessary to determine what potential subskills and attributes are necessary for demonstrating proficiency in a unidimensional assessment's measured construct (Lee, Park, & Taylan, 2011).

The third issue regarding assessment design is that in some cases attributes are not specified in a way that produces reliable results or makes them identifiable for a given test (Chen, Liu, Xu, & Yang, 2015; Tatsuoka, 1995; Templin & Bradshaw, 2014). For example, reliable results would not be attained if only a single item measures an attribute. In that case the only evidence for skill mastery would be an examinee's correct or incorrect response to that item. An example of a case where attributes are not identifiable

would be if two attributes were always paired together across items. In such a case there would be insufficient evidence to determine whether or not a single attribute has been mastered, since all items that measure them require both attributes. Consequently, it is recommended that each attribute should be measured both as singletons and paired with multiple attributes per item (Liu, Huggins-Manley, Bradshaw, 2016).

Regarding statistical quality, Liu, Huggins-Manley, & Bulut (2017) single out three issues. The first of these issues is that attributes obtained from unidimensional assessments tend to be highly correlated (VanderVeen et al., 2007). Highly correlated attributes can be problematic since it contradicts the multidimensional assumption of the diagnostic measurement framework that each attribute is its own separate construct. The second issue regarding statistical quality that may arise is that fitting DCMs to unidimensional assessment data can result in poor model fit (Gierl & Cui, 2008). As with most psychometric models, poor model fit typically is considered evidence that the mathematical model is not appropriate to be applied to a particular set of data and in the case of DCM retrofitting, wherein the objective is usually to find evidence that DCMs are not only appropriate for the data but can provide more desirable results than unidimensional models, this can also be problematic for establishing result validity and reliability (Chen, de la Torre, & Zhang, 2013; Hu, Miller, Huggins-Manley, & Chen, 2016; Lei & Li, 2016). Finally, the third issue of statistical quality is that correlations between attributes and the total score of an exam may be weak or even negative (Buck & Tatsuoka, 1998; Svetina, Gorin, & Tatsuoka, 2011). Weak or negative correlations between assessment attributes and total score are an issue since such findings would contradict the diagnostic measurement framework assumption that mastery of the

multidimensional attributes are necessary for demonstrating proficiency in the overall construct.

Therefore it is the aim of this dissertation to develop techniques for using DCMs to analyze rubric scored assessments to improve the diagnostic ability of such assessments. Much of the current research that involves retrofitting DCMs has focused on producing diagnostic information from assessments and items that were not intended for diagnostic purposes. However, little research has gone into formalizing and improving the diagnostic ability of rubric-scored constructed-response assessment, which by their design specify attributes and associations of attribute mastery with levels of skill and attribute proficiency.

2 LITERATURE REVIEW

In this chapter a review of the literature on attribute specification, the diagnostic classification measurement framework, the statistical nature of diagnostic classification models, common diagnostic classification models, assessing model fit, rubrics as diagnostic guides, and item response theory models for polytomous data are presented in six sections. This is followed by a summary of that literature and a general discussion as to how these concepts relate to the issue of diagnostic classification modeling rubric-scored data. The final section of the chapter then provides the exact goals of the present study, including the intended theoretical contributions.

2.1 Attribute Specification

In order for an assessment to be able to fit to the diagnostic framework it is essential to delineate the skills or “attributes” required for proficiency in the assessment’s overall construct. Each item in the assessment is assumed to require either a single or multiple attributes in order to answer that item correctly. When developing DCAs from the ground up, these attributes are typically delineated prior to test construction. Furthermore, each attribute is represented as a dichotomous ordinal variable that represents either mastery or non-mastery.

Liu, Huggins-Manley, and Bulut (2017) identify three considerations that should be heeded when conducting attribute specification from items that were not designed in

the diagnostic measurement framework. The first is that items that measure single attributes are more likely to have higher classification accuracy than items that measure multiple attributes (Madison & Bradshaw, 2015). Therefore, in cases where two attributes are consistently required together, it may be beneficial to combine them into a single attribute. The second consideration is that relationships between attributes can be correlated or even dependent upon each other (Templin & Bradshaw, 2014; Liu & Huggins-Manley, 2016). As discussed earlier, highly correlated attributes may indicate evidence of unidimensionality. Attributes that are dependent upon each other require that one attribute be mastered before another can be mastered, a condition that is also known as an attribute hierarchy (Leighton, Gierl, & Hunka, 2004; Liu & Huggins-Manley, 2016).

The third and final consideration that should be taken into account is that it is recommended by multiple researchers that it is ideal to try to minimize the total number of attributes. As the number of attributes increases, so does the difficulty in parameter estimation and also in interpretation of results (Embretson & Yang, 2013; Xu & Zhang, 2016). As a rule of thumb it is recommended that an assessment measure a maximum of 10 attributes (DiBello, Roussos, & Stout, 2007). According to research done by Liu, Huggins-Manley, & Bulut (2017), most retrofitting studies they encountered specified three to five attributes.

Once the skills/attributes that are necessary for mastering the general construct assessed by the DCA have been defined and determined, each item must then be constructed so that responses demonstrate skill mastery or non-mastery. The Q-matrix is a matrix that encodes which skills/attributes are required to be mastered by each item in

order to answer that item correctly (Tatsuoka, 1983). While the Q-matrix functionally acts as a component in the computation of skill mastery probabilities, it also can act as a roadmap of a DCA's design. For example, in Table 1 we observe how a list of six theoretical attributes can be encoded for a DCA with six items:

Table 1. Example of a Q-matrix

Item	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Attribute 6
1	1	0	1	0	0	0
2	1	0	0	0	0	0
3	0	1	0	1	0	1
4	0	1	0	0	0	0
5	0	0	0	1	0	0
6	0	0	0	1	1	1

In this example, answering each item correctly represents mastery of a different attribute. However, it is possible for an item to demonstrate complete or partial mastery (depending on whether or not the DCM used is noncompensatory or compensatory, respectively) in multiple attributes. Likewise, it is also possible for an attribute to be represented across multiple items.

In essence, attributes are characterized as being either discrete or dichotomous and are therefore either present in or absent in a response (Madison & Bradshaw, 2015). The combination of attributes that a respondent ultimately possesses is defined as an attribute pattern, \mathbf{a} , and is represented as a latent vector of length K . In the attribute pattern, \mathbf{a} , a 0 represents non-mastery of the k th attribute and a 1 represents mastery of the k th attribute. Additionally, each item has a corresponding q-vector, \mathbf{q} , which is also of length K , and indicates which skills/attributes are necessary to solve each particular item (Tatsuoka, 1995). For example, the corresponding q-vector for Item 3 in Table

1 is $\{0,1,0,1,0,1\}$. Also observable in Table 1 is the fact that a Q-matrix will always be of length J , where J is the number of items on the assessment, and of width K , where K is the number of attributes measured by the assessment. Furthermore, due to the dichotomy of the presence of skill/attribute mastery, there will also always be 2^K possible attribute combinations. For example, Table 1 has 64 possible combinations of attribute patterns. Therefore, the complexity of a DCM's computation increases exponentially as the number of measured attributes increases.

Finally, it is essential that when designing the Q-matrix that each attribute is measured by an adequate number of items, regardless of it be represented as a singleton or conjoined with multiple attributes (Madison & Bradshaw, 2015). As was stated earlier, items that measure single attributes tend to have higher classification accuracy. In fact, many researchers concur that in order for a Q-matrix to be complete, each attribute that exists in a Q-matrix there must be at least one item that only measures that attribute (Casella & Berger, 2002; Chiu, Douglas, & Li, 2009; DeCarlo, 2011; DiBello, Stout, & Roussos, 1995; Tatsuoka, 1991; Xu & Zhang, 2016). Furthermore, it is critical that the Q-matrix be specified correctly since a misspecification of the Q-matrix can lead to detrimental effects to model fit as well as producing flawed inferences regarding the classification accuracy of latent traits (Chen, Liu, Xu, & Ying, 2015; Chiu, 2013; Kunina-Habenicht, Rupp, & Wilhelm, 2012; Madison & Bradshaw, 2015; Rupp & Templin, 2008a, 2008b).

2.2 The DCM Framework and the Statistical Nature of DCMs

Diagnostic classification models (DCMs) are statistical models capable of making diagnostic analyses (Nichols, Chipman, & Brennan, 2009). In statistical terms, DCMs can be defined as confirmatory multidimensional latent-variable models with categorical latent variables. DCMs are the frameworks upon which diagnostic classification assessments, (DCAs) assessments that provide diagnostic information, are built (DiBello, Roussos, & Stout, 2007). In order for an assessment to be cognitively diagnostic, it is required that the assessment be designed to measure knowledge components or cognitive skills that are deemed necessary for proficiency in a determined domain. An assessment designed as a DCA must also assume that its design is consistent with up-to-date developments in cognitive theories and other relevant fields in order to maintain its validity and accuracy (de la Torre & Minchen, 2014). In other words, if an assessment is designed within the framework of a DCM and provides diagnostic information, it can be classified as a DCA. It is also the objective of all DCMs to classify respondents according to selected latent variables, which ultimately leads to the creation of attribute profiles for DCA respondents and the production of diagnostic information. As a result, DCMs have also been referred to as latent response models, restricted latent class models, multiple classification latent class models, structured located latent class models, and structured item response theory models although they all refer to the same construct (Rupp, Templin, & Henson, 2010).

Currently there are a wide variety of DCMs that have been developed, each one assuming a distinct cognitive model that specifies that exact nature of the relationship between examinees' attributes and their ability to answer an item on a DCA correctly. Rupp, Templin, and Henson (2010) identify 18 different core DCMs although this paper

focuses on four that most sharply contrast each other in order to highlight the salient characteristics that separate DCMs from each other categorically. The aforementioned researchers also identified three classifications from which to categorize DCMs: manifest response variable type (dichotomous or polytomous), latent predictor variable type (dichotomous or polytomous), and model type (noncompensatory and compensatory) (Rupp & Templin, 2008b). DCMs can be members of any number classifications, for example, the *log-linear cognitive diagnosis model* (LCDM) can be categorized according to any combination of the three classifications. A model such as the *deterministic inputs, noisy “and” gate* (DINA) model can only be applied to data with dichotomous observed response and latent predictor variables wherein the nature of the relationship between the attributes and the items is noncompensatory.

2.2.1 *Compensatory vs. Non-Compensatory*

There are two types of DCMs: *general* and *specific*. The distinctions between these two types of DCMs are the assumptions that the models make regarding the compensatory nature of the fine-grained skills and attributes measured by the exam. DCMs can be either *compensatory* or *non-compensatory*. If a DCM is compensatory, it assumes that the lack of mastery in one skill or attribute measured by the exam can be made up for by the presence of mastery in another skill or attribute measured by the exam. Contrarily, if a DCM is non-compensatory, it assumes that the lack of mastery in one skill or attribute measured by the exam cannot be made up for by the presence of another skill or attribute measured by the exam. DCMs that can only be either compensatory or non-compensatory

within a single exam are ‘specific’ DCMs. DCMs that allow for both compensatory and noncompensatory relationships within a single exam are ‘general’ DCMs (Ravand, 2016). General DCMs operate by allowing each item to pick the model that best fits the data as opposed to imposing one single model for all items.

2.2.2 *Conjunctive vs. Disjunctive*

Additionally, DCMs are also categorized as being either *conjunctive* or *disjunctive*. A conjunctive DCM assumes that, in order to answer an item correctly, all of the fine-grained skills and attributes measured by the exam must be mastered by the examinee. Contrarily, a disjunctive DCM assumes that not all of the fine-grained skills and attributes measured by the exam are required to be mastered in order to answer an item correctly. Therefore, due to the nature of the compensatory and conjunctive assumptions, DCMs can be conjunctive and non-compensatory, disjunctive and compensatory, or disjunctive and non-compensatory, but not conjunctive and compensatory.

2.2.3 *Guess vs. Slip Parameters*

Another distinction that is made between different DCMs is the use of the *guess* and *slip* parameters. The *guess* and *slip* parameters operate similarly to the guessing parameter used in the 3-parameter IRT model. In the DCM framework, a *guess* represents a case in which a respondent has insufficient attribute mastery to answer an item correctly but does so anyway. Similarly, a *slip* represents a case in which a respondent has sufficient

mastery to answer an item correctly but answers incorrectly (Junker & Sijtsma, 2001). The probability that either of these events occur is estimated by DCMs and become the guessing and slipping parameters. The difference in these variables between different DCMs is at what point in the model structure they are taken into account and how they are to be applied. In essence, the guessing and slipping parameters can be applied at the item level across all attributes, at the attribute level across all items, or at both the item and attribute levels (de la Torre, 2011; de la Torre & Minchen, 2014). Once the skills/attributes that are necessary for mastering the general construct assessed by the DCA have been defined and determined, each item must then be constructed so that responses demonstrate skill mastery or non-mastery.

2.2.4 Theoretical Frameworks for DCA Design

With regard to DCAs, Rupp, Templin, & Henson (2010) identify two different frameworks for the design and development of a DCA. The first framework is the cognitive design system (CDS), and it primarily follows current research on cognitive mechanics as they relate to assessments that measure basic cognitive abilities such as spatial rotation and/or general reasoning. The goal of an assessment designed with a CDS framework is to be able to delineate attributes that pertain to the skills measured by the assessment so accurately and in such a way that they can be stored while the test is not in use and applied to the evaluation of performance on the assessment instantaneously. Seven steps must be followed in order to design a DCA within the principled assessment design process of the CDS framework.

The first two steps of the CDS framework are to determine the construct that the assessment is to measure and identify necessary/relevant attributes for accurate measurement of the construct. The next three steps of principled assessment design according to the CDS framework is to develop a cognitive model for performance on assessment tasks, generate items according to the developed cognitive model, and evaluate the cognitive model empirically by evaluating each item's individual performance (Embretson, 1994, 1998). The final two steps of the CDS framework is to bank the items according to cognitive complexity and validate the model by ensuring that the model is indeed only measuring the specified construct.

The first two initial steps already demonstrate the potential complexity that can go into the design and development of a DCA. Aside from identifying the general construct that is to be measured, identifying the skills/attributes that are necessary for measuring that construct, depending on the nature of the construct, may require collaboration of multiple subject matter experts of various fields (de la Torre & Minchen, 2014). For example, if the designated construct to be measured on an assessment were to be a subject that is commonly taught in public education such as biology, there would most likely need to be a great deal of research that would need to be done in order to establish which attributes/skills were most relevant features of that domain as well as agreement from multiple subject matter experts as biology encompasses a vast array of different subjects, of which all can be considered to have equal importance. Furthermore, once these steps have been completed, even more experts from other fields may still be needed to contribute to the design of the DCA.

In the next three steps we can see that this will require experts from both a test construction design standpoint, as well as an expert on relevant cognitive theories. Depending on the context of the administration of the exam (virtual or physical domain), other experts and resources may be required that have the potential limit the scope of what the tasks are able to assess. However, once these five initial steps are complete the rest of the DCA design is relatively straightforward.

At this point in the design of the assessment, which are the final two steps of the CDS framework, no further experts are required. However, the determination of the validity of the assessment may require parallel research and so it is not necessarily guaranteed even at the final stage of the assessment design that the assessment is valid with 100% certainty. That being said, the CDS framework highlights the importance of designing an assessment in conjunction with underlying theory in order to maximize the defensibility of the assessment's validity. Also, while a CDS framework of design for DCAs may require a number of different experts and resources, this depends on the scope of the construct that the DCA is designed to measure (Embretson, 1994, 1998). The more specific the construct being measured by the assessment, the less time/resource consuming and number of experts are necessary for its construction.

Rupp, Templin and Henson (2010) also discuss the evidence-centered design (ECD) framework. The main crux of the ECD framework is that items should be constructed in a way that they elicit specific behaviors that are most indicative of an individual's latent ability structure (Mislevy, Steinberg, & Almond, 2003; Mislevy, Steinberg, & Lukas, 2006). An ECD framework consists of five different models that operate partially hierarchically and partially in conjunction with each other, unlike the

CDS framework that operates procedurally following step after step. At the bottom level of the ECD framework is the respondent model and the task model. The respondent model specifies which attributes are relevant to be measured by the assessment in order to demonstrate skill mastery (Mislevy, Steinberg, & Lukas, 2006). The task model specifies how those specified skills are to be measured by the assessment. These models, like the first few stages of the CDS framework, may require the collaboration of multiple experts in order to accomplish.

The evidence model links both the respondent model and task model together by determining which parts of a response are most salient and what statistical models will be used to make assertions on the performance of the respondent model (Mislevy, Steinberg, & Lukas, 2006). In other words, the evidence model determines what aspects of the responses given, to items that were designed according to the task model on constructs specified by the respondent model, are most noteworthy for indicating skill mastery and what metrics are used to justify the validity of the items. This means that the evidence model operates as a mediator between the respondent model and the task model, with each model influencing the other through the evidence model. The final two models of the ECD framework encapsulate and are a product of the three aforementioned models. These are the assembly model and presentation model.

The assembly model determines what combination of the evidence model, respondent model, and task model will be used for a particular assessment or subsection of an assessment (Rupp, Templin, & Henson, 2010). This means that, in an ECD framework, a test can measure multiple constructs that comprise an overarching construct. In other words, instead of a single result to indicate overall construct mastery,

multiple results may be given from multiple subsections. Finally, the presentation model specifies whether task modes change for different subsections and, if they do change, what are the implications of these changes (Rupp, Templin, & Henson, 2010). For example, if the mode of one subsection is multiple-choice and another requires constructed responses this can have implications for changes with regards to the task and evidence model.

Ultimately the ECD framework for DCA design highlights the importance of the statistical models used in the evidence model. These statistical models become the lens through which patterns of behavior from respondents are perceived and interpreted and yet the decision of which statistical model to use is at the complete discretion of the DCA developer. In other words, DCA developers must make informed and careful decisions on which attributes should be represented via variables in the statistical models they choose (Rupp, Templin, & Henson, 2010). Although there theoretically is not one correct choice for most DCMs, choosing the most appropriate statistical model is possible and can depend on a number of different factors including the construct being measured, cognitive theory, pedagogical theory, and learning theory.

2.3 Common Diagnostic Classification Models

This section provides a description of the purpose, design, and statistical properties of commonly cited and used DCMs. Models that are presented in pairs are typically noncompensatory and compensatory or disjunctive and conjunctive analogs of each other, meaning that they both hold very similar mathematical structures, but for some

critical difference in composition are used to represent different kinds of attribute relationships. Other DCMs presented include families of DCMs that may not always fit the typical mold of most DCMs and may only be used for unique attribute or data situations, such as attribute hierarchies.

2.3.1 The DINA and DINO Models

The *deterministic inputs, noisy “and” gate* (DINA) model and the *deterministic inputs, noisy “or” gate* (DINO) model are noncompensatory and compensatory analogs of each other, respectively. In the case of the DINA model, it is conjunctive and noncompensatory (Heartel, 1989; Junker & Sijtsma, 2001; de la Torre & Douglas, 2004). In other words, the model assumes that examinees must possess all required skills in order to answer an item efficiently and the presence of one skill cannot compensate for the lack of presence of another skill. As a result, the DINA model separates participants into two latent groups per item: those that have mastered the required attributes of the item and those that have not mastered the required attributes of the item. Therefore, respondents that lack any of the required attributes in order to answer a particular item correctly all have an equal probability of answering the item incorrectly (de la Torre & Minchen, 2014).

For respondent i answering item j , the probability of an examinee having the most ideal attribute pattern for that respondent to answer the item correctly, is defined by the DINA model as

$$\eta_{ij} = \prod_{k=1}^K \alpha_{ij}^{q_{jk}}$$

(de la Torre, 2009). In essence, $n=1$ if and only if examinee i has mastered all the required attributes for item j . The slipping parameter s and guessing parameter g for the DINA model are given by

$$s_j = P(X_{ij} = 0 | \eta_{ij} = 1)$$

$$g_j = P(X_{ij} = 1 | \eta_{ij} = 0)$$

Therefore, the probability function for an item's response is given by

$$P(X_{ij} = 1 | \eta_{ij}) = (1 - s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}}$$

Thus, in the DINA model, the probability of respondent i answering item j correctly, given that they have mastered all the required attributes, is equal to the probability that they have mastered all the required attributes and did not answer incorrectly times the probability that they have mastered all the required attributes and did not guess correctly.

Contrarily, the DINO model (Templin & Henson, 2006) is conjunctive and compensatory. This means that, similar to the DINA model, examinees are separated according to those that either have all of or none of the necessary attributes in order to answer an item correctly. However, this also means that, contrary to the DINA model, the DINO model allows for the possibility that the lack of presence in a required skill can be completely compensated for by the presence of just one other required skill.

The probability of an examinee having the most ideal attribute pattern for that respondent to answer the item correctly, according to the DINO model is given as:

$$v_{ij} = \prod_{k=1}^K (1 - \alpha_{ik})^{q_{jk}}$$

In essence, the presence of just one of the designated attributes is required for a respondent to have a high probability of having a correct response to a particular item

(Templin & Henson, 2006). The item response function for the DINO model is essentially identical to that of the DINA model, and is given as such:

$$P(X_{ij} = 1|v_{ij}) = (1 - s_j)^{v_{ij}} g_j^{1-v_{ij}}$$

The difference in interpretation for the DINO model's response function is that, in this case, the probability that an examinee answers an item correctly, assuming that they did not slip or guess is also given the probability that they have mastered *at least* one of the designated attributes.

If a Q-matrix is designed in a way so that all items are singletons, there will be no difference between fitting the DINA model and fitting the DINO model, its disjunctive counterpart. Assigning all items as singletons essentially removes the effect that a difference in “condensation” rule would have on the determination of whether respondents in latent class c have mastered all measured attributes required for item i ($\xi_{ic} = 1$) or not ($\xi_{ic} = 0$). The conjunctive kernel that creates the latent response variable ξ_{ic} can be expressed mathematical as:

$$\xi_{ic} = \prod_{a=1}^A \alpha_{ca}^{q_{ia}}$$

If an attribute is not measured by an item, then $q_{ia} = 0$, which implies that $\alpha_{ca} = 1$, indicating that it does not matter whether or not that attribute is mastered by the respondent in order to answer the item correctly. If an attribute is measured by them item, then $q_{ia} = 1$, implying that it matters whether a respondent in latent class c has mastered the measured attribute $\alpha_{ca} = 1$ or not $\alpha_{ca} = 0$ in order to answer item i correctly. As a result, $\xi_{ic} = 1$ only if all product terms are 1, which means that all measured attributes for item i have been mastered by respondents in latent class c .

The difference between the DINA model and the DINO model is in the calculation of its latent class variable, which in the DINO model is denoted as ω_{ic} . Contrarily to the DINA model, the DINO model uses a disjunctive condensation rule to determine if at least one measured attribute is present (as opposed to all measured attributes). This is expressed mathematically as:

$$\omega_{ic} = 1 - \prod_{a=1}^A (1 - \alpha_{ca})^{q_{ia}}$$

Each parameter in the equation above, functions identically to the parameters in the previous equation. Therefore we can see that in the case wherein each item only measures a single attribute, the disjunctive condensation rule reduces to the conjunctive condensation rule, since if for all other attributes besides being measured $q_{ia} = 0$, there is no effect on ω_{ic} from other attributes.

2.3.2 The G-DINA Model

The *generalized* DINA (G-DINA) model is both disjunctive and compensatory, making it different from both the DINA and DINO models since both of those models are conjunctive. This means that the G-DINA model does not require that all attributes need to be present in order for an item to be answered correctly and assumes that the presence of one attribute can compensate for the lack of presence in another attribute (de la Torre, 2011). Furthermore, the DINA and DINO models separate examinees into two latent groups, those that have either mastered the necessary attributes or have not, whereas the G-DINA model partitions candidates into every possible combination of mastery

presence. In other words, the G-DINA model partitions candidates into $2^{K_j^*}$ groups, where K_j^* is the number of attributes required for item j . This means that the G-DINA model increases exponentially in computational complexity as the number of attributes recorded by the assessment increases. It is also worthy to note at this point that, if $K_j^* = 1$, the DINA and G-DINA model become the same model, since there are only 2 possible groups that can exist.

In order to define the item response function for the G-DINA model, we must first assume that the first K_j^* attributes are required for item j and, subsequently, α_{ij}^* must be defined as the attribute vector which constitutes the first K_j^* elements of α_i . Thus the item response function for the G-DINA model can be defined as follows (de la Torre, 2011):

$$P(\alpha_{ij}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk} \alpha_{ik} + \sum_{k^*=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jk k^*} \alpha_{ik^*} \alpha_{ik} + \dots + \delta_{j12L\dots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{ik}$$

In this equation δ_{j0} represents the intercept, δ_{jk} represents the main effect due to α_k , $\delta_{jk k^*}$ represents the two-way interaction effect between α_{k^*} and α_k , and $\delta_{j12L\dots K_j^*}$ represents the K_j^* -way interaction effect due to α_1 through α_{k^*} (de la Torre, 2011). In terms of interpretation, the intercept represents the minimum probability of success when none of the required attributes are present, the main effect represents the change in the probability of success when one attribute is mastered, and the interaction effects represent the change in the probability of success when more than one attribute is simultaneously mastered.

2.3.3 Hierarchical Diagnostic Classification

The study on Bayesian networks raises the possibility of nested attributes or, in other words, the possibility of attribute hierarchies. A study by Templin & Bradshaw (2014) sought to address this by forging a link between the *Attribute Hierarchy Method* (AHM) (Leighton, Gierl, & Hunka, 2004) and the *Log-linear Cognitive Diagnosis Model* (LCDM) to create the *Hierarchical Diagnostic Classification Model* (HDCM). The LCDM can be either conjunctive and non-compensatory or disjunctive and compensatory, subsuming the DINA, DINO, NIDA, and NIDO models. In an example for an item that measures two attributes, where $q_{j1} = 1$ and $q_{j2} = 1$, conditional on an examinee i 's attribute profile for these two attributes, $\alpha_i = [\alpha_{i1}, \alpha_{i2}]$ the LCDM item response function is as follows (Templin & Bradshaw, 2014):

$$P(X_{ij} = 1 | \alpha_i) = \frac{\exp(\lambda_{j,0} + \lambda_{j,1,(1)} \alpha_{i1} + \lambda_{j,1,(2)} \alpha_{i2} + \lambda_{j,2,(1,2)} \alpha_{i1} \alpha_{i2})}{1 + \exp(\lambda_{j,0} + \lambda_{j,1,(1)} \alpha_{i1} + \lambda_{j,1,(2)} \alpha_{i2} + \lambda_{j,2,(1,2)} \alpha_{i1} \alpha_{i2})}$$

The attribute pattern parameters (α_{i1}, α_{i2}) are equal to either 0 or 1, depending on whether or not the examinee i has mastered the attribute. The item intercept ($\lambda_{j,0}$) is interpreted as the log-odds of a correct response for a respondent that has not mastered any of the required attributes. Subsequently, the main effects ($\lambda_{j,1,(1)}$ and $\lambda_{j,1,(2)}$) represent the change in log-odds for a correct response if an examinee has mastered the respective attribute. The two-way interaction between the two attributes ($\lambda_{j,2,(1,2)} \alpha_{i1} \alpha_{i2}$) then enables the log-odds of a correct response to change in the event that both attributes have been mastered. In terms of the subscripts for the λ parameters, the first represents the corresponding item i , the second indicates the parameter type (in this example it is 0

for intercept, 1 for main effect, and 2 for two-way interaction), and the third indicates which specific attribute the λ parameter pertains to (Templin & Bradshaw, 2014). Thus the LCDM follows a factorial ANOVA model wherein the attributes can be thought of as crossed-factors and it is assumed that all combinations of attributes are possible.

The LCDM can be measured up to A attributes although its computation increases exponentially as the number of attributes increases. Therefore, the general form of the LCDM item response function is as follows (Templin & Bradshaw, 2014):

$$P(X_{ij} = 1 \mid \alpha_i = \alpha_c) = \frac{\exp(\lambda_{j,0} + \lambda_j^T h(\alpha_i, q_j))}{1 + \exp(\lambda_{j,0} + \lambda_j^T h(\alpha_i, q_j))}$$

The term $h(\alpha_i, q_j)$ represents a vector-valued function of size $(2^A - 1) \times 1$, containing the information on if a required attribute has been mastered. Formulaically, the response function of attributes is (Templin & Bradshaw, 2014):

$$\lambda_j^T h(\alpha_i, q_j) = \sum_{a=1}^A \lambda_{j,1,(k)} \alpha_{ik} q_{jk} + \sum_{a=1}^{A-1} \sum_{b>a} \lambda_{j,2,(a,b)} \alpha_{ia} \alpha_{ib} q_{ja} q_{jb}$$

Monotonicity constraints are placed upon the elements of λ_j so that the probability of a positive response increases if an examinee demonstrates mastery on additional required attributes as delineated by the Q-matrix. Therefore the LCDM is a constrained version of a more general latent class model wherein the full model contains the attribute distribution information in base-rate parameters (π_c) that represent the probability of a given respondent from a population has a particular attribute mastery pattern c ($c = 1, \dots, 2^A$). When merged with the item response function, the marginal LCDM likelihood function for binary items with binary attributes for a single respondent then becomes (Templin & Bradshaw, 2014):

$$P(X_i) = \sum_{c=1}^{2A} \pi_c \prod_{j=1}^J P(X_{ij} = 1 | \alpha_i)^{X_{ij}} (1 - P(X_{ij} = 1 | \alpha_i))^{1-X_{ij}}$$

However, so far this has so far only demonstrated half of the components that are necessary for forming the HDCM. The other half of the HDCM formulation requires the AHM, which is a probabilistic approach for classifying respondents and requires a formal representation of the relationships between the measured attributes of an assessment (Templin & Bradshaw, 2014). An AHM analysis begins with the construction of matrix of possible attribute profiles given the structure of the attribute hierarchies (Templin & Bradshaw, 2014). This is similar to how many DCMs require the construction of a Q-matrix. Next, in a step that is similar to the creation of DAG representations when conducting a Bayesian network analysis, the assessment is mapped onto possible attribute profiles. AHM then uses various statistical tests and an index of classification reliability in order to determine if the selected attribute profiles matched the data that is observed, to assess model fit. However, AHM is ultimately more of a pattern-recognition approach to measurement than a statistical measurement model (Templin & Bradshaw, 2014).

In order to bridge the gap between LCDM and AHM, HDCM combines the factorial ANOVA model components of the LCDM and the structural framework of the AHM. More specifically, in congruence with the fundamentals of the LCDM, the HDCM assumes that every possible combination of attributes in the population being measured exists (Templin & Bradshaw, 2014). Simultaneously, keeping in line with the structural framework of AHM, the HDCM considers attributes to be nested factors instead of fully crossed factors, such as in the LCDM. This latter step has the potential to greatly reduce computational complexity, compared to the LCDM, as nesting attribute profiles reduces

the number of possible attribute profiles. As a result, the HDCM changes the parameterization of the LCDM to avoid over-parameterization and reflect the nested structure of the attributes (Templin & Bradshaw, 2014).

For example, if an item measures two attributes a and b , and attribute b is nested within attribute a , the HDCM item response function for an examinee i on item j is as follows (Templin & Bradshaw, 2014):

$$P(X_{ij} = 1 | \alpha_i^*) = \frac{\exp(\lambda_{j,0} + \lambda_{j,1,(a)} \alpha_{ia} + \lambda_{j,2,(b(a))} \alpha_{ia} \alpha_{ib})}{1 + \exp(\lambda_{j,0} + \lambda_{j,1,(a)} \alpha_{ia} + \lambda_{j,2,(b(a))} \alpha_{ia} \alpha_{ib})}$$

The three parameters of the item response function for the HDCM are the intercept ($\lambda_{j,0}$), the main effect for attribute a ($\lambda_{j,1,(a)} \alpha_{ia}$), and the interaction for attribute b nested within attribute a ($\lambda_{j,2,(b(a))} \alpha_{ia} \alpha_{ib}$). Since an attribute hierarchy suggests that certain attribute profiles are impossible, there are no longer 2^A base rate (π_c) parameters such as in the LCDM. However, much like how the nested structure of the Bayesian network had to have been predetermined via a modeling sample, the attribute hierarchy in a HDCM must be determined before diagnostic classification can begin.

Templin & Bradshaw (2014) demonstrate how the LCDM can be used prior to the implementation of the HDCM to discover hierarchies in attributes. In a process similar to conducting a factor analysis, the parameter estimates for λ garnered using the LCDM provide evidence as to what may be “suspected” hierarchies within the attributes. Simply put, by comparing the parameter estimates for the intercepts, main effects, and interaction effects (when possible) for each item across all examinees, it can be determined whether or not there is a linear hierarchy amongst the attributes. Via this method, the researchers determined a suspected attribute hierarchy using a sample of 2,922 examinees that had

taken the Examination of Proficiency in English (ECPE) (Templin & Bradshaw, 2014). Once the suspected hierarchy in the study was codified into the HDCM, a simulation study was conducted in order to assess the HDCM's ability to detect attribute hierarchies, model efficiency, and model fit compared to the LCDM.

The conditions of the simulation study included 3000 examinees, 30 items, and 3 attributes with 500 replications (Templin & Bradshaw, 2014). The results of the study indicated that the HDCM's item parameter estimates were nearly identical to the LCDM's item parameter estimates, with a Pearson correlation 0.999. However, the HDCM displayed more stable behavior, with fewer extreme values (item main effects estimates that were close to their item intercept estimates) and significantly smaller standard errors. The authors then formulated a hypothesis test to determine whether or not an attribute hierarchy existed, by constructing a deviance test statistic, using the naïve distribution, and 100 simulations to obtain the correct p -value. Based on the results of that test, the researchers found no significant difference between the LCDM and HDCM, concluding that an attribute hierarchy was indeed present in the data (Templin & Bradshaw, 2014). However, the HDCM's model fit was then also compared to other unidimensional DCMs (such as the DINA and DINO models) and it was not found to be among the best fitting models.

The authors conclude by recommending that HDCMs be used in conjunction with other DCMs, along with the LCDM, as a litmus test for detecting attribute hierarchies. This top-down approach will enable researchers to become aware of any potential hierarchies that may exist within their attributes before creating Q-matrices and applying diagnostic models to data that are unable to distinguish hierarchies (such as the DINA

and DINO models). In other words, the HDCM is a new psychometric model that can theoretically be used to detect attribute hierarchies and restructure model attributes (Templin & Bradshaw, 2014). Compared to the DCMs previously discussed in this review, this model is not recommended as a stand-alone model for diagnosis, since it is most informative when used with the LCDM.

2.4 Assessing Model Fit

Determining the most appropriate or efficient DCM for a particular set of data is neither a straightforward nor simple process. Multiple models must be applied to the data in order to compare their performance. Measures of model performance include reliability estimates, validity estimates, and model fit statistics. However, unlike other psychometric models, there is a not general consensus on the best model fit statistics and procedures for assessing the fit of DCMs (Sinharay & Almond, 2007). A case study done by Sinharay and Almond (2007) examines a variety of different approaches to assess the model fit of DCMs, using Bayesian networks as an example.

Bayesian networks were selected as an example for that study because they are able to model both complex relationships among proficiency variables and dependencies between observable variables that are measured by the same task (Sinharay & Almond, 2007). However, the study focuses on analysis of the Q-matrix and not the design of the Bayesian network. Therefore, the results from the case study are applicable for other DCMs that utilize binary or discrete latent variables, such as the DINA, NIDA, and other general diagnostic models. The three techniques used by Sinharay and Almond to assess

the model fit of the Bayesian networks were Bayesian residual plots, item fit plots, and an item fit test statistic. The deviance information criterion (DIC), was also suggested as a practical tool for comparing model performance as well, although it applies only to models fitted via an MCMC algorithm (Sinharay & Almond, 2007).

Bayesian residual plots have the ability to demonstrate model fit along a spectrum of skill proficiency by visualizing the posterior mean of the realized residual R_i versus the posterior mean expected number of correct responses $E(Y_i|\theta_i, \pi, \lambda)$ for each examinee i (Sinharay & Almond, 2007). The expected number of correct score of an examinee i is computed by $E(Y_i|\theta_i, \pi, \lambda) = \sum_{j=1}^J \pi_j \delta_{i(s)}$, where $\delta_{i(s)}$ is a binary indicator that denotes whether examinee i has mastered the skills needed for tasks requiring the s th attribute from the Q-matrix. The value of $\delta_{i(s)}$ is determined by θ_i (Sinharay & Almond, 2007). The realized residual $R_i = Y_i - E(Y_i|\omega)$, where $\omega = (\omega_1, \omega_2, \dots, \omega_M)$ represents the vector of model parameters. An MCMC algorithm from a previous study, which the current case study seeks to replicate various aspects of, was used to estimate the values of the parameters and latent class variables. For each examinee, the plot shows how well the model predicts the number of correct responses they received, according to their respective $E(Y_i|\theta_i, \pi, \lambda)$ (Sinharay & Almond, 2007). Depending on the length of the 95% posterior credible interval, the location of R_i demonstrates whether or not the model has over- or under-predicted an examinee's correct number of responses, while each examinee is placed along an axis of posterior mean expected number of correct scores.

With regard to the item fit plots, the same MCMC algorithm used for calculating the Bayesian residual plots is used to generate values of θ_i . For each iteration of the algorithm, p_{kji} is the proportion of the examinees that belong to latent class k and answer

item j correctly, and \widehat{p}_{kj} represents the median p_{kji} over all iterations of the MCMC algorithm (Sinharay & Almond, 2007). Functionally, \widehat{p}_{kj} acts as an observed proportion correct for an item j and latent class k in the item fit plot. Each item fit plot is thus a comparison of \widehat{p}_{kj} versus p_{kji} for each item. For each latent class k , the item fit plot visualizes the proportion of examinees that belong to that latent class k that answered that item correctly. By comparing the Q-matrix, specifically the assigned attributes to each item, it is possible to determine based on the assigned attributes to each item and the mastered attributes of the latent class members whether or not an item is performing efficiently.

The item fit test statistic suggested by the authors is also similar to this method in that it calculates a statistic represent a proportion of correct responses with regard to specific latent classes. Essentially, the test statistic quantifies what is shown in the item fit plots and is denoted as χ^2 (Sinharay & Almond, 2007). For the calculation of this statistic, \widehat{p}_{kj} now acts the observed proportion correct for item j for examinees in latent class k . An observed number of examinees for that latent class, denoted as O_{kj} is then calculated by $O_{kj} = \widehat{p}_{kj}N_k$, where N_k represents the number of members of latent class k . A corresponding predicted number of examinees for the same latent class E_{kj} can be obtained by $E_{kj} = P_{kj}N_k$ where P_{kj} is the predicted proportion correct for item j for examinees in latent class k . The function for the item fit statistic is then as follows (Sinharay & Almond, 2007):

$$\chi_j^2 = \sum_{k=1}^9 \frac{(O_{kj} - E_{kj})^2}{E_{kj}} + \sum_{k=1}^9 \frac{[(N_k - O_{kj}) - (N_k - E_{kj})]^2}{N_k - E_{kj}} = \sum_{k=1}^9 \frac{N_k(O_{kj} - E_{kj})^2}{E_{kj}(N_k - E_{kj})}$$

The authors conclude that, while these approaches and techniques may be useful tools in gauging model fit, they risk being too conservative since their observed values are not actually observed but are estimated from the data (Sinharay & Almond, 2007). Another limitation of these techniques is that they are experimental and require more research across multiple models and datasets in order to determine their reliability and validity. As DCMs are still a burgeoning subject within the field of psychometrics, much research is still being conducted as to what the most reliable measures of model fit are for these statistical models.

2.5 Rubrics as Diagnostic Guides

Rubrics are defined as detailed scoring guides that articulate the expectations for an assignment by delineating the assessment's criteria and describing the levels of quality in relation to each of those criteria (Reddy & Andrade, 2010). Rubrics are unique in that they are inherently diagnostic in that their objective is to identify a respondent's location on a graded scale in terms of proficiency on at least one attribute, and so they are able facilitate formative assessment by their very nature (Panadero & Jonsson, 2013). Scriven (1991) identifies three characteristics of diagnostic inference that highlight the inherent diagnostic ability of rubrics:

1. Diagnosis requires that the inherent features of poor performances be determined and reported.
2. The process of diagnosis should result in the classification of cognitive skills via an appropriate reporting system.

3. The objective of diagnosis is classification.

Similarly, Wolf & Stevens (2007) identify three steps for designing effective rubrics:

1. Clearly identify the specific criteria that are necessary for demonstrating expertise in the construct. Three to six are recommended in order maximize reliability, although increasing the number of criteria increases the formative ability of the assessment.
2. Determine the number of performance levels that are appropriate for the assessment. Again, three to six levels are recommended in order to maximize reliability, although increasing the number of performance levels increases the formative ability of the assessment.
3. Identify subskills within each criterion that provide sufficient guidance to determine a response's level. It is critical that these subskills maintain a parallel structure throughout their respective criteria. Describing the subskills at each level using similar language, form, and content can increase parallelism across subskills.

Clearly, rubrics share significant parallels in their purpose, objective, and even procedure for design with diagnostic assessment. However, just as there is a dearth in DCA development, research in rubric analysis and performance is scarce as well (Hafner & Hafner, 2010). Much of the literature on improving rubric performance employs intuitive or qualitative methodologies in order to make recommendations for design and development (Hawkey & Barker, 2004; Janssen, Meier, & Trace, 2015; Jonsson & Svingby, 2007; Knoch, 2011; McNamara, 1996; Panadero & Jonsson, 2013). Of the studies that do attempt to use quantitative methods to improve rubric performance, many

use the “many facets Rasch model” (MFRM) approach (Linacre, 1989), which is an extension to the 1PL Rasch model which allows for additional “facets” to be included in parameter estimation such as rater severity/leniency and rating scale step difficulty (Janssen, Meier, & Trace, 2015; Knoch, 2009; Meier, 2013). While this approach may lend unique insights into the differences in scoring tendencies by raters and rater reliability, it does not focus on the diagnostic ability of the rubric itself. Furthermore, the MFRM is limited in the scope of its application since it currently can only be run using the software FACETS (Linacre, 2010).

As a result, there exists a need for data based quantitative methods for designing and developing rubrics (Clauser, 2000; Harsch & Martin, 2012; Janssen, Meier, Trace, 2015; Knoch, 2009), just as there is a need for more DCAs. Conversely, Banerjee, Yan, Chapman, & Elliott (2015) developed a method for revising a rating scale for the writing section of a large-scale diagnostic assessment. However, the authors acknowledged that discussions of scale design and development are rare and that the approach they used was both singular and “extremely unusual”. Still, despite the current lack of research in rubric revision or development, rubric-scored tests that are currently in use must be monitored and modified regularly in order to ensure score reliability and validity (Banerjee, Yan, Chapman, & Elliot, 2015).

Therefore, it seems that there simultaneously exists a need to develop methods of improving the diagnostic ability of existing rubrics as well as a need to develop methods of retrofitting DCMs to existing assessments. It is then curious that these two dimensions of measurement, that are so similar in objectives and goals as well as their lack of investigation, are not more intertwined in the current literature. Reasons for the lack of

connection between the study of rubric revision and retrofitting DCMs are somewhat unclear but could be related to either their relative novelty in the field of psychometrics and/or the relatively complex, sometimes stringent, and unknown nature of the DCM framework or the unidimensional nature of rubric-scored assessments.

Indeed, rubric-scored items are unlike multiple-choice items and even many constructed-response items in that they produce graded polytomous data. In research wherein DCMs are retrofitted to assessments that have constructed-response items or items that produce polytomous data, it is not uncommon for researchers to determine a method of dichotomizing the item so that it simply produces answers that are considered correct or incorrect (Close, 2012; Lee, Park, & Taylan, 2011; Svetina, Dai, & Wang, 2017). If that same technique were to be applied to a rubric however, it may result in the loss of significant diagnostic information that rubrics provide as a result of their multilevel graded structure.

2.5.1 Hierarchical Attribute Structures in DCMs

Earlier it was discussed that a crucial matter to be aware of when conducting DCM attribute specification is the potential for attributes to have dependent relationships. In such cases, the mastery of one attribute depends upon the mastery of another attribute. When this occurs, a hierarchical attribute structure is present. Hierarchical attribute structures can have significant effects on Q-matrix design, classification accuracy, model fit, and test validity (Liu, Huggins-Manley, & Bradshaw, 2016; Liu & Huggins-Manley, 2016; Templin & Bradshaw, 2014). However, if their presence is known, then there are

techniques that can be applied that can actually lead to deeper insights regarding the nature and the relationships within those attribute hierarchies.

For example, Leighton, Gierl, and Hunka (2004) developed what is known as the Attribute Hierarchy Method (AHM), which is a DCM in which it is assumed that the measured cognitive attributes have hierarchical relationships and are dependent. The AHM is a variation on the rule-space method (Tatsuoka, 1983) in that it still observed response patterns of masters and non-masters according to a set of attributes and used Q-matrices to represent those attribute structures. The motivation to develop the AHM came from the authors' belief that modeling cognitive attributes necessitated the specification of a hierarchy and the desire to link cognitive theory to the rule-space method (Leighton, Gierl, & Hunka, 2004).

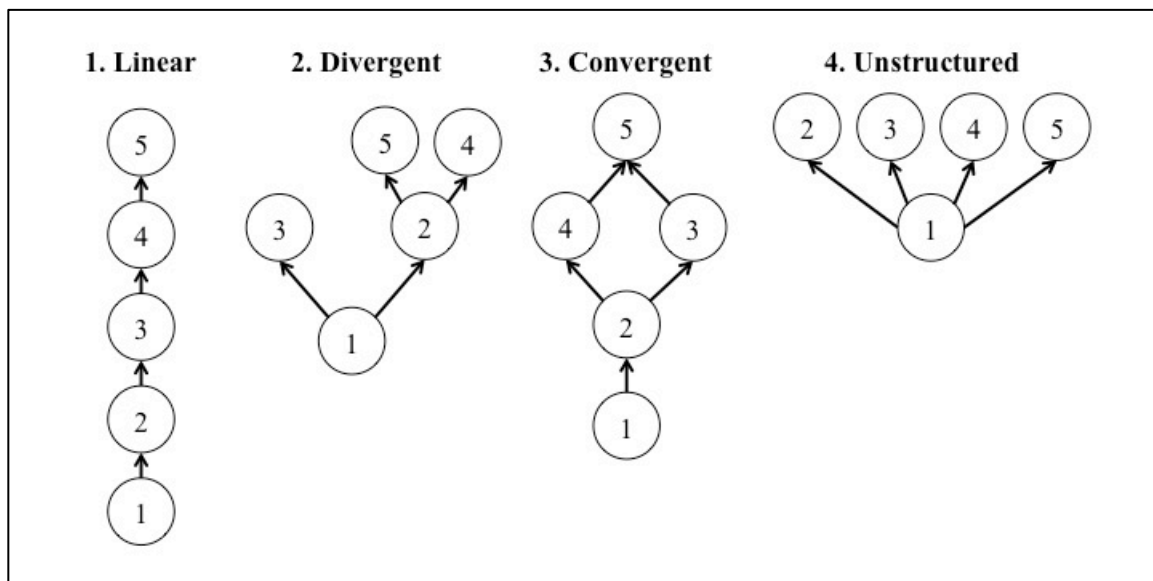


Figure 1. Four forms of hierarchical attribute structures.

The authors also identified four forms of hierarchical structures represented by five attributes, as seen in Figure 1. The *linear hierarchy* indicates that all five attributes are in succession, which further indicates that in order to master an attribute, it is required that all the attributes that precede it are also mastered. For example, in order to

master Attribute 3, an examinee must first master Attribute 1 and Attribute 2. In the *divergent hierarchy*, multiple paths separate from the initial parent attribute. This type of hierarchy could be used to represent assessments where there are multiple components that could be considered correct or incorrect, such as in constructed-response items (Leighton, Gierl, & Hunka, 2004).

In a *convergent hierarchy*, at some point in the hierarchy the path of attribute mastery diverges from attribute to multiple, and then converges to a common attribute. Such an attribute hierarchy may be used to represent an assessment wherein there is one desired outcome and multiple paths must be taken to produce it. Finally, the *unstructured hierarchy* represents a case where the only the first attribute is required to be mastered to master all other attributes. In this case, there is no unique relationship between the attribute structure and the total score.

One of the earliest examples of a linear attribute hierarchy structure was developed by Louis Guttman (1944, 1950) and was known as the *Guttman scale* or ‘Guttman scaling’ (Abdi, 2010). Guttman scaling essentially is a process of determining the degree of unidimensionality within a set of items (Andrich, 1985). The process is done by first defining the construct that the assessment is intended to measure. Next, dichotomous items must be developed and then ordered so if an item is to be marked as 0, all items that followed that item will also be marked as 0, and all items that precede it will be marked as 1. As a result of this process, attribute hierarchies can be clearly identified within a group of items (Schultz & Siegel, 1961). The example shown in Table 2 represents a perfect Guttman scale, in which a unidimensional Q-matrix has its items

ranked in order from the least amount of attributes required to answer correctly to the most.

Table 2. Example of a Q-matrix as a Perfect Guttman Scale

Item	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Attribute 6
1	1	0	0	0	0	0
2	1	1	0	0	0	0
3	1	1	1	0	0	0
4	1	1	1	1	0	0
5	1	1	1	1	1	0
6	1	1	1	1	1	1

If a Guttman model is valid, the marginal sum of each row or column (depending on which represents the attributes) will indicate the order in which they ascend. In practical usage perfect Guttman scales are rare since in most cases the construct that is being measured is actually multidimensional (Abdi, 2010). Since its conception, Guttman scales have been used to identify unidimensional attribute hierarchies in many diverse fields such as education, economics (Guest, 2000), social issues (Vimalraj Kumar, Mathialagan, & Sabarathnam, 2016), and anthropology (Peregrine, Ember, & Ember, 2004).

Returning to the concept of linear hierarchies, since the Guttman scale is sequential and requires that all attributes that precede an attribute must be mastered in order to master the attributes ahead of it in the sequence, it falls under the category of a linear hierarchy. Furthermore, this kind of Guttman scale linear hierarchy can also describe the attribute structure of most rubrics. Such an attribute structure may also be able to be represented by a Q-matrix. However, the Guttman scale type linear hierarchy is not the only simple attribute structure that can be represented by a Q-matrix.

Liu, Huggins-Manley, & Bradshaw (2016) present three approaches for parameterizing Q-matrices that represent attributes that follow a linearly hierarchical structure: the *independent approach*, the *adjacent approach*, and the *reachable approach*. The researchers also conducted a simulation study in which the three approaches were investigated for their effects on classification results, using fixed sample sizes ($N = 2,000$) over 1,000 replications per condition, controlling for guess and slip parameter values as well as the tetrachoric correlation between attributes. For example, Table 3 shows an example of an Independent Approach Q-matrix (Liu, Huggins-Manley, & Bradshaw, 2016).

Table 3. Example of an Independent Approach Q-matrix

Item	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Attribute 6
1	1	0	0	0	0	0
2	0	1	0	0	0	0
3	0	0	1	0	0	0
4	0	0	0	1	0	0
5	0	0	0	0	1	0
6	0	0	0	0	0	1

The *independent approach* models a particular form of attribute relationship where each item only measures one attribute. Although this may be considered an extreme method of attribute isolation, isolating attributes increases diagnostic classification accuracy. Specifically, if each attribute is measured the same number of times, and in isolation, classification accuracy will be higher than in other Q-matrix designs (Liu, Huggins-Manley, & Bradshaw, 2016). It should also be noted that the Q-matrix in Table 3 is also considered a “balanced” Q-matrix since each attribute is measured the same number of times and types (Liu & Huggins-Manley, 2016).

Furthermore, increasing the number of times each attribute is measured was also found to increase classification accuracy as well.

However, while measuring attributes in isolation may be the most efficient at identifying attribute mastery, it is also rare or unrealistic in some cases, as well as unsuitable for making inferences about the compensatory/non-compensatory or conjunctive/disjunctive relationships between attributes. Conversely, Liu, Huggins-Manley, & Bradshaw (2016) recommend that the single attribute item element of the independent approach be incorporated into the design of a Q-matrix in order to most effectively identify attribute mastery in situations where linear hierarchies exist within the attribute structure. Specifically, the researchers commend combining the design of the independent approach with the two other approaches for Q-matrix design presented in the study, the *adjacent approach* and the *reachable approach* shown in Table 4 and Table 5, respectively (Tatsuoka, 1983, 2009).

Table 4. Example of an Adjacent Approach Q-matrix

Item	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Attribute 6
1	1	1	0	0	0	0
2	0	1	1	0	0	0
3	0	0	1	1	0	0
4	0	0	0	1	1	0
5	0	0	0	0	1	1
6	0	0	0	0	0	1

The adjacent approach stipulates that each item measures a maximum of two directly linked attributes wherein one attribute is a prerequisite for mastering the other attribute. Take, for example, the Q-matrix in Table 4. In this example Attribute 1 could be a prerequisite of mastery for Attribute 2, and in order to answer Item 1 correctly, both attributes need to be mastered. However, in order to answer Item 2 correctly, both

Attribute 3 and Attribute 2 have to be mastered, and Attribute 2 is a prerequisite for Attribute 3. Therefore, Attribute 1 must be mastered in order to master Attribute 2, in order to master Attribute 3. Hence, the linear hierarchy that follows throughout.

The results of the study found that classification results for the adjacent approach either resembled or were better than those of the independent approach in some cases, when tests were of similar length. In conditions where item quality was low, the adjacent approach also produced better classification results than both the independent and reachable approaches as well (Liu, Huggins-Manley, & Bradshaw, 2016). The reason the adjacent approach performs better than the independent and reachable approaches is that the adjacent approach decreases the number of parameters required to be estimated by the model by eliminating impossible attribute mastery patterns as well as limiting the number of attributes an item can measure to two.

Table 5. Example of a Reachable Approach Q-matrix

Item	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Attribute 6
1	1	0	0	0	0	0
2	1	1	0	0	0	0
3	1	1	1	0	0	0
4	1	1	1	1	0	0
5	1	1	1	1	1	0
6	1	1	1	1	1	1

Similarly, the reachable approach (Liu, Huggins-Manley, & Bradshaw, 2016), an example of which is shown in Table 5, extends the adjacent approach by permitting the maximum number of attributes a single item can measure to be equal to the largest number of attributes possible. It should also be noted that this approach, unlike the independent and adjacent approaches, directly represents a perfect Guttman scale. In the example shown in Table 5, Attribute 1 must be mastered before Attribute 2, and so on

and so forth, and so the rules of dependency or rather, the nature of the hierarchical relationship between attributes is still the same as it was for the adjacency approach. As a result, a major differentiation of the design of the reachable approach to the independent approach and adjacent approach is that the reachable approach allows an item to measure the maximum possible number of attributes possible.

Moreover, the study by Liu, Huggins-Manley, & Bradshaw (2016) showed that the reachable approach design was the most affected by item quality. In other words, in cases where items had high guess or slip parameter estimates, it becomes more difficult to estimate the interactions between attributes as the number of attributes measured by the item increases. As a result, the reachable approach for Q-matrix design was the least recommended approach by the authors for representing attribute hierarchies.

2.5.2 Alternative Models for Rubric-Scored Tests

Polytomous-IRT models such as the Graded Response Model (GRM) (Samejima, 1969), Partial Credit Model (Masters, 1982), and the Generalized Partial Credit Model (G-PCM) (Muraki, 1992,1993) are all IRT models used to model items that produce graded polytomous data. Graded item responses refer to cases wherein item responses are divided into ordered categories in which the lowest category contributes the least to a person's test score and the highest category contributes the most (Baker & Kim, 2004). Although these models have existed for at least as long as DCMs, their diagnostic potential/capability has not yet been fully researched, although their mathematical theory is firmly grounded in IRT. For example, polytomous-IRT models still follow many of the

same assumptions as most other IRT models such as the assumption of unidimensionality (Immekus & Imbrie, 2008). Furthermore the results from these models are not easily translatable to individuals unfamiliar with IRT and therefore may not be the most beneficial for students or test administrators, although they may be quite useful for test developers. Nonetheless, polytomous-IRT models provide a reasonable alternative to analyzing rubric performance and so they will also be applied to the rubric-scored data so that the usefulness and depth of their insights may be compared to the proposed method of retrofitting.

2.6 Item Response Theory Models for Polytomous Data

In item response theory (IRT), most models are designed to measure and evaluate the performance of test items that produce dichotomously scored data, with responses being either correct or incorrect. The assumption behind this scoring procedure is that the only data available in an item response is the correctness of that response (Baker & Kim, 2004). In other words, most IRT models assume that the only salient information to be gained from an item response is whether or not that item has been answered correctly or incorrectly. However, this assumption does not hold true for all test items, especially for cases wherein responses are graded, meaning cases wherein responses have varying “degrees of correctness” or are nominal, such as rubric-scored test items and items from personality tests. In these cases, no response is either correct or incorrect and so the salient information is which specific response has been selected.

In cases where responses are graded, item responses are divided into ordered categories so that the lowest category contributes the least to a person's test score and the highest category contributes the most (Baker & Kim, 2004). In order to model these cases that produce graded response data (or any data that is characterized as ordered categorical responses), the graded response model (GRM) was developed (Samejima, 1969). However, the GRM was only intended to model ordered categorical data in general, without taking into account the different forms of ordered categorical data that exist. Masters (1982) delineated four different types of graded responses: repeated trials, counts, rating scales, and partial credit.

Masters argued that partial credit data comes from an observation format that requires the prior identification of multiple ordered levels that indicate success on an item and for which the partial completion of would award corresponding credit (Masters, 1982). In order to model this type of graded response data, the Partial Credit Model or PCM was developed. However, since the PCM model was developed it has also been extended by Muraki (1992) to include additional parameters in order to create a more flexible PCM, which then became known as the generalized partial credit model (G-PCM). However, although these polytomous IRT models have been around for some time, there is a dearth of research in their theory and application, especially with rubric-scored data.

Real rubric-scored data is difficult to produce, since it requires the employment of a scorer to evaluate and score each response. Rubric scores are often more preferable to multiple-choice scores for students as they are able to provide more detailed feedback. By modeling rubric-scored data using polytomous IRT models such as the graded response

model (GRM) or the generalized partial credit model (G-PCM), we may be able to make inferences about the performance of the test items the rubric is based on, the test population, and the rubric itself. Furthermore, what is unique about rubric data compared to multiple-choice data, is that rubric-data does not come directly from examinees. Rather, a scorer must be involved in order to determine an examinee's score. Therefore, applying polytomous IRT models to rubric-scored data may also lend insights into the performance of the scorers as well.

2.6.1 The Graded Response Model

In the GRM each item (i) is described by a slope parameter (α_i) and $j = 1 \dots m_i$ “between-category threshold parameters” (β_{ij}), where the number of item response categories equals $m_i + 1 = K_i$ (Embretson & Reise, 2000). Item slope parameter (α_i) represents the degree to which an item is able to distinguish how well an individual will perform on the exam, much like in dichotomous IRT models (Nering & Ostini, 2010). It is important to note that it is not necessary for items to have the same number of response categories and no complications arise from having items with different response formats. In cases where α_j remains constant across items, the GRM is deemed a *constrained graded response model* (CGRM) and is the model that is most commonly used. In cases where α_j varies between items, the model is deemed an *unconstrained graded response model* (UGRM) (Samejima, 1969).

To compute the category response probabilities in the GRM, two stages must be followed. Let us consider an example where a test item has $K = 6$ response options,

where examinees receive item scores of $x = 1 \dots 6$. Since there are 6 response options, there are $m_i = 5$ thresholds ($j = 1 \dots 5$) between the response options. Thus, one of the main issues in fitting the GRM is to determine the location of these thresholds on the latent trait continuum (Embretson & Reise, 2000). The first step in using the GRM to estimate response probabilities requires computing m_i curves for each item using the following equation:

$$P_{ix}^*(\theta) = \frac{e^{\alpha_i(\theta - \beta_{ij})}}{1 + e^{\alpha_i(\theta - \beta_{ij})}}$$

where $x = j = 1, \dots, m_i$. Each curve $P_{ix}^*(\theta)$ is known as a *boundary characteristic curve* (BCC) and represents the probability of an examinee's raw item response (x) falling in or above a given category threshold ($j = 1 \dots m_i$) conditional on trait level (θ) (Samejima, 1969). A visual example of these boundary characteristic curves (BCC's) can be seen in Figure 2, note that in Table 6, we can also see the values of each line's *between category threshold parameter* (β_{ij}), which indicates the location on the latent trait continuum wherein an examinee has a 0.50 probability of responding in or above category $j = x$.

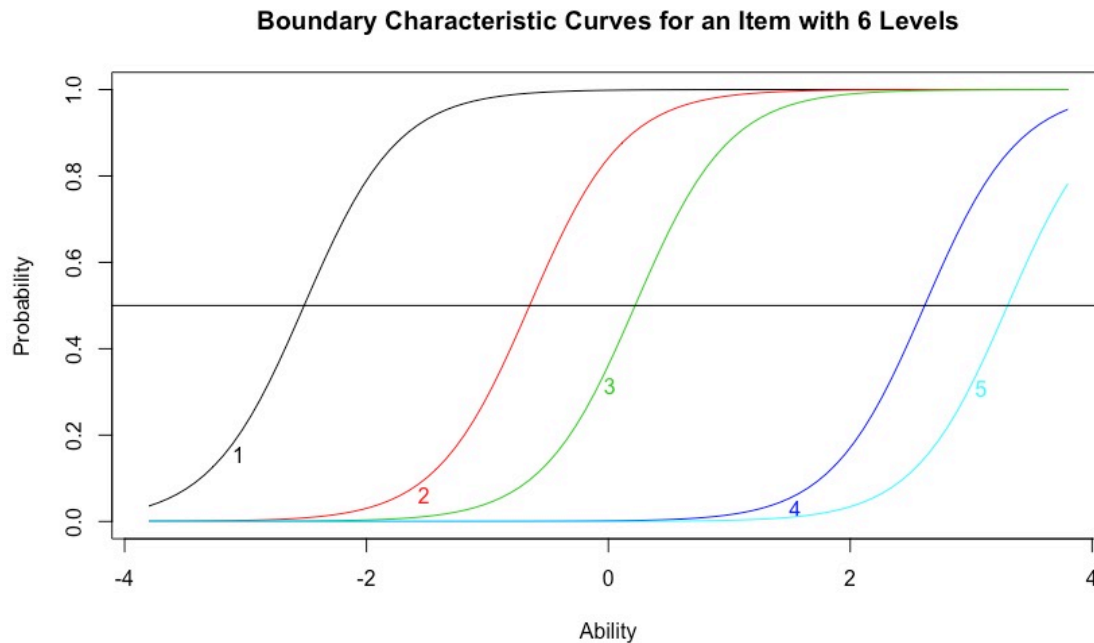


Figure 2. Example of boundary characteristic curves.

Table 6. Between Category Threshold Parameters for Figure 2

β_{i1}	β_{i2}	β_{i3}	β_{i4}	β_{i5}
-2.516	-0.651	0.221	2.616	3.301

For each between category threshold parameter (β_{ij}) a BCC must be estimated and so, in this example where there are six response categories, five β_{ij} parameters are estimated with one common item slope parameter (α_i). Interpretively, the value of each β_{ij} parameter represents the trait level necessary to respond above threshold j with 0.50 probability (Baker & Kim, 2004). In essence, the GRM treats each item as a series of $m_i = K - 1$ dichotomies (e.g., 1 vs 2,3,4,5,6; 1,2 vs 3,4,5,6; 1,2,3 vs 4,5,6; 1,2,3,4 vs 5,6; 1,2,3,4,5 vs 6) and 2PL IRT models are estimated for each dichotomy assuming that the slopes of the BCC's ($P_{ix}^*(\theta)$) are equal within each item (Embretson & Reise, 2000). Once each $P_{ix}^*(\theta)$ is estimated, the actual category response probabilities for $x = 1 \dots 6$ are calculated using the following equation:

$$P_{ix}(\theta) = P_{ix}^*(\theta) - P_{i(x+1)}^*(\theta)$$

where the probability of responding in or above the lowest category response is $P_{i1}^*(\theta) = 1.0$, and the probability of responding above the highest category response is $P_{i6}^*(\theta) = 0.0$, by definition. Thus, in the current example, the probability of selecting each of the six possible category response options is:

$$\begin{aligned} P_{i1}(\theta) &= 1.0 - P_{i2}^*(\theta) \\ P_{i2}(\theta) &= P_{i2}^*(\theta) - P_{i3}^*(\theta) \\ P_{i3}(\theta) &= P_{i3}^*(\theta) - P_{i4}^*(\theta) \\ P_{i4}(\theta) &= P_{i4}^*(\theta) - P_{i5}^*(\theta) \\ P_{i5}(\theta) &= P_{i5}^*(\theta) - P_{i6}^*(\theta) \\ P_{i6}(\theta) &= P_{i6}^*(\theta) - 0. \end{aligned}$$

Each $P_{ix}(\theta)$ is known as an *item category response curve* (ICRC) and represents the probability of an examinee selecting a particular response category given their trait level.

To summarize, the GRM item parameters determine the shape and location of the ICRC's ($P_{ix}(\theta)$) and BCC's ($P_{ix}^*(\theta)$). Each BCC ($P_{ix}^*(\theta)$) represents the probability of an examinee's response falling in or above a given category threshold conditional on trait level, and is estimated using between category threshold parameters (β_{ij}). Between category threshold parameters (β_{ij}) represent the point on the latent trait scale at which examinees have a 0.50 probability of responding above a category threshold or, in other words, responding in or above category $j = x$ (Embretson & Reise, 2000). Subsequently, each ICRC ($P_{ix}(\theta)$) represents the probability of responding in each category ($x = 1 \dots 6$) conditional on examinee trait level.

In order to fit the GRM to data, the item parameters must be estimated. Popular parameter estimation algorithms include the joint maximum likelihood estimation (JMLE) method and the marginal maximum likelihood estimation (MMLE) method

(Bock & Lieberman, 1970). Once a preferred parameter estimation algorithm is chosen, a numerical optimization method, such as the Newton-Raphson method or expectation-maximization (EM) algorithm must be chosen as well. Software such as Parscale (Muraki & Bock, 2003) and Multilog (Thissen, 2003) are commonly used for estimating polytomous IRT model parameters. More recently R (R Core Team, 2016) software packages have also become popular in estimating polytomous IRT model parameters such as ‘mirt’ (Chalmers, 2012) and ‘ltm’ (Rizopoulos, 2006).

2.6.2 The Partial Credit Model

The partial credit model (PCM) was originally designed to analyze test items that require multiple steps for which it is essential to award partial credit for completing at least one or some of those steps in the solution process (Embretson & Reise, 2000). As a result, the PCM is ideal for describing item responses to achievement tests where items award partial credit, as well as attitude or personality scale items where subjects rate their opinions or respond to statements on a multi-point scale. Unlike the GRM, in the PCM the probability of responding with a particular category is written directly as an exponential divided by the sum of exponentials. Subsequently, the PCM can be considered an extension of the 1PL IRT model, having all the standard Rasch model features such as the ability to distinguish separate person and item parameters (Masters, 1982). Using the same example from the previous section on the GRM, let us assume that item i is scored $x = 1, \dots, m_i$ for an item with $K_i = m_i + 1$ response categories. Thus for $x = j$ the ICRC for the PCM is written as:

$$P_{ix}(\theta) = \frac{e^{[\sum_{j=0}^x(\theta - \delta_{ij})]}}{\sum_{r=0}^{m_i} [e^{\sum_{j=0}^r(\theta - \delta_{ij})}]}$$

where $\sum_{j=0}^0(\theta - \delta_{ij}) \equiv 0$.

The *step difficulty* term δ_{ij} ($j = 1, \dots, m_i$) is associated with a category score of j and indicates that a higher value of δ_{ij} , representing a more difficult level, relative to other levels within an item. A more direct interpretation of δ_{ij} would be that it is the point on the latent trait scale at which two consecutive category response curves intersect (Embretson & Reise, 2000). In other words, the step difficulty term (δ_{ij}) represents the position on the latent trait scale associated with the transition from one category response level to the next, and within each item there are m_i step difficulty terms for an item with $m_i + 1$ response categories. Thus the step difficulty term (δ_{ij}) replaces the function of the category threshold parameters (β_{ij}) in the GRM, but does not represent a point on the latent trait continuum where one has a .50 probability of responding above a category threshold. Instead, the step difficulty term (δ_{ij}) represents the relative difficulty of each level within an item (Embretson & Reise, 2000).

As a result, within an item, some category response levels may be less difficult than other category response levels, even though they are ordered higher on the graded scale and vice versa. For example, it may be possible for it to be more difficult for an examinee to go from a score of one to two than it is for an examinee to go from two to three. This phenomenon of un-ordered step parameters is known as a *reversal* (Dodd & Koch, 1987). In general, if the step difficulty terms (δ_{ij}) are ordered within an item, then every category response option is most probable on at least one position on the latent trait continuum. Alternatively, if there is a “reversal” in step difficulty parameters within an

item, it is guaranteed that there will be at least one category response option that is never the most likely option at any position on the trait continuum (Andrich, 1988).

Essentially, the δ_{ij} parameter indicates the location on the latent trait continuum where category response curves intersect, thus indicating where on the latent trait continuum where one category becomes relatively more probable than the previous category (Masters, 1982). A visual example of an item category response curve can be seen in Figure 3, how the δ_{ij} parameter values in Table 7 correspond to the location of the line intersections.

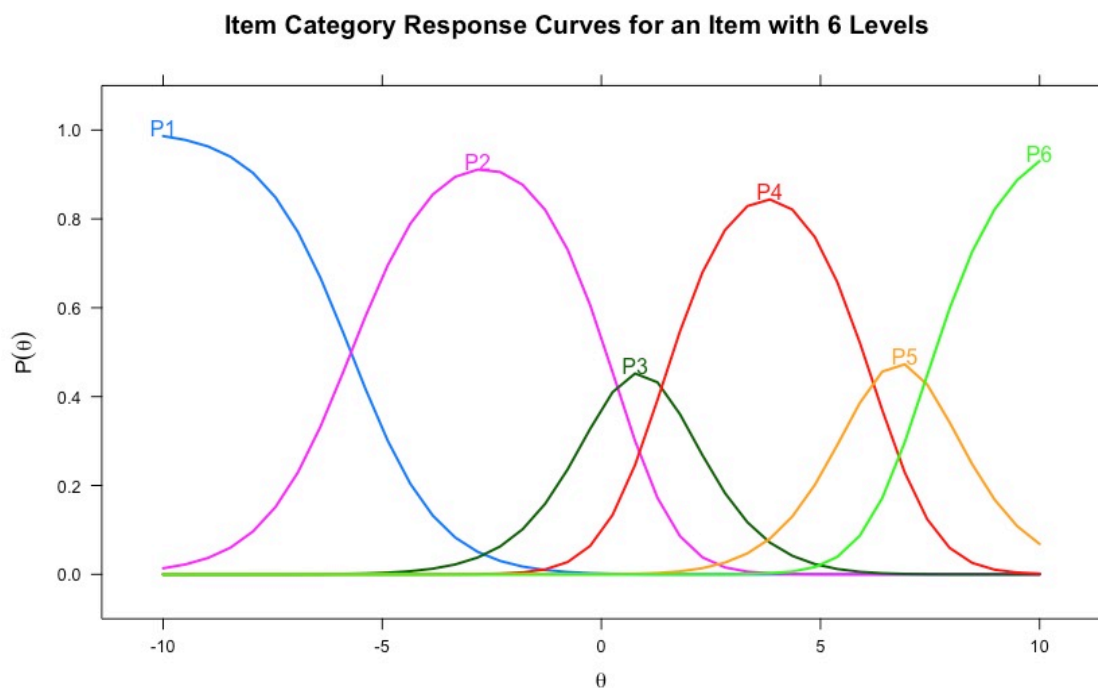


Figure 3. Example of item category response curves.

Table 7. Between Category Threshold Parameters for Figure 3

δ_{i1}	δ_{i2}	δ_{i3}	δ_{i4}	δ_{i5}
-5.713	0.360	1.377	6.199	7.387

Interpretively, the δ_{ij} parameter points represent where the completion of a step becomes more likely than non-completion, given that an examinee has already completed the previous steps (Embretson & Reise, 2000). In other words, the ICRC's calculated by the PCM can be used to compute the most probable response for examinees at any point on the latent trait continuum.

One unique benefit of the mathematical structure of the PCM is that it is possible to use the item parameters to graph the expected score for an item along the latent trait continuum, thus representing the expected raw item score for examinees at any particular trait level. The following equation represents the mathematical procedure for such calculations:

$$E(X) = \sum_{x=0}^{m_i} xP_x(\theta)$$

As a result, an examinee's raw score becomes a sufficient statistic for estimating examinee latent trait ability, which also means that examinees with the same raw score on a set of items that fit the PCM are also estimated to possess equal positions on the latent trait continuum. However, it must be assumed that all items are uniformly associated to the underlying latent trait or, in other words, represent different components of the same overarching test construct.

2.6.3 The Generalized Partial Credit Model

In order to account for items within a scale that differ in slope (i.e. items with different levels of item discrimination), Muraki (1992; 1993), developed a generalization of the

PCM known as the generalized partial credit model (G-PCM). The G-PCM expands on the original equation of the PCM by instituting a slope parameter (α_i):

$$P_{ix}(\theta) = \frac{e^{[\sum_{j=0}^x \alpha_i(\theta - \delta_{ij})]}}{\sum_{r=0}^M [e^{\sum_{j=0}^r \alpha_i(\theta - \delta_{ij})}]}$$

where $\sum_{j=0}^0 \alpha_i(\theta - \delta_{ij}) \equiv 0$.

The step difficulty parameters (δ_{ij}) in the G-PCM do not change in their interpretation from the PCM, and continue to represent the intersection point of two adjacent category response curves (Embretson & Reise, 2000). In other words, the step difficulty parameters (δ_{ij}) still indicate the points on the latent trait continuum where one category response option becomes more probable than the preceding category response option, given that the examinee has completed the previous steps. However, the interpretation of the slope parameter (α_i) is different from its interpretation in dichotomous IRT models. In the G-PCM, the slope parameters (α_i) indicate the degree to which category response option probabilities change as the latent trait variable θ changes (Muraki, 1992). Consequently, as α_i decreases the ICRC's flatten and as α_i increases the ICRC's become more peaked.

2.7 Summary of the Literature

To summarize, the need for formative assessments has led to the development of a psychometric framework known as the diagnostic measurement framework. The diagnostic measurement framework seeks to maximize the diagnosticity of assessments

in order to provide students, stakeholders, and administrators with rich diagnostic information on test taker's specific strengths and weaknesses on a set of skills determined necessary to be mastered within a given construct. Rich diagnostic feedback is obtained using mathematical models known as diagnostic classification models (DCMs). Ideally, an assessment would be designed with the diagnostic measurement framework in mind so that the data produced from the assessment is intended to fit to DCMs, thus producing the most optimally diagnostic information. Assessments that are designed in this way are known as diagnostic classification assessments (DCAs).

However, development in DCM theory has far outpaced development in DCAs. Accordingly, psychometricians began exploring the concept of retrofitting, which is the process of applying DCMs to data from exams that were not designed with the diagnostic measurement framework in mind.

One of the major challenges that arose from this endeavor was the issue of attribute specification. The diagnostic measurement framework specifies that in an ideal situation, the relevant attributes to an assessment should be specified before the assessment is developed, in order to ensure that item/attribute measurement construct validity. However, in retrofitting the items already exist, so the task then became how best to extract or determine what the attributes are from the assessment.

The need for more formative assessment has also led researchers to seek ways of improving the diagnostic ability of rubric-scored tests. However, similar to the shortage of development in DCAs, there is a shortage of development in mathematical techniques for improving rubric diagnosticity. Thus, we might ask if it is possible for DCMs to be

retrofitted to rubrics and rubric-scored items in a way that improves their diagnostic capability.

2.8 Goals of the Present Study / Theoretical Contributions

Therefore, it is the theoretical goal of this study to propose methods for retrofitting DCMs to rubric-scored constructed-response items. Past research has shown that DCMs can be successfully retrofitted to multiple-choice items that were not designed within the diagnostic measurement framework, if the necessary attributes are specified correctly. However, there is little or no research on retrofitting DCMs to rubrics and rubric-scored items. Developing methods for doing so constitutes the theoretical contributions this study aims to make.

The first specific contribution is to examine whether or not a rubric-scored item can reasonably be conceived of as a DCA. Most rubrics, by their design, have multiple sub-areas that have even finer-grained subskills within them that a respondent must demonstrate mastery of in order to be rated as proficient or higher in those sub-areas. Furthermore, each sub-area represents a different factor for demonstrating mastery in the same overarching construct, thus being inherently both unidimensional and multi-dimensional at the same time. Moreover, the graded structure of a rubric's sub-areas translates into a Q-matrix as a nested structure of attributes known as a linear hierarchy and in fact, is a Guttman scale.

The second theoretical contribution is to describe the conditions under which a numeric or merely ordinal rubric score can be represented by a Q-matrix. By “conditions”

we are referring to the type of DCM model being retrofitted to the data and the specific Q-matrix structure. We will also examine the set of nominal skills that define the rubric, in order to comment on the related assessment problem, which is when can a set of nominal skills described in a scoring rubric be assessed with a unidimensional rubric score in a way that maximizes its diagnostic potential? The set of nominal attributes within each sub-area defines a large space of possible knowledge states, but the rubric scoring defines only some of these as possible. Therefore by retrofitting DCMs to the rubric and analyzing the fit, parameter estimates, and potential inferences, we may be able to define how well the rubric is actually able to diagnose the nominal-attribute knowledge states it is designed to measure.

The third theoretical contribution is to gain insights regarding the possibility of using methods for retrofitting DCMs to rubric-scored data in order to guide future rubric design. In other words, if the perceived attribute structure of the rubric can be represented within the mathematical space of the Q-matrix, can the results of a DCM analysis of the data provide implications for redesigning the Q-matrix in a more optimal way for the data and, if so, can the new design of the Q-matrix then have implications for the redesign of the rubric. If Q-matrices can be redesigned based on the results of retrofitting DCMs to rubric-scored data, by retrofitting DCMs to rubric scored data, it may be feasible to then redesign the rubric itself, therefore improving its diagnostic ability. This final contribution may be considered the “reverse problem”, in essence the initial quandary was to determine whether or not rubrics can viably be represented as a Q-matrix within the DCM framework, therefore the reverse situation is to determine if a Q-matrix within a DCM framework can then be represented as a rubric.

3 Method

This section describes the methods that are proposed for retrofitting DCMs to rubric-scored constructed-response items, beginning with an overview of each step in the process. Next a description is given of the measures used, participants that comprise the data, including the demographics of each of the samples. This is followed by an explanation of the exact details of the proposed methods for retrofitting a DCM to rubric-scored constructed-response items, the software used to conduct the analysis, and a description on some of the parameters that are examined in the analysis.

3.1 Overview

The goal of the present study is to develop general methods by which a rubric-scored essay-based test can be “retro-fit” as a diagnostic test within the context of the diagnostic measurement framework, and illustrate these methods using the Collegiate Learning Assessment+ (CLA+). This goal will be achieved in six steps.

1. Describe methods to transform polytomous constructed-response scores into dichotomous “pseudo-items” for the Q-matrix by adapting Tutz’s (1997) sequential response mechanism, so that the data can be fit to a dichotomous DCM framework.
2. Design Q-matrices for the constructed-response section of the CLA+ based on the hierarchical structure of specific skills described in the test rubric specification for the three scoring sub-areas.

3. As an alternative method, design Q-matrices using an automatic M-attribute method that automatically generates Q-matrices based on the number of score levels of the rubric.
4. Apply constrained versions of the DINA model for data with linear attribute hierarchies (de la Torre, 2009; de la Torre & Karelitz, 2009) to the constructed-response section using Q-matrices derived by the above two methods, gathering model fit and item fit statistics, attribute correlation, information-based item discrimination indices for DCMs, (Rupp, Templin, & Henson, 2010), and skill mastery classification estimates. Five different conditions of parameter constraints will be applied in which both the guessing and slipping parameters are constrained to zero, only the guessing parameter is constrained to zero, only the slipping parameter is constrained to zero, the guessing parameter is constrained to always be less than the slipping parameter, and no constraints are placed on either parameter.
5. Analyze and compare the results of the two attribute coding methods by which DINA models can be applied to constructed-response tests in order to evaluate the effectiveness of the proposed rubric score conversion methods.
6. For comparison purposes, apply a polytomous IRT Generalized Partial Credit Model (G-PCM) to the unconverted constructed-response section.
7. Analyze and compare the results between the dichotomous DINA model and the G-PCM in order to evaluate the utility and effectiveness of the proposed methods for retrofitting constructed-responses tests as DCMs

3.2 An Application – The Collegiate Learning Assessment +

All participants completed the Collegiate Learning Assessment+ (CLA+), which is a two-part assessment that measures critical-thinking and written-communication for diagnostic purposes and for studying growth in those skills within postsecondary education institutions, developed and administered by the Council for Aid to Education (CAE) (Steedle, 2012; Zahner, 2014). Traditionally used as an instrument to measure the value-added growth of an institution (Benjamin, 2014), the CLA+ is currently being used in a number of ways, including as a performance-based assessment of generic thinking and writing skills, which means that it does not require any specific prior content knowledge, but instead is used to demonstrate the examinee's current level of skill mastery. Each section presents the student with a prompt or scenario, which both the constructed essay response and the multiple-choice items are based on, so that the multiple-choice items are scenario-based as well. Since its inception in 2002, almost half a million students from over 750 institutions have taken the CLA and CLA+ (Lehrfeld, Muller, & Zahner, 2017). The assessment is comprised of two sections (multiple-choice and constructed-response), administered online, and takes a maximum of 90 minutes to complete in total. A sample form of the assessment can be found in Appendix II.

The raw scores of each multiple-choice subsection are scaled via a linear transformation, in order to correct for variation in difficulty between test versions, and equated to the performance of the original multiple-choice section population. Similarly, the constructed response section raw scores are summed across all three sub-areas measured by the rubric, and scaled by linear transformation and equated to the

performance of the norm population. The scaled constructed-response and multiple-choice section scores are then averaged together to form a raw total CLA+ score, which is then also scaled to the total test scores of the norm population as well (Council for Aid to Education, 2017). CLA+ test scores have been found to be both reliable at the student level (Zahner, 2013) as well as predictive of college GPA (Zahner, Ramsaran, & Steedle, 2012) and positive post-college outcomes such as employment, salary, and enrollment (Zahner & James, 2016).

3.2.1 The Rubric-Scored Constructed-Response Item Section

The constructed-response section is known as the *performance task* (PT), and measures examinees according to three sub-areas: *analysis and problem solving* (APS), *writing effectiveness* (WE), and *writing mechanics* (WM). The PT first presents the examinee with a problem based on a real-world scenario. It is left entirely up to the examinee to decide on a course of action and to justify the decision with information from a document library. The document library contains six to eight different documents that they can use to provide evidence for and strengthen their argument. These documents range in form and include such references as technical reports, data tables, newspaper articles, office memoranda, and/or emails (Council for Aid to Education, 2018). Each examinee is given one hour to construct their response. Upon completion of the exam, their response is scored according to the three sub-areas (APS, WE, WM).

The *analysis and problem solving* (APS) sub-area measures the degree to which an examinee has made a logical decision or conclusion and supported it by utilizing

information from the documents provided. Scorers judge the submitted response on whether or not it has a definitive or discernable position on the issue at hand and how much analysis has been given on the subject. The *writing effectiveness* (WE) sub-area measures the degree to which the examinee has constructed a well-organized and logical response for their argument. Scorers judge a response on whether or not the examinee has elaborated on any of their facts or statements provided and whether or not the response has been constructed in a clear and coherent manner. Finally, the *writing mechanics* (WM) sub-area measures the degree to which the examinee's response demonstrates facility with the conventions of standard written English and control of the English language, including syntax and vocabulary complexity. Each category is measured on a scale of 1 to 6, with 1 being the lowest score possible and 6 being the highest score possible.

Each response is graded by two scorers on each of the three sub-areas according to the rubric, producing six scores (Council for Aid to Education, 2017). Each pair of sub-area scores are then averaged together and summed across sub-areas to produce a final total score for the constructed-response performance-test (PT) section. If the response given by a participant does not meet the minimum requirements for a score of one on the APS sub-area, the constructed response is considered inadmissible and the student does not receive a score for their response at all. As a result, the lowest total score an individual may receive on the PT section is a three and the highest score they may receive is an 18. However, in order for the Q-matrix to be as identifiable as possible (Chen, Liu, Xu, & Ying, 2015; Groß & George, 2014) i.e. in order to estimate a class of knowledge states in which no attributes are mastered, respondents that did not meet the

minimum requirements for a score of one in the APS sub-area (and therefore both the WE and WM sub-areas as well) will be included in the analysis.

Before becoming scorers, all scorers must undergo an extensive training process lead by CAE item editors and psychometricians (Council for Aid to Education, 2017). All operational performance tasks use a combination of automated and human scorers, unless the automated system identifies an irregular pair of responses, in which case the response is flagged and is sent to a human scorer to be scored by a second human instead. To ensure scorer calibration verification system is set in place wherein scorers that fail to accurately score multiple verification responses are removed from the scoring system and placed in recalibration (Council for Aid to Education, 2017). If a scorer continues to score inaccurately, they are either trained further or removed from scoring.

3.2.2 Participants

The participants consist of a total of 1,618 college freshmen and 330 college seniors who took the same form of the CLA+ in the fall of 2013 and spring of 2014 from 12 different US post-secondary institutions, across two different versions. The total population is divided into two samples, according to test version. Sample A consists of the students who took Test Form A, specifically 998 freshmen who took the exam in the fall of 2013. Sample B consists of the students who took Test Form B, 612 freshmen who took the exam in the fall of 2013 and 314 seniors who took the exam in the spring of 2014.

3.2.3 Sample A Demographics

Of those students in Sample A, 64% self-reported as female, 34% male, and 2% declined to state a gender. The racial composition of the sample was 62% white, 12% Hispanic or Latino, 12% African-American, 6% Asian, 1% American Indian/Alaska Native/Indigenous, 3% other, and 4% declined to state any race. For 89% of the participants, English was their primary language, while the remaining 11% reported that a language other than English was their primary language. In terms of parent's highest level of education, 3% of participants reported less than high school, 19% reported high school, 24% reported some college, 35% reported a bachelor's degree, and the remaining 20% reported graduate or post-grad degree. In terms of field of study, 11% of the sample had majors in social sciences, 24% in sciences & engineering, 11% in humanities and languages, 13% in business, 27% in helping services, and 14% reported their major as "undecided/other/not-applicable".

3.2.4 Sample B Demographics

Of those students in Sample B, 60% responded as female, 38% male, and 2% declined to state a gender. The racial composition of the sample was 54% white, 15% Hispanic or Latino, 13% African-American, 9% Asian, 1% American Indian/Alaska Native/Indigenous, 5% other, and 4% declined to state any race. For 82% of the participants, English was their primary language, while the remaining 18% reported that a language other than English was their primary language. In terms of parent's highest

level of education, 17% of participants reported less than high school, 19% reported high school, 25% reported some college, 29% reported a bachelor's degree, and the remaining 19% reported graduate or post-grad degree. In terms of field of study, 13% of the sample had majors in social sciences, 24% in sciences & engineering, 13% in humanities and languages, 13% in business, 26% in helping services, and 11% reported their major as "undecided/other/not-applicable".

3.3 Retrofitting the DINA Model to Constructed-Response Items

In this section each step of the proposed methods for retrofitting DCMs to rubric-scored constructed-response items is detailed. First described is the proposed process to convert polytomous graded scores into dichotomous variables so that they can be modeled using dichotomous DCMs. Next, two methods for developing Q-matrices for rubric-scored data are explained, beginning with a method that is based on the language of the rubric ("Rubric Coding") and a method that is based on the number of levels of the rubric ("Full-Score Coding").

3.3.1 Polytomous Score to Dichotomous Vector Conversion

Although the multiple-choice section of the CLA+ can be readily analyzed with the DINA model, it is not straightforward to do so for the constructed response section.

While the constructed-response section produces polytomous rubric scores, the DINA model requires items to have only binary “correct” or “incorrect” results. In order to resolve this problem, the study will employ nested or hierarchical attribute sets (Leighton, Gierl, & Hunka, 2004), which can be used in this case given the graded structure of rubric scores. Such nested attribute sets have been used to model a *sequential response mechanism* as used in the *sequential model* (Tutz, 1997), which was proposed to model ordered response categories that represent consecutive steps in problem solving. In other words, the sequential model exploits problems in which the next step solving a problem can only be performed successfully if all prior steps have also been completed successfully.

Let the graded response for item i and person j be given by the response variable $U_{ij} \in \{0, \dots, m_i\}$ and assume item i has levels $0, \dots, m_i$, where 0 represents the lowest and m_i represents the highest level. The sequential model assumes that each item is solved sequentially. Let $U_{ijh}, h = 1, \dots, m_i$, represent the step from level $h - 1$ to level h , where $U_{ijh} = 1$ represents a successful transition and $U_{ijh} = 0$ represents an unsuccessful transition. The first step in the mechanism always begins at level 0. If the transition to level 1 is unsuccessful, the process stops and the examinee’s score is $U_{ij} = 0$. If the transition to level 1 is successful, the examinee’s score is at least level 1 and $U_{ij} = 1$. Therefore, $U_{ij} = 0$ if $U_{ij1} = 0$ and $U_{ij} \geq 1$ if $U_{ij1} = 1$.

The value of U_{ij} is *at least* 1 if the individual is successful on the transition from level 0 to level 1, because they may also be successful on the next transition, unless otherwise indicated. For example, if $U_{ij1} = 1$ the individual progresses to the next transition, which is from level 1 to level 2. If the transition from level 1 to level 2 is

successful, then $U_{ij} \geq 2$, otherwise, if the transition from level 1 to level 2 is unsuccessful, $U_{ij2} = 0$ and $U_{ij} = 1$. Simply put if the transition from level 1 to level 2 were to be unsuccessful, $U_{ij} = 1$ given $U_{ij} \geq 1$ if $U_{ij2} = 0$. Otherwise, if the transition from level 1 to level 2 were to be successful, $U_{ij} \geq 2$ given $U_{ij} \geq 2$ if $U_{ij2} = 1$. As a result, conditioning on $U_{ij} \geq h$ is an integral component of the mechanism, since the next step is only considered if the previous is successful.

In essence, if the transition from a primary level to a secondary level is unsuccessful, the response variable is equal to the primary level. Otherwise, the response variable is greater than or equal to the primary level if the transition to a secondary level is successful. Thus, the sequential response mechanism (Tutz, 1997) is modeled using the equation:

$$U_{ij} = h \text{ given } U_{ij} \geq h \text{ if } U_{ij,h+1} = 0.$$

Table 8 shows how the six possible score levels from the CLA+ scoring rubric are represented using a nested attribute structure, as in the sequential model. Each column of Table 8 represents a “pseudo-item”, constructed to represent a hypothetical level of achievement on the scoring rubric. In this way, the level of each sub-area in the rubric becomes a pseudo-item, and each score received indicates how many of those pseudo-items can be considered to have been answered “correctly” in sequential order. As a result, each sub-area’s score on the rubric corresponds to a particular pseudo-item response pattern

Table 8. Table for Converting Rubric Sub-Area Scores into Pseudo-Item Response Patterns

Score	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	1	0	0	0	0	0
2	1	1	0	0	0	0

3	1	1	1	0	0	0
4	1	1	1	1	0	0
5	1	1	1	1	1	0
6	1	1	1	1	1	1

3.3.2 Rubric to Q-matrix Conversion (Rubric Coding)

The process of converting the CLA+ rubric (found in Appendix I), and therefore any rubric, can be broken down into three steps: identify and define the sub-areas within the rubric, define each sub-area's attributes and specify the attribute structure of each sub-area, and determine whether or not at least some degree of mastery must be demonstrated for each attribute at each level of the rubric. In the case of the CLA+ the first step is relatively straightforward as the rubric is scored according to three sub-areas: Analysis and Problem Solving (APS), Writing Effectiveness (WE), and Writing Mechanics (WM). The sub-areas are also predefined according to the rubric and their definitions are observed in Figure 4.

Sub-Area	Description
APS	Making a logical decision or conclusion (or taking a position) and supporting it by utilizing appropriate information (facts, ideas, computed values, or salient features) from the Document Library
WE	Constructing organized and logically cohesive arguments. Strengthening the writer's position by providing elaboration on facts or ideas (e.g., explaining how evidence bears on the problem, providing examples, and emphasizing especially convincing evidence)
WM	Demonstrating facility with the conventions of standard written English (agreement, tense, capitalization, punctuation, and spelling) and control of the English language, including syntax (sentence structure) and diction (word choice and usage)

Figure 4. CLA+ scoring rubric sub-area definitions.

For each identified sub-area, a separate Q-matrix is constructed, thereby in effect treating each sub-area score as its own ‘pseudo-exam’ within the context of a DCM framework. Using the graded polytomous score to dichotomous item response conversion method discussed earlier, the number of levels for scoring each sub-area on the rubric then becomes the number of pseudo-items in each pseudo-exam. For example, in the case of the CLA+’s scoring rubric, there are three sub-areas scored on six levels each. Therefore the CLA+ rubric converts to three “sub-exams” that are each six pseudo-items long.

The second step in the process for conversion is to define within each sub-area the individual components that makes up the rubric’s scoring level gradient, treating each component as a separate attribute. Each defined attribute should in effect be a summary of what that component is thought to measure in the rubric-scoring gradient. In doing so each component addressed within the rubric’s sub-area becomes an attribute in that sub-area’s Q-matrix. As a result, this step also determines the number of attributes measured by the Q-matrix. For example, Figure 5 shows the separate components identified within the Analysis and Problem Solving (APS) sub-area’s scoring gradient. Three attributes have been identified and defined, therefore there are three attributes represented in the APS sub-area’s pseudo-exam Q-matrix.

Attribute	Description
1	Stating or implying a decision/conclusion/position
2	Providing analysis as support by comprehensively addressing relevant documents
3	Addressing contradictory information or alternative decisions/conclusions/positions

Figure 5. Analysis and problem solving (APS) sub-area attribute definitions.

Once the attributes have been identified and defined, each sub-area's attribute specifications must be delineated according to each level, that is, the association of attributes with scoring levels of the rubric. This step is relatively straightforward as this process consists of specifying which attribute's criterion applies to which level, and is provided by the rubric. The Analysis and Problem Solving (APS) sub-area's attribute specifications can be seen in Table 9.

The next and final step is less straightforward. It is the process of determining what degree of mastery must be demonstrated for each attribute to be successfully displayed at each level, based on the description of each attribute's criterion at each level. In essence, at each level and for each attribute, it must be determined, based on the language of the rubric and the attribute definitions, whether or not at least some degree of mastery of that attribute must be demonstrated in a response in order for a response to meet the criteria of that level.

Table 9. Analysis and Problem Solving (APS) Sub-Area Attribute Specifications

Level	Attribute 1	Attribute 2	Attribute 3
1	May state or imply a decision/conclusion/position	Provides minimal analysis as support (e.g., briefly addresses only one idea from one document) or analysis is entirely inaccurate, illogical, unreliable, or unconnected to the decision/conclusion/position	N/A
2	States or implies a decision/conclusion/position	Provides analysis that addresses a few ideas as support, some of which is inaccurate, illogical, unreliable, or unconnected to the decision/conclusion/position	N/A
3	States or implies a decision/conclusion/position	Provides some valid support, but omits or misrepresents critical information, suggesting only superficial analysis and partial comprehension of the documents	May not account for contradictory information (if applicable)
4	States an explicit	Provides valid support that	May attempt to address

	decision/conclusion/position	addresses multiple pieces of relevant and credible information in a manner that demonstrates adequate analysis and comprehension of the documents; some information is omitted	contradictory information or alternative decisions/conclusions/positions (if applicable)
5	States an explicit decision/conclusion/position	Provides strong support that addresses much of the relevant and credible information, in a manner that demonstrates very good analysis and comprehension of the documents	Refutes contradictory information or alternative decisions/conclusions/positions (if applicable)
6	States an explicit decision/conclusion/position	Provides comprehensive support, including nearly all of the relevant and credible information, in a manner that demonstrates outstanding analysis and comprehension of the documents	Thoroughly refutes contradictory evidence or alternative decisions/conclusions/positions (if applicable)

For example, observing Level 1 in Table 9, Attribute 1 can be determined to be required since it is required that at a minimum a response at least state or imply a decision/conclusion/position. It can be inferred from the description of the attribute at that level given by the rubric that at least some degree of mastery must be demonstrated in order to qualify as a Level 1 response. Therefore Attribute 1 is indicated as required in the Analysis and Problems Solving (APS) Q-matrix for Pseudo-Item 1. The language for Attribute 2 at Level 1 however, may not be considered to appear to indicate that Attribute 2, which is defined as “providing analysis as support by comprehensively addressing relevant documents” as shown in Figure 5, is required to be demonstrated as a requirement for Level 1. Therefore Attribute 2 will not be indicated as required in APS’s Q-matrix for Pseudo-Item 1. Lastly, Attribute 3 is not addressed as part of the criteria for a response to meet the requirements for a response to be considered at Level 1, nor is it indicated as required for Pseudo-Item 1 in APS’s Q-matrix. Using this same process for

each level, a Q-matrix is constructed that represents the rubric's Analysis and Problem Solving sub-area, final product of which can be observed in Table 10.

Table 10. Q-matrix Design for Analysis and Problem Solving (APS) Sub-Area

Pseudo-Item	Attribute 1	Attribute 2	Attribute 3
1	1	0	0
2	1	1	0
3	1	1	0
4	1	1	1
5	1	1	1
6	1	1	1

The process is then applied to the remaining two sub-areas measured by the scoring rubric, beginning at the step of attribute definition. Again, it must be defined within the sub-area the individual dimensions that comprises the rubric's level gradient, considering each dimension a separate attribute. The attribute definitions for Writing Effectiveness (WE) are shown in Figure 6.

Attribute	Description
1	Developing convincing, logical, and cohesive argument
2	Providing valid and comprehensive elaboration on relevant information

Figure 6. Writing effectiveness (WE) sub-area attribute definitions.

In this case, only two separate attributes are identified that make up the criterion used to determine whether or not a response is at a particular level on the rubric. Next, the attribute specifications for Writing Effectiveness must be delineated, which is shown in Table 11.

Table 11. Writing Effectiveness (WE) Sub-Area Attribute Specifications

Level	Attribute 1	Attribute 2
1	Does not develop convincing arguments; writing may be disorganized and confusing	Does not provide elaboration on facts or ideas
2	Provides limited, invalid, over-stated, or very unclear arguments; may	Any elaboration on facts or ideas tends to be vague, irrelevant, inaccurate, or unreliable

	present information in a disorganized fashion or undermine own points	(e.g., based entirely on writer's opinion); sources of information are often unclear
3	Provides limited or somewhat unclear arguments. Presents relevant information in each response, but that information is not woven into arguments	Provides elaboration on facts or ideas a few times, some of which is valid; sources of information are sometimes unclear
4	Organizes response in a way that makes the writer's arguments and logic of those arguments apparent but not obvious	Provides valid elaboration on facts or ideas several times and cites sources of information
5	Organizes response in a logically cohesive way that makes it fairly easy to follow the writer's arguments	Provides valid elaboration on facts or ideas related to each argument and cites sources of information
6	Organizes response in a logically cohesive way that makes it very easy to follow the writer's arguments	Provides valid and comprehensive elaboration on facts or ideas related to each argument and clearly cites sources of information

When deciding whether or not to indicate each attribute as required for a chosen level, the decision must again be based on the language used in the rubric for each identified attribute. For example, in Table 11 we see that for Attribute 2, Level 1's description is that the response given "does not provide elaboration on facts or ideas". Given that the definition of Attribute 2 is "providing valid and comprehensive elaboration on relevant information", it may be inferred that it is not necessary to demonstrate some degree of mastery of Attribute 2 in a response in order for that response to meet the requirements of Level 1. As a result, Attribute 2 will not be indicated as required for Pseudo-Item 1 in the Writing Effectiveness (WE) Q-matrix. Similarly, the language for Attribute 2 in Level 3 states that a response "provides elaboration on facts or ideas a few times, some of which is valid; sources of information are sometimes unclear".

This particular case is an example wherein the determination of whether or not an attribute is required for a particular level may need to be judged on the language of preceding or subsequent levels. While the description for Attribute 2 in Level 3 indicates

that, while the response does provide elaboration, not all of the elaboration is valid and the sources are somewhat unclear. On its own, the description of Attribute 2's criterion at Level 3 may not make for a definitive judgment on whether or not at least some degree of mastery is being demonstrated of "providing valid and comprehensive elaboration on relevant information". However, when compared to the description of the preceding level's criterion "any elaboration on facts or ideas tend to be vague, irrelevant, inaccurate, or unreliable" and its subsequent level's criterion "provides valid elaboration on facts or ideas several times", one could make the case that "provides elaboration on facts or ideas a few times" is *at least some* degree of mastery of the attribute. Based on this logic, and it is indicated that Attribute 2 is required to be mastered for Pseudo-Item 3 in the Writing Effectiveness (WE) pseudo-exam Q-matrix. The resulting Q-matrix for the Writing Effectiveness (WE) is shown in Table 12.

Table 12. Q-matrix Design for Writing Effectiveness (WE) Sub-Area

Pseudo-Item	Attribute 1	Attribute 2
1	1	0
2	1	0
3	1	1
4	1	1
5	1	1
6	1	1

Following the same process as the Analysis and Problem Solving (APS) and Writing Effectiveness (WE) sub-areas, the Writing Mechanics (WM) sub-area's attributes are also defined. Figure 7 shows the attributes defined to be the Writing Mechanics (WM) sub-areas three skills/attributes measured by the rubric gradient. Here, as in the Analysis and Problem Solving (APS) sub-area, three attributes have been identified and defined.

Attribute	Description
1	Demonstrating control of grammatical conventions
2	Writing well-constructed, complex, and varied sentence structure
3	Displaying an adept use of vocabulary

Figure 7. Writing mechanics (WM) sub-area definitions.

Next, the attribute structure is specified and is observed in Table 13. Again these attribute specifications will be used to determine, based on the language of the criterion for each attribute and at each level, whether or not at least some degree of mastery of each attribute must be demonstrated in order a response to qualify at each level.

Table 13. Writing Mechanics (WM) Sub-Area Attribute Specifications

Level	Attribute 1	Attribute 2	Attribute 3
1	Demonstrates minimal control of grammatical conventions with many errors that make the response difficult to read or provides insufficient evidence to judge	Writes sentences that are repetitive or incomplete, and some are difficult to understand	Uses simple vocabulary, and some vocabulary is used inaccurately or in a way that makes meaning unclear
2	Demonstrates poor control of grammatical conventions with frequent minor errors and some severe errors	Consistently writes sentences with similar structure and length, and some may be difficult to understand	Uses simple vocabulary, and some vocabulary may be used inaccurately or in a way that makes meaning unclear
3	Demonstrates fair control of grammatical conventions with frequent minor errors	Writes sentences that read naturally but tend to have similar structure and length	Uses vocabulary that communicates ideas adequately but lacks variety
4	Demonstrates good control of grammatical conventions with few errors	Writes well-constructed sentences with some varied structure and length	Uses vocabulary that clearly communicates ideas but lacks variety
5	Demonstrates very good control of grammatical conventions	Consistently writes well-constructed sentences with varied structure and length	Uses varied and sometimes advanced vocabulary that effectively communicates ideas
6	Demonstrates outstanding control of grammatical conventions	Consistently writes well-constructed complex sentences with varied structure and length	Displays adept use of vocabulary that is precise, advanced, and varied

In the case of Attribute 2, based on the language of the criterion for a Level 1 response, “writes sentences that are repetitive or incomplete, and some are difficult to understand”, and the attribute’s definition, which is “writing well-constructed, complex, and varied sentence structure”, it was not determined that at least some degree of mastery of Attribute 2 was required to be demonstrated in a Level 1 response for Writing Mechanics. Similarly, the language of the criterion for Attribute 3 at Level 1 and Level 2 also indicated that demonstration of at least some degree of mastery of that attribute was not required at those levels. Specifically, the descriptions of the criterion in both levels that a response at those levels used vocabulary in a way that “makes meaning unclear” contributed to the decision not to mark those attributes as required for those levels. However, based on the description of their respective criterion, all other combinations of attributes and levels were indicated as requiring at least some degree of mastery to be demonstrated in a response. The resulting Q-matrix for Writing Mechanics (WM) is shown in Table 14.

Table 14. Q-matrix Design for Writing Mechanics (WM) Sub-Area

Pseudo-Item	Attribute 1	Attribute 2	Attribute 3
1	1	0	0
2	1	1	0
3	1	1	1
4	1	1	1
5	1	1	1
6	1	1	1

It is important to note that the three of the resulting Q-matrices shown in Table 10, Table 12, and Table 14 all simultaneously represents a perfect Guttman scale (Guttman, 1944, 1950), linear hierarchy (Leighton, Gierl, & Hunka, 2004), and reachability matrix (Tatsuoka, 1983, 2009). Therefore the DINA model must be constrained when being fit

to the data as the number of identifiable latent classes becomes limited as a result (de la Torre, 2009). This restriction, however, was expected given the design of rubrics, as it would be theoretically impossible to identify mastery of a single attribute that is only specified within pairs in the Q-matrix.

Finally, a Q-matrix that contains all three sub-areas combined can be designed once all three sub-areas have been individually specified and coded. As a result, sub-areas can be treated as either their own independent pseudo-exam or as one combined pseudo-exam. Since each level is treated as its own independent item as well, the order in which the sub-areas are combined does not affect the model fit, item fit, or parameter estimates as long as the attribute structure is correctly specified and each item in the Q-matrix still corresponds to its pertinent pseudo-item. However, in order to simplify the interpretability of the results, sub-areas should remain grouped together. An example of such a Q-matrix is shown below in Table 15.

Table 15. Example of Q-matrix Design for All Sub-Areas Combined

Pseudo-Item	A1	A2	A3	A4	A5	A6	A7	A8
1	1	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0
3	1	1	0	0	0	0	0	0
4	1	1	1	0	0	0	0	0
5	1	1	1	0	0	0	0	0
6	1	1	1	0	0	0	0	0
7	0	0	0	1	0	0	0	0
8	0	0	0	1	0	0	0	0
9	0	0	0	1	1	0	0	0
10	0	0	0	1	1	0	0	0
11	0	0	0	1	1	0	0	0
12	0	0	0	1	1	0	0	0
13	0	0	0	0	0	1	0	0
14	0	0	0	0	0	1	1	0
15	0	0	0	0	0	1	1	1
16	0	0	0	0	0	1	1	1
17	0	0	0	0	0	1	1	1
18	0	0	0	0	0	1	1	1

In the version of the combined sub-area Q-matrix shown in Table 15, attributes one through three represent the attributes measured by the APS sub-area, attributes four and five represent the attributes measured by the WE sub-area, and attributes six through eight represent the attributes measured by the WM sub-area. While the order in which the sub-areas are combined does not affect model fit or item parameter estimates, combining the sub-areas into a single Q-matrix greatly increases the number of parameters estimated by the DINA model (de la Torre, 2009) as well as the number of possible attribute mastery profiles. For example, a DINA model, constrained for the relevant attribute hierarchies, applied to the single-sub-area Q-matrix for APS has a total of four possible attribute mastery profiles $\{0,0,0; 1,0,0; 1,1,0; 1,1,1\}$. Comparatively, the combined sub-area DINA model shown in Table 15, also constrained for the relevant attribute hierarchies, has a total of 48 possible attribute hierarchies, one for each combination of all three sub-area's possible attribute mastery profiles (four for APS, three for WE, and four for WM). The combined sub-area design therefore significantly affects both model and item fit negatively as it increases the number of parameters estimated by the model and affects item and attribute discrimination positively as it increases the number of possible attribute mastery profiles.

It should be noted at this point that the specification of the Q-matrix is an integral component of DCM analysis, and that the misspecification of the Q-matrix can be detrimental to the validity and interpretability of the results, especially in the presence of attribute hierarchies (Chen, Liu, Xu, & Ying, 2015; Liu & Huggins-Manley, 2016; Groß & George, 2014; Liu, Huggins-Manley, & Bradshaw, 2016; Madison & Bradshaw, 2015; Rupp & Templin, 2007; Templin & Bradshaw, 2014). Therefore in practice subjective

inference alone should not be the only method used in Q-matrix construction. Various other methods such as the consultation of subject-matter experts (Choi, Lee, & Park, 2014; Lee, Park, & Taylan, 2011) as well as statistical methods for Q-matrix validation and design (Chiu, 2013; Cui, Gierl, & Chang, 2012; de la Torre, 2008; Xu & Zhang, 2016) should also be employed.

3.3.3 An Automatic M -attribute Method (Full-Score Coding)

Similar to the technique proposed for converting rubric scores into dichotomous pseudo-items, the proposed automatic M -attribute (aMa) method is an adaptation of the sequential response mechanism (Tutz, 1997) that offers a method for creating Q-matrices for rubric sub-areas. In the automatic M -attribute method, a rubric r is represented as having M scoring levels that range from $\{1, \dots, m_r\}$, where 1 represents the lowest score level in the rubric and m_r represents the highest score level. The aMa method then assumes that each sub-area s in the rubric r , is measured by the same number of levels, and can be represented by a $L \times C$ Q-matrix (Q_s) of m_r length and m_r width, where L represents each level in the rubric and C represents which level's criteria must be met by the response in order for the response to be classified as level l . Each level l in the sub-area Q-matrix corresponds to a level in the rubric, and has a corresponding q -vector that represents the criteria for which level that have been met by the response. If level l requires that criteria for $\{1, \dots, l\}$ levels be met, then $q_{l\{1, \dots, c\}} = 1$, otherwise $q_{lc} = 0$.

As a result, a perfect Guttman scale (Guttman, 1944, 1950) is formed. An example of a sub-area Q-matrix for a rubric with eight levels is shown in Table 15. The theoretical rubric sub-area Q-matrix as conceived by the aMa method can be expressed in the following equation:

$$Q_s = \begin{pmatrix} q_{11} & \cdots & q_{1c} \\ \vdots & \ddots & \vdots \\ q_{l1} & \cdots & q_{lc} \end{pmatrix} \text{ where } c = \{1 \dots m_r\} \text{ and } l = \{1 \dots m_r\}.$$

Table 16. Example of a Sub-Area Q-matrix for a Rubric with Seven Levels, Designed using the Automatic *M*-Method

Level	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
1	1	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0
3	1	1	1	0	0	0	0	0
4	1	1	1	1	0	0	0	0
5	1	1	1	1	1	0	0	0
6	1	1	1	1	1	1	0	0
7	1	1	1	1	1	1	1	0
8	1	1	1	1	1	1	1	1

The CLA+ scoring rubric scores all three sub-areas according to six levels. As a result, the Q-matrix designed using the automatic *M*-method, shown in Table 16, can be applied to all three sub-areas.

Table 17. CLA+ Rubric Sub-Area Q-matrix Designed using the Automatic *M*-Method

Level	c_1	c_2	c_3	c_4	c_5	c_6
1	1	0	0	0	0	0
2	1	1	0	0	0	0
3	1	1	1	0	0	0
4	1	1	1	1	0	0
5	1	1	1	1	1	0
6	1	1	1	1	1	1

Finally, as was shown in the previous section, all three sub-areas can be combined to form a $M \times M$ Q-matrix. Again, the effect of such a design greatly increases both the number of parameters estimated by the model and the number of possible attribute mastery profiles. Also, as long as the attribute hierarchy is specified correctly, the order in which the sub-areas are combined will not affect the model fit, item fit, or parameter estimates, although sub-areas should remain grouped together for ease of interpretability.

3.4 Software for Fitting the Proposed Models

The computer program R (R Core Team, 2016) was the tool used to conduct the following analysis. Specifically the study will be using the R software package ‘CDM’ (Robitzsch, Kiefer, George, & Uenlue, 2017) to apply various constrained-DINA models (de la Torre, 2009) to the converted constructed-response data. Parameter estimation for the DINA model followed the steps outlined in the EM algorithm presented by de la Torre (2009). In order to fit the G-PCM to the polytomous constructed-response data, the R software package ‘mirt’ (Chalmers, 2012) is used. G-PCM parameters were estimated using the standard EM algorithm described by Bock & Aitken (1981).

3.5 Diagnostic Classification Model Analysis

This section describes the analyses used to assess the methods for retrofitting DCMs to rubric-scored constructed-response data. First, an overview of the conditions for DINA

analysis are presented, followed by the specific model fit statistics used, with guidelines for their interpretation. Next, a description of DCM framework specific item fit statistics are shown, also including guidelines for their interpretation. Finally, a brief description of attribute tetrachoric correlation and information-based item discrimination indices for DCMs and their significance in DCM analysis is discussed.

3.5.1 Conditions for DINA Model Analysis

The DINA model was fit to the data in eighteen conditions between three factors: Q-matrix coding method, single sub-area or combined sub-area Q-matrix design, and guessing and slipping parameter constraints. The Q-matrix coding factor contains two conditions, in which rubrics are designed using either the rubric based method (rubric coding) or the automatic M-method (full-score coding). Likewise, the single sub-area or combined sub-area Q-matrix design factor also involves two conditions, one in which each sub-area is treated as its own pseudo-exam and thus has its own Q-matrix and one in which all sub-areas are combined and thus one Q-matrix is used to represent all three sub-areas simultaneously. Lastly, the guessing and slipping parameter constraints factor consists of five conditions in which both the guessing and slipping parameters are constrained to zero, only the guessing parameters are constrained to zero, only the slipping parameters are constrained to zero, the guessing parameter is constrained to be less than the slipping parameter, and no constraints are placed on either parameters. Table 18 exhibits the intersection of the three factors that create the eighteen conditions explored in this study.

Table 18. Conditions in DINA Model Analysis

Parameter Constraints	Q-matrix Design			
	Rubric-Coded		Full-Score Coded	
	Single Sub-Area	Combined Sub-Area	Single Sub-Area	Combined Sub-Area
Guess and Slip to Zero	-	-	Condition 1	Condition 2
Guess to Zero	Condition 3	Condition 4	Condition 5	Condition 6
Slip to Zero	Condition 7	Condition 8	Condition 9	Condition 10
Guess < Slip	Condition 11	Condition 12	Condition 13	Condition 14
No Constraints	Condition 15	Condition 16	Condition 17	Condition 18

The cells for the rubric-coded Q-matrix designs are left blank for the ‘Guess and Slip to Zero’ parameter constraints condition because the DINA model will not converge if both the guess and slip parameters are set to zero, unless the latent response vector is identical to the manifest or observed variable (de la Torre, 2009). This is due to the first two steps in the EM algorithm that is used for DINA model parameter estimation. The first step of the algorithm begins with setting initial values for the guessing parameter g and slip parameter s . The next step in the algorithm involves solving the two equations shown below:

$$\hat{g}_j = \frac{R_{jl}^{(0)}}{I_{jl}^{(0)}}$$

$$\hat{s}_j = [I_{jl}^{(1)} - R_{jl}^{(1)}]I_{jl}^{(1)}$$

where $I_{jl}^{(0)}$ represents the expected number of examinees lacking at least one attribute for item j , $R_{jl}^{(0)}$ represents the expected number of examinees among $I_{jl}^{(0)}$ that correctly answered item j , $I_{jl}^{(1)}$ represents the expected number of examinees that have mastered all the required attributes for item j , and $R_{jl}^{(1)}$ represents the expected number of examinees among $I_{jl}^{(1)}$ that correctly answered item j , for all attribute mastery states l

(de la Torre, 2009). Based on the initial values for g and s , $I_{jl}^{(0)}$, $R_{jl}^{(0)}$, $I_{jl}^{(1)}$, and $R_{jl}^{(1)}$ are calculated.

Interpretively, the second step of the algorithm stipulates that in order for both the guess and slip parameters to be zero, two states must be met simultaneously. The first state is that none of the examinees that lacked at least one of the required attributes for item j answered item j correctly. The second state is that all of the examinees that mastered all of the required attributes for item j also answered item j correctly. Therefore, both conditions can only exist concurrently if the attribute mastery pattern is identical to the observed item response pattern, since the attribute mastery pattern is merely a reflection of the item response pattern, and so this is the only condition in which both states are a certainty. As a result, only the full-score coded Q-matrix designs can be fit to this condition. Furthermore, both the ‘Guess and Slip to Zero’ and ‘Guess < Slip’ parameter constraint conditions represents forced conditions of monotonicity $g < 1 - s$, which is the assumption that the probability of answering an item correctly without mastering its required attributes is less than the probability of answering an item correctly, having mastered its required attributes (Rupp, Templin, & Henson, 2010).

In the ‘Guess < Slip’ condition, the guess parameters have been constrained to a minimum value of zero and a maximum value of 0.20 while the slip parameters have been constrained to a minimum of 0.20 and a maximum of one. These values were chosen based on the simulation study done by de la Torre (2009). Interpretively, in this condition there is a maximum of a 20% chance that a level may be scored without a respondent having actually mastered the required attributes and a minimum of a 20% chance that a candidate may not have been scored at a particular level although they have

mastered the respective attributes. Initial values for the EM algorithm were then set as 0.10 and 0.30 for the guess and slip parameters, respectively.

3.5.2 Model Fit Statistics

In order to determine the quality of model fit, the study examined the absolute model fit of the data by estimating model fit indices for each test subsection and comparing those estimates. Traditional IRT relative model fit indices such as the corrected Akaike information criterion (AICc; Akaike 1974) and Bayesian information criterion (BIC) (Schwarz, 1976) were used to compare model fit between test subsections. Absolute model fit indices were also used in order to objectively examine the quality of the model fit to the data:

1. Mean Absolute Difference for the Item-Pair Correlations (MADcor)

- The average absolute difference between the calculated observed r_{ij} and model-predicted \hat{r}_{ij} item correlations for item pairs (i, j) (DiBello, Roussos, & Stout, 2007):

$$MADcor = \frac{1}{J(J-1)/2} \sum_{i < j} |r_{ij} - \hat{r}_{ij}|$$

2. Standardized Root Mean Square Residual (SRMSR)

- Designed by Maydeau-Olivares (2013) to estimate the approximate fit of large models for ordinal data while being unaffected by the number of items if all other factors are held constant. It is recommended that a cut-off

of SRMR values ≤ 0.05 be used as an indication of good model fit (Maydeu-Olivares, 2013). Like the MADcor statistic, the SRMSR is also based on comparing the observed and predicted pairwise correlation between item pairs (i, j) :

$$SRMSR = \sqrt{\frac{1}{J(J-1)/2} \sum_{i < j} (r_{ij} - \hat{r}_{ij})^2}$$

3. Q3 Statistic (MADQ3)

- The Q3 statistic represents the mean absolute values of pairwise correlations of the difference between the observed and model-predicted responses for each examinee (Yen, 1984). In order to compute the value of $Q3_{ij}$ for item pairs (i, j) , residuals of observed and expected responses for examinees n and item i (i.e. $\varepsilon_{ni} = X_{ni} - e_{ni}$) and item j (i.e. $\varepsilon_{nj} = X_{nj} - e_{nj}$) are first calculated. Next, the correlation is found between residuals ε_{ni} and ε_{nj} and over examinees n so that $Q3_{ij} = r_{\varepsilon_{ni}\varepsilon_{nj}}$ where $r_{\varepsilon_{ni}\varepsilon_{nj}} = \frac{cov(\varepsilon_{ni}, \varepsilon_{nj})}{s_{\varepsilon_{ni}} s_{\varepsilon_{nj}}}$. MADQ3 is then estimated by averaging the absolute value of the calculated Q3 values for all item pairs. MADQ3 values below 0.05 indicate good model fit.

Previous studies have shown that such indices can be useful in model as well as Q-matrix selection in DCM settings (Galeshi & Skaggs, 2014; Hu, Miller, Huggins-Manley, & Chen, 2016; Lei & Li, 2016). In the case of the MADcor and SRMSR, the absolute difference between observed and predicted correlations r_{ij} is based on the formula from Chen, de la Torre, & Zhang (2013):

$$r_{ij} = |Z[\text{Corr}(X_i, X_j)] - Z[\text{Corr}(\tilde{X}_i, \tilde{X}_j)]|$$

where $Z[\cdot]$ is the fisher transformation, X_i and \tilde{X}_i are the observed and predicted response vectors for item i , respectively, and X_j and \tilde{X}_j are the observed and predicted response vectors for item j , respectively.

3.5.3 Item Fit/Parameter Estimates

The root mean square error of approximation (RMSEA; Kunina-Habenicht, Rupp, & Wilhelm, 2009) is used as an item fit statistic and indicates how well an item harmonizes with the model. As a rule of thumb, it is recommended that items with fit indices below .05 indicate good model fit, while items below .10 indicate moderate fit, and items with indices greater than .10 indicate poor fit. In the DINA model the additional constraint of $g_j < 1 - s_j$ dictates that the probability of answering an item without having mastered the required skills is less than the probability of answering an item correctly when the required skills have been mastered. This constraint can be checked using the item discrimination index (IDI) where $IDI_j = 1 - s_j - g_j$ (Lee, de la Torre, & Park, 2012).

A negative value for IDI indicates a violation of this restraint. In this way, the IDI can be viewed as a diagnostic index, indicating how well that item discriminates between respondents that possess all the required skills (i.e. a response probability of $1 - s_j$) and respondents that do not possess all the required skills (i.e. a response probability of g_j). As a rule of thumb, an IDI value that is close to 1 indicates good item discrimination or

diagnosticity and an IDI value that is close to 0 indicates low item discrimination or diagnosticity (Kunina-Habenicht, Rupp, & Wilhelm, 2009).

3.5.4 Attribute Tetrachoric Correlation

The calculation of tetrachoric correlation between the estimated attributes is based on two assumptions. The first assumption is that latent continuous variables underlie the latent dichotomous skill variables. The second assumption is that the correlation between two skill variables is equal to the correlation between the two underlying continuous variables (Templin & Henson, 2006; Templin, Henson, Templin, & Roussos, 2008). In other words, the tetrachoric correlation between skills posits to represent the correlation between two attributes in the Q-matrix. Strong positive tetrachoric correlation relationships may be considered evidence of unidimensionality or attribute hierarchy (Templin & Bradshaw, 2014).

Past research has found that setting tetrachoric correlations between attributes to fixed values of 0.70 is reasonable in an educational context as values at or below 0.70 appear to indicate that the constructs being measured may still be multidimensional (Bradshaw & Templin, 2014; Cui, Gierl, & Chang, 2012; Liu, Huggins-Manley, & Bradshaw, 2016; Sinharay, Puhon, & Haberman, 2011). However, if it is assumed that attributes are part of a hierarchical structure, tetrachoric correlations should not be low and in fact, may result in an increased number of iterations for model convergence or non-convergence (Liu, Huggins-Manley, & Bradshaw, 2016). Therefore we expect that the tetrachoric correlations between attributes in this study will not be low and in fact, as

we assume the presence of an attribute hierarchy for all attributes, will have strong positive relationships.

3.5.5 Information-Based Item Discrimination Indices for DCMs

Another approach to determining the discriminatory ability of an item other than item-fit and model parameter estimates is the IRT concept of *statistical information*. Statistical information refers to the amount of information that an item can provide regarding different values of latent variables (Rupp, Templin, & Henson, 2010). In IRT frameworks, statistical information is calculated using the *Fisher information* algorithm, and provides evidence for locations on a latent trait continuum where an item provides the most information. The amount of information an item provides is determined by the item-discrimination parameter, with higher estimates of item information indicating higher estimates of discrimination (Embretson & Reise, 2000). However, the Fisher information algorithm is defined for models that measure continuous latent variables, and cannot be applied to DCMs, which measure discrete latent variables.

Therefore, in order to devise an information statistic that can be used in the DCM measurement framework, researchers have suggested using the *Kullback-Leibler information* (KLI) statistic (Kullback & Leibler, 1951) as an alternative for estimating the ability for DCM items to discriminate between attribute masters and non-masters (Henson & Douglas, 2005; Henson, Roussos, Douglas, & He, 2008). The KLI measures the amount of difference between a target distribution $f(X)$ and a reference distribution $g(X)$, by computing the expected value of the natural logarithm of the ratio between the

density of the target distribution $f(X)$ and the reference distribution $g(X)$ (Rupp, Templin, & Henson, 2010):

$$KLI(f(X), g(X)) = E_f \left[\ln \left[\frac{f(X)}{g(X)} \right] \right].$$

As a result, if $f(X)$ and $g(X)$ are equal, $KLI = 0$ and KLI does not require that either function be continuous, thus allowing them to be applied to the DCM framework.

In order to apply the KLI to the DCM measurement framework, the target distribution $f(X)$ and the reference distribution $g(X)$ are defined as the *conditional distribution* $f(X | \alpha_u)$ and $g(X | \alpha_v)$ where α_u and α_v represent two attribute profiles and X represents an observed item response pattern. The objective of this restructuring is to set up the KLI equation to compare the expected item response pattern on a DCA between respondents with attribute profiles α_u and α_v . Thus, the KLI equation can be rewritten as follows: $KLI(f(X | \alpha_u), g(X | \alpha_v)) = E_f \left[\ln \left[\frac{f(X | \alpha_u)}{g(X | \alpha_v)} \right] \right]$. As a result, in the context of dichotomously scored items the KLI is the sum of the two distinct outcomes for the item ($X_i = 0$ and $X_i = 1$) (Henson, Roussos, Douglas & He, 2008):

$$KLI(f(X_i | \alpha_u), g(X_i | \alpha_v)) = P(X_i = 1 | \alpha_u) \ln \left[\frac{P(X_i=1|\alpha_u)}{P(X_i=1|\alpha_v)} \right] + P(X_i = 0 | \alpha_u) \ln \left[\frac{P(X_i=0|\alpha_u)}{P(X_i=0|\alpha_v)} \right].$$

Interpretively, the magnitudes of the KLI values represent the unique diagnostic power of each DCA for determining skill mastery classification. In order to compute the overall discriminatory power of a DCA, all possible comparisons of KLI values between attribute profiles must be represented within a matrix D by computing a D_i matrix for each item and summing them together (Henson, Roussos, Douglas & He, 2008):

$$D = \sum_{i=1}^I D_i$$

where D is a $2^K \times 2^K$ matrix for K attributes, representing $2^K(2^K - 1)$ possible comparisons of KLI values between attributes profiles and

$$D_i = E_f \left[\ln \left[\frac{P(x_i|\alpha_u)}{P(x_i|\alpha_v)} \right] \right].$$

By being computed in this way, the KLI statistic operates much like the Fisher information statistic how it identifies the locations on a latent trait continuum where psychometric information is the highest or lowest for an item by identifying which items contribute the most towards attribute mastery discrimination. However, a major limitation of this method is that matrix D_i expands exponentially as the number of attributes increases.

As a solution, Henson and Douglas (2005) propose simplifying the information in each D_i by defining an index called the *Cognitive Diagnostic Index* (CDI), which calculates the average of all the values in D_i while weighting each comparison between different attribute profiles by the number of attributes by which the two profiles do not share. The CDI is therefore defined by the following equation:

$$C_i = \frac{\sum_{u \neq v} h(\alpha_u, \alpha_v) D_{i,uv}}{\sum_{u \neq v} h(\alpha_u, \alpha_v)^{-1}}$$

where $D_{i,uv}$ is the entry in D_i associated with attribute profiles α_u and α_v and

$$h(\alpha_u, \alpha_v) = \sum_{k=1}^K |\alpha_{ua} - \alpha_{va}|.$$

Interpretively, $h(\alpha_u, \alpha_v)$ represents the frequency of the number of attributes that are not shared between two attribute profiles. As a result, the final discriminatory power of a diagnostic assessment can be computer as:

$$C = \sum_{i=1}^I C_i.$$

DCAs with higher estimates of C are expected to be able to correctly classify masters and non-masters of individual attributes more frequently than DCAs with lower estimates of C (Henson & Douglas, 2005). Similarly, items with higher C_i values are expected to be more useful in identifying masters and non-masters of the attributes measured by the DCA than items with lower C_i values. Furthermore, in addition to global item discrimination the KLI can also estimate attribute-specific item discrimination as well.

The objective of attribute-specific item discrimination is to identify which items in a DCA provide the most diagnostic information in determining an attribute's mastery or non-mastery (Kuo, Pai, de la Torre, 2016). The calculation of this index is similar to that of C_i in that it is based on elements of D_i except in this case the only values that are relevant are those cells that contain KLI values related to differences in attribute mastery profiles wherein the targeted attribute is included. In addition, in calculating D_i in the attribute-specific case for item discrimination, only attribute patterns of mastery that only differ by one attribute will be used since attribute patterns that differ by only one component are the most difficult to discriminate (Henson, Roussos, Douglas, & He, 2008). Consequently the attribute-specific item discrimination index is defined as:

$$C_{ik} = \frac{1}{2^k} \sum_{all\ relevant\ cells} D_{i,uv}.$$

3.6 Generalized Partial Credit Model (G-PCM) Analysis

Researchers have noted that the practical justification for employing parametrically complex DCMs is limited, and is still an ongoing debate (Gorin, 2009; Haberman, von

Davier, & Lee, 2008; Rupp & Templin, 2009; Sinharay, Puhon, & Haberman, 2011; von Davier, 2009; von Davier & Haberman, 2014). This is especially true when these methods are compared to simpler and more established IRT models. Recent research has also shown that a combination of DCM and IRT analysis may be beneficial for improving the diagnostic ability of assessments as well (Bradshaw & Templin, 2014). Therefore the polytomous IRT generalized partial credit model (G-PCM) will also be applied to the data in order to analyze the performance of the rubric-scored data as a comparison for the efficacy and practicality of the DCM method.

4 Empirical Results/Application to the CLA+

The results from the selected-response section and constructed-response section are presented in this chapter. First, the results of the constructed-response section will be presented, starting with an examination of the two proposed techniques for representing each rubric sub-area as a Q-matrix for the DCM framework, followed by an analysis of polytomous IRT models applied to the constructed response data with each sub-area represented as an item.

4.1 Descriptive Statistics

The descriptive statistics shown in Table 19 and Table 20 both show that a majority of test scores in both samples are either a score of two, three, or four, for all sub-areas.

Moreover, it appears that at least 90% of the data in all three sub-areas for both samples. This disproportionate distribution of scores also appears to be weighted towards particular score levels in some sub-areas. For example, in Test Form B approximately 49% of the scores for Writing Mechanics (WM) are a score of four, compared to only 0.86% for level five.

Table 19. Test Form A Sub-Area Score Proportions

Sub-Area	0	1	2	3	4	5	6
APS	0.30%	3.31%	38.28%	24.85%	31.56%	1.40%	0.30%
WE	0.30%	4.21%	46.79%	19.24%	27.05%	2.20%	0.20%
WM	0.30%	1.30%	29.66%	23.85%	41.88%	2.40%	0.60%

N = 988

Table 20. Test Form B Sub-Area Score Proportions

Sub-Area	0	1	2	3	4	5	6
APS	1.30%	5.18%	33.91%	28.40%	29.16%	1.84%	0.22%
WE	1.30%	4.75%	31.32%	27.97%	32.72%	1.73%	0.22%
WM	1.30%	1.30%	15.44%	31.75%	49.14%	0.86%	0.60%

N = 926

Similarly, the proportion of scores in both samples for all sub-areas for level five are on average 1.74%, compared to level four for all sub-areas, which is 35.25%.

In Table 21 we can see that Cronbach's alpha for the exam overall was high for both samples at 0.93 for Form A and 0.92 for Form B, indicating strong test reliability. However, reliability appeared to drop noticeably with the exclusion of APS and WE, and appeared to be largely unaffected with the exclusion of WM. Furthermore, the exclusion of WM appears to approve the overall reliability of test Form B.

Table 21. Cronbach's Alpha for Overall Test Reliability

	Form A	Form B
All Items	0.93	0.92
Excluding APS	0.87	0.88
Excluding WE	0.88	0.86

Excluding WM

0.93

0.94

On the whole, the different sub-areas, being treated as separate polytomous items in this case, appear to have strong correlations, which would suggest that the IRT assumption of local independence has been violated. However, the two components to the assumption of local independence are that only one latent trait is being measured by the assessment and that the response to one item is not contingent upon the response to another question (Yang & Kao, 2014). By the very nature of the rubric, the items (which are in fact sub-areas of the rubric) are measuring separate components of the same overarching construct (which is the examinee's overall performance on the constructed-response). Similarly, each sub-area of the rubric is scored based on that sub-area's criterion alone and not the criterion or the examinee's score in the other sub-areas. Therefore it would appear that the assumption of local independence is not violated despite strong item correlations, due to the operational parameters of the rubric.

Table 22. Form A: Item (Sub-Area) Correlations

	APS	WE	WM
APS	-	0.86	0.78
WE	<0.001	-	0.77
WM	<0.001	<0.001	-

N = 998

*upper diagonal contains correlation coefficient estimates

*lower diagonal contains corresponding p-values

Table 23. Form B: Item (Sub-Area) Correlations

	APS	WE	WM
APS	-	0.88	0.73
WE	<0.001	-	0.76
WM	<0.001	<0.001	-

N = 926

*upper diagonal contains correlation coefficient estimates

*lower diagonal contains corresponding p-values

4.2 Constructed-Response Section Retrofit Results

In this section the results from retrofitting a constrained DINA model to the rubric-scored constructed-response data are presented, beginning with an overview of the model fit. In each subsection, the results for both proposed methods for designing Q-matrices for rubric-scored items are shown separately for both test forms A and B. Next, item fit statistics are shown, displaying the behavior of typical DCM parameters when constrained to the parameters of rubric-scored data. Tetrachoric correlations between attributes are then shown next, which may lend insight into the dimensionality of each rubric sub-area. Finally, skill classification estimates are presented, providing an illustration of the amount of diagnostic information that can be yielded from the proposed methods.

4.2.1 Model Fit

Observing the model fit statistics for the Rubric Coded Q-matrix designs in Table 24 and Table 25, the relative fit indices indicate little difference between the sub-areas in both Forms A and B, although the WM sub-area in Form B appears to have better model fit than the other sections. Furthermore, the combined sub-area Q-matrix designs all appear to have substantially poorer model fit than the single sub-area Q-matrix designs.

Similarly, the No Constraints condition appears to have the best model fit universally in both samples and for both single sub-area and combined sub-area Q-matrix designs. The absolute fit statistics for rubric coded Q-matrices in both Form A and Form B indicate poor model fit as the mean absolute differences between the mean absolute Q3 values (MADQ3) and standardized root mean square residual (SRMSR) values for all three sub-areas are over 0.05. Furthermore, the mean absolute difference of the item-pair correlations (MADcor) absolute fit statistics for the single sub-area Q-matrix designs, with the exception of WM in Form B, indicate good model fit, as they are less than 0.05. However, overall the absolute model fit statistics do not suggest that the DINA model fits the rubric scored data parsimoniously, although some conditions and sub-areas may be more so than others.

Table 24. Form A Rubric Coded Q-matrix Design Model Fit Statistics

Parameter Constraints	Q-matrix Design	AICc	BIC	MADcor	MADQ3	SRMSR
Guess to Zero	Single APS	3049.844	3093.813	0.075	0.069	0.182
	Single WE	2653.208	2692.308	0.055	0.050	0.102
	Single WM	2600.667	2644.636	0.041	0.051	0.118
	Combined	7193.869	7503.537	0.100	0.115	0.204
Slip to Zero	Single APS	3104.224	3148.193	0.088	0.077	0.179
	Single WE	3230.126	3269.226	0.101	-	0.196
	Single WM	3095.642	3139.611	0.088	-	0.175
	Combined	9246.517	9556.185	0.146	-	0.263
Guess < Slip	Single APS	3829.383	3902.480	0.135	0.126	0.214
	Single WE	3421.098	3489.351	0.089	0.083	0.127
	Single WM	3611.259	3684.356	0.119	0.100	0.182
	Combined	9659.979	10051.900	0.161	0.115	0.266
No Constraints	Single APS	2628.977	2702.075	0.045	0.083	0.125
	Single WE	2610.387	2678.640	0.028	0.080	0.076
	Single WM	2600.081	2673.179	0.040	0.084	0.115
	Combined	6192.776	6584.698	0.085	0.106	0.196

Table 25. Form B Rubric Coded Q-Matrix Design Model Fit Statistics

Parameter Constraints	Q-matrix Design	AICc	BIC	MADcor	MADQ3	SRMSR
Guess to Zero	Single APS	2924.234	2967.515	0.063	0.062	0.154
	Single WE	2681.694	2720.184	0.063	0.059	0.120
	Single WM	2210.496	2253.777	0.037	0.050	0.117
	Combined	6889.442	7193.472	0.099	0.128	0.208
Slip to Zero	Single APS	3044.144	3087.426	0.103	0.086	0.190
	Single WE	3100.020	3138.510	0.108	0.093	0.187
	Single WM	2548.680	2591.961	0.099	0.076	0.173
	Combined	8401.290	8705.320	0.151	0.178	0.259
Guess < Slip	Single APS	3626.853	3698.788	0.152	0.138	0.223
	Single WE	3392.789	3459.960	0.133	0.108	0.190
	Single WM	3185.765	3257.701	0.165	0.136	0.255
	Combined	9352.006	9736.408	0.182	0.126	0.289
No Constraints	Single APS	2643.483	2715.419	0.047	0.080	0.128
	Single WE	2601.634	2668.805	0.034	0.082	0.092
	Single WM	2261.390	2333.326	0.060	0.088	0.141
	Combined	5973.937	6358.339	0.096	0.120	0.209

In cases where the value for MADQ3 is not shown, the pseudo-exam contained item pairs (which are in this case represent levels), where either one item had no difference between respondent observed and expected score residuals or both items had no difference between respondent observed and expected score residuals. If for an item there is zero difference between respondents in the difference between the observed and expected scores, the average difference is zero, and thus the estimate for MADQ3 becomes infinity once the correlation is calculated, as zero will divide the covariance of the residuals (i.e. $r_{\varepsilon_{ni}\varepsilon_{nj}} = \frac{cov(\varepsilon_{ni}, \varepsilon_{nj})}{s_{\varepsilon_{ni}}s_{\varepsilon_{nj}}} = \infty$ when $s_{\varepsilon_{ni}}s_{\varepsilon_{nj}} = 0$). This is what occurs in the Slip to Zero and No Constraints conditions in Form A. In the Guess and Slip to Zero condition, both the covariance of the item pair residuals $cov(\varepsilon_{ni}, \varepsilon_{nj})$ and standard deviations $s_{\varepsilon_{ni}}s_{\varepsilon_{nj}}$ are equal to zero, producing non-real numbers.

Table 26. Form A Full-Score Coded Q-matrix Design Model Fit Statistics

Parameter Constraints	Q-matrix Design	AICc	BIC	MADcor	MADQ3	SRMSR
Guess and Slip to Zero	Single APS	2577.039	2606.389	0.000	-	0.000
	Single WE	2553.936	2583.286	0.000	-	0.000
	Single WM	2529.484	2558.834	0.000	-	0.000
	Combined	6312.945	7632.526	0.000	-	0.000
Guess to Zero	Single APS	2589.552	2648.104	0.000	0.406	0.001
	Single WE	2566.218	2624.771	0.000	0.372	0.001
	Single WM	2541.982	2600.534	0.000	0.497	0.000
	Combined	6398.819	7756.852	0.000	0.294	0.000
Slip to Zero	Single APS	2589.348	2647.900	0.000	-	0.001
	Single WE	2566.261	2624.814	0.000	-	0.001
	Single WM	2542.023	2600.576	0.000	0.054	0.001
	Combined	6398.985	7757.018	0.000	-	0.000
Guess < Slip	Single APS	3491.666	3579.271	0.081	0.061	0.118
	Single WE	3418.419	3506.023	0.075	0.056	0.109
	Single WM	3571.908	3659.513	0.094	0.073	0.146
	Combined	9622.647	11014.140	0.108	0.073	0.199
No Constraints	Single APS	2601.755	2689.360	0.000	0.041	0.001
	Single WE	2578.652	2666.256	0.000	0.044	0.001
	Single WM	2554.390	2641.995	0.000	0.043	0.001
	Combined	6489.662	7881.154	0.000	-	0.000

Table 27. Form B Full-Score Coded Q-matrix Design Model Fit Statistics

Parameter Constraints	Q-matrix Design	AICc	BIC	MADcor	MADQ3	SRMSR
Guess and Slip to Zero	Single APS	2567.776	2596.670	0.000	-	0.000
	Single WE	2549.239	2578.133	0.000	-	0.000
	Single WM	2176.762	2205.656	0.000	-	0.000
	Combined	5969.191	7218.928	0.000	-	0.000
Guess to Zero	Single APS	2580.089	2637.717	0.000	0.486	0.001
	Single WE	2561.565	2619.194	0.000	0.226	0.001
	Single WM	2189.420	2247.049	0.001	0.008	0.001
	Combined	6062.962	7342.041	0.000	0.297	0.000
Slip to Zero	Single APS	2580.122	2637.751	0.000	0.037	0.000
	Single WE	2561.594	2619.222	0.000	0.039	0.000
	Single WM	2189.125	2246.754	0.000	0.022	0.000
	Combined	6062.978	7342.058	0.000	0.085	0.000
Guess < Slip	Single APS	3367.486	3453.688	0.108	0.084	0.163
	Single WE	3376.587	3462.788	0.111	0.087	0.169
	Single WM	3166.286	3252.487	0.134	0.101	0.223
	Combined	9087.607	10389.868	0.131	0.090	0.226

	Single APS	2592.488	2678.689	0.000	0.038	0.001
No	Single WE	2573.960	2660.162	0.000	0.039	0.001
Constraints	Single WM	2201.694	2287.895	0.000	0.140	0.001
	Combined	6162.648	7464.909	0.000	0.123	0.000

The MADcor and SRMSR absolute model fit statistics for the full-score coded q-matrices in shown in Table 26 and Table 27 both indicate that, with the exception of the Guess < Slip parameter constraints condition, the data has fit well with the model. In almost all conditions where the slip parameter was left unconstrained the MADQ3 statistic values appear to rise significantly, especially when the guess parameter is also constrained in some way, either to zero or to less than the slip parameter. This may be evidence that the inclusion of the slip parameter has a detrimental effect to model fit in cases where the Q-matrix represents a full-score coded rubric structure. In both samples the Guess < Slip parameter constraint condition has the highest absolute model fit estimates within the sample, indicating that these parameter constraints fit the full-score coded Q-matrix data the least parsimoniously. Contrarily, both the absolute and relative model fit statistics in both samples indicate that the Guess and Slip to Zero parameter constraint conditions have the lowest model fit statistics for both the single sub-area and combined Q-matrix designs.

4.2.2 Item Parameter Estimates

For space, the results presented will now focus on only the No Constraints parameter constraint conditions for rubric-coded and full-score coded Q-matrices. For example,

Table 28 and Table 29 both compare the item parameter estimates for rubric-coded Q-matrices in the no constraint condition.

Table 28. Form A Rubric-Coded Q-matrix Design Item Parameter Estimates without Guess or Slip Parameter Constraints

Item	Single Sub-Area Q-matrix				Combined Sub-Area Q-matrix			
	Guess	Slip	RMSEA	IDI	Guess	Slip	RMSEA	IDI
APS1	0.965	0.000	0.000	0.035	0.000	0.000	0.000	1.000
APS2	0.914	0.000	0.012	0.086	0.914	0.000	0.194	0.086
APS3	0.002	0.000	0.000	0.998	0.000	0.000	0.000	1.000
APS4	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000
APS5	0.000	0.949	0.000	0.051	0.000	0.949	0.006	0.051
APS6	0.000	0.991	0.000	0.009	0.000	0.991	0.001	0.009
WE1	0.934	0.000	0.000	0.066	0.933	0.000	0.075	0.067
WE2	0.014	0.000	0.000	0.986	0.000	0.000	0.000	1.000
WE3	0.272	0.000	0.084	0.728	0.069	0.000	0.068	0.931
WE4	0.000	0.003	0.000	0.997	0.000	0.344	0.258	0.656
WE5	0.000	0.919	0.000	0.081	0.000	0.946	0.031	0.054
WE6	0.000	0.993	0.000	0.007	0.000	0.996	0.003	0.004
WM1	0.465	0.000	0.000	0.535	0.000	0.000	0.000	1.000
WM2	0.004	0.000	0.000	0.996	0.004	0.000	0.000	0.996
WM3	0.433	0.000	0.078	0.567	0.389	0.000	0.207	0.611
WM4	0.000	0.000	0.000	1.000	0.000	0.080	0.040	0.920
WM5	0.000	0.933	0.000	0.067	0.000	0.938	0.043	0.062
WM6	0.000	0.987	0.000	0.013	0.000	0.988	0.009	0.012

Table 29. Form B Rubric Coded Q-matrix Design Item Parameter Estimates without Guess or Slip Parameter Constraints

Item	Single Sub-Area Q-matrix				Combined Sub-Area Q-matrix			
	Guess	Slip	RMSEA	IDI	Guess	Slip	RMSEA	IDI
APS1	0.857	0.000	0.000	0.143	0.000	0.000	0.000	1.000
APS2	0.840	0.000	0.047	0.160	0.840	0.000	0.256	0.160
APS3	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000
APS4	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000
APS5	0.000	0.934	0.000	0.066	0.000	0.934	0.005	0.066
APS6	0.000	0.993	0.000	0.007	0.000	0.993	0.001	0.007
WE1	0.786	0.000	0.000	0.214	0.786	0.000	0.143	0.214
WE2	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000
WE3	0.001	0.000	0.001	0.999	0.167	0.000	0.146	0.833
WE4	0.000	0.446	0.000	0.554	0.000	0.372	0.317	0.628
WE5	0.000	0.969	0.000	0.031	0.000	0.965	0.033	0.035
WE6	0.000	0.997	0.000	0.003	0.000	0.996	0.004	0.004

WM1	0.165	0.000	0.000	0.835	0.000	0.000	0.000	1.000
WM2	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000
WM3	0.628	0.000	0.147	0.372	0.578	0.000	0.226	0.422
WM4	0.000	0.026	0.001	0.974	0.000	0.124	0.065	0.876
WM5	0.000	0.979	0.000	0.021	0.000	0.981	0.019	0.019
WM6	0.000	0.996	0.000	0.004	0.000	0.996	0.004	0.004

The results in both Table 28 and 29 indicate that the usage of the combined sub-area Q-matrix results in an increases in IDI values for most sub-area levels. For example, the IDI values across Q-matrix designs for level one for APS increase from 0.035 to 1.000 and 0.143 to 1.000 in Forms A and B, respectively. In other cases, however, IDI decreases in the combined sub-area Q-matrix design, such as in level four for WE in Form A or level three for WE in Form B. These changes reflect the increased number of possible attribute skill mastery profiles that occur as a result of the combination of the sub-area Q-matrices.

However, they also may be considered evidence of to what degree the diagnostic ability of a sub-area is subject change when viewed in the context of the other sub-areas. Furthermore, it appears that for all cases in which IDI is low, the cause appears to be either high values for the guess parameter or high values for the slip parameter, with no levels showing moderate values for both parameters simultaneously. This may indicate that if a respondent were to be assigned or not assigned a level inaccurately, the cause is either due to one or the other, without a strong probability of both. The same effect wherein combining each sub-area into a single Q-matrix appears to improve model estimates for diagnostic ability can also be observed in the full-score coded Q-matrix design item parameter estimates shown in Table 30 and Table 31.

Table 30. Form A Full-Score Coded Q-matrix Design Item Parameter Estimates without Guess or Slip Parameter Constraints

Item	Single Sub-Area Q-matrix				Combined Sub-Area Q-matrix			
	Guess	Slip	RMSEA	IDI	Guess	Slip	RMSEA	IDI
APS1	0.689	0.000	0.000	0.311	0.000	0.000	0.000	1.000
APS2	0.011	0.000	0.001	0.989	0.000	0.000	0.000	1.000
APS3	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000
APS4	0.000	0.001	0.000	0.999	0.000	0.000	0.000	1.000
APS5	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000
APS6	0.000	0.481	0.000	0.519	0.000	0.000	0.000	1.000
WE1	0.739	0.000	0.000	0.261	0.000	0.000	0.000	1.000
WE2	0.014	0.000	0.001	0.986	0.000	0.000	0.000	1.000
WE3	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000
WE4	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000
WE5	0.000	0.003	0.000	0.997	0.000	0.000	0.000	1.000
WE6	0.000	0.687	0.000	0.313	0.000	0.000	0.000	1.000
WM1	0.464	0.000	0.000	0.536	0.000	0.000	0.000	1.000
WM2	0.004	0.000	0.000	0.996	0.000	0.000	0.000	1.000
WM3	0.006	0.000	0.001	0.994	0.000	0.000	0.000	1.000
WM4	0.000	0.001	0.000	0.999	0.000	0.000	0.000	1.000
WM5	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000
WM6	0.000	0.442	0.000	0.558	0.000	0.004	0.000	0.996

Table 31. Form B Full-Score Coded Q-matrix Design Item Parameter Estimates without Guess or Slip Parameter Constraints

Item	Single Sub-Area Q-matrix				Combined Sub-Area Q-matrix			
	Guess	Slip	RMSEA	IDI	Guess	Slip	RMSEA	IDI
APS1	0.439	0.000	0.000	0.561	0.000	0.000	0.000	1.000
APS2	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000
APS3	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000
APS4	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000
APS5	0.000	0.008	0.000	0.992	0.000	0.000	0.000	1.000
APS6	0.000	0.631	0.000	0.369	0.000	0.000	0.000	1.000
WE1	0.418	0.000	0.000	0.582	0.000	0.000	0.000	1.000
WE2	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000
WE3	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000
WE4	0.000	0.001	0.000	0.999	0.000	0.000	0.000	1.000
WE5	0.000	0.008	0.000	0.992	0.000	0.000	0.000	1.000
WE6	0.000	0.616	0.000	0.384	0.000	0.000	0.000	1.000
WM1	0.161	0.000	0.000	0.839	0.000	0.000	0.000	1.000
WM2	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000
WM3	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000
WM4	0.000	0.012	0.002	0.988	0.000	0.001	0.000	0.999

WM5	0.000	0.003	0.000	0.997	0.000	0.003	0.000	0.997
WM6	0.000	0.445	0.000	0.555	0.000	0.000	0.000	1.000

In these cases, the single sub-area Q-matrix design parameter estimates indicate that level one has observable levels of guess parameter estimates and while level six has significant levels of slip parameter estimates as well. However, these results are due to large proportion of students that received a score of at least a one in all three sub-areas, the low proportion of students that received a score of six in any of the three sub-areas, and the small number of possible attribute mastery class profiles. Consequently, once the number of potential attribute mastery class profiles was increased, as seen in the combined sub-area parameter estimates, all levels for the full-score coded Q-matrix design have perfect diagnostic ability. The unconstrained full-score coded combined sub-area results also essentially mirror the parameter estimates from the Guess and Slip to Zero parameter constraint condition.

4.2.3 Cognitive Diagnostic Indices (CDI)

Similar to the item parameters shown in subsection 4.2.2, the CDI values indicate the amount of diagnostic information that is produced by each level. Figure 8 and Figure 9 compare the amount of diagnostic information produced by each level by the single and combined sub-area Q-matrix designs. Similar to the estimates for item discrimination shown in the item parameter estimates, the estimates for the amount of diagnostic information produced by each increases substantially for the first level APS 1.

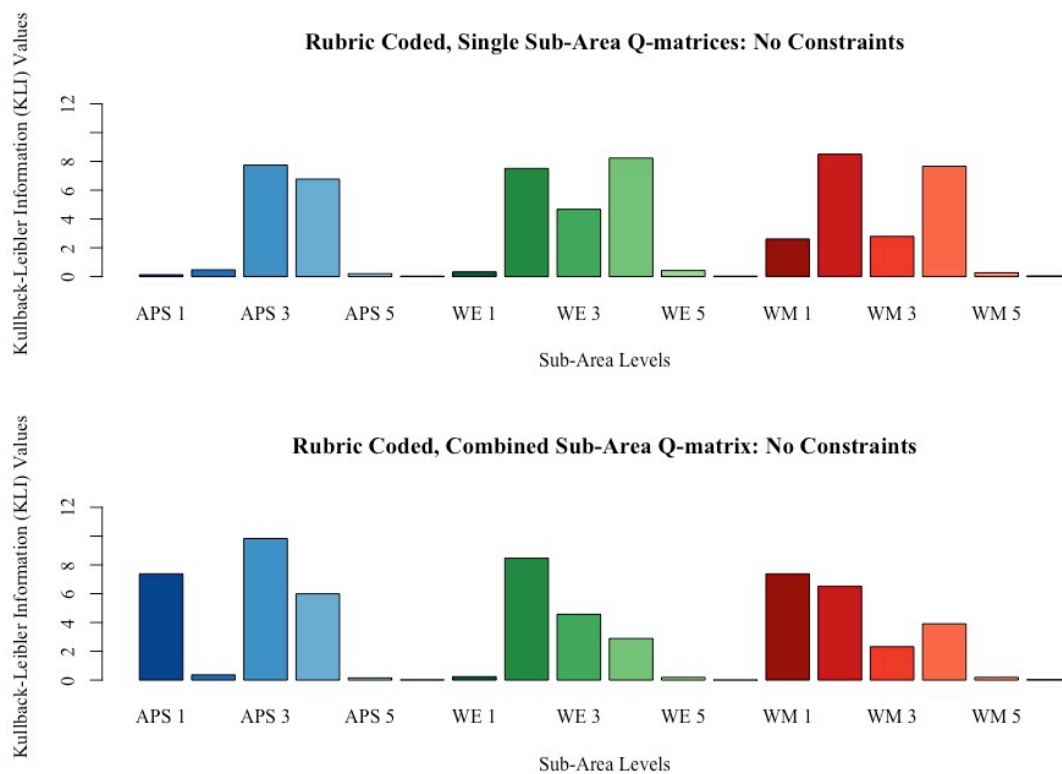


Figure 8. Form A Test Diagnostic Information

Furthermore, in Form A it appears that the levels that produce the most diagnostic information for each section appear to be level three for APS, level four for WE, and level two for WM in the single sub-area Q-matrix designs. However, in the combined sub-area Q-matrix design, the levels that produce the most diagnostic information for the exam are also level three for APS, but level two for WE, and level one for WM. There is also a clear gradient in the amount of diagnostic information produced by each level in WE, where levels one, five, and six appear to produce very little diagnostic information, while level two produces the most diagnostic information, and then levels three and four both produce less and less.

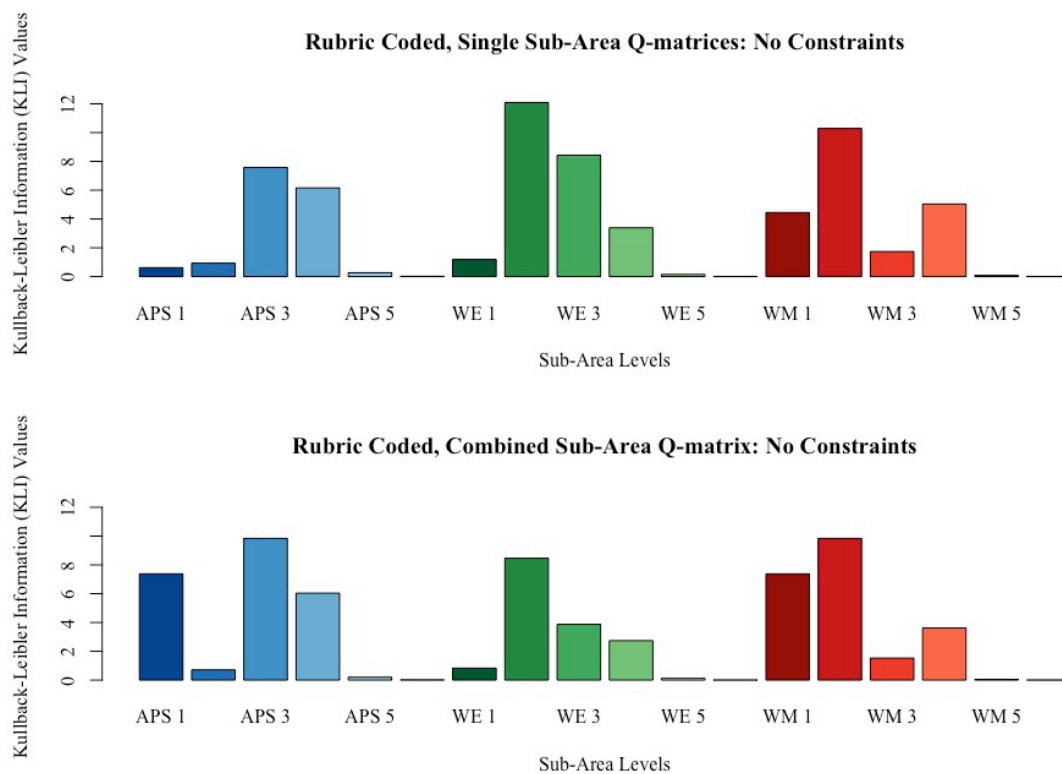


Figure 9. Form B Test Diagnostic Information

Similarly, the amount of diagnostic information produced by APS level 1 increase significantly in the combined Q-matrix designs. However, unlike Form A, there seems to be relatively small changes in diagnostic information otherwise. The only other significant change in diagnostic information is the increase in WM level 1; again from single to combined sub-area Q-matrix design. The item attribute discriminations shown in Table 32 provide a more complete perspective of the amount of information provided by each level, by showing the degree of discrimination each level is in identifying masters and non-masters of each attribute. For example, in both samples, level two for WE is the most discriminating for determining masters and non-masters of Attribute 1 for WE.

Table 32. Item Attribute Discrimination Values for Rubric Coded Combined Sub-Area Q-matrix Designs with no Parameter Constraints

Item	Form A			Form B		
	Attribute 1	Attribute 2	Attribute 3	Attribute 1	Attribute 2	Attribute 3
APS1	16.12	0.00	0.00	16.12	0.00	0.00
APS2	0.00	0.59	0.00	0.00	1.16	0.00
APS3	0.00	16.10	0.00	0.00	16.12	0.00
APS4	0.00	0.00	13.09	0.00	0.00	13.19
APS5	0.00	0.00	0.34	0.00	0.00	0.44
APS6	0.00	0.00	0.05	0.00	0.00	0.04
WE1	0.45	0.00	-	1.59	0.00	-
WE2	16.12	0.00	-	16.12	0.00	-
WE3	0.00	8.71	-	0.00	7.38	-
WE4	0.00	5.50	-	0.00	5.23	-
WE5	0.00	0.35	-	0.00	0.23	-
WE6	0.00	0.02	-	0.00	0.02	-
WM1	16.12	0.00	0.00	16.12	0.00	0.00
WM2	0.00	10.68	0.00	0.00	16.12	0.00
WM3	0.00	0.00	5.06	0.00	0.00	3.34
WM4	0.00	0.00	8.53	0.00	0.00	7.91
WM5	0.00	0.00	0.41	0.00	0.00	0.11
WM6	0.00	0.00	0.07	0.00	0.00	0.02

4.2.4 Tetrachoric Correlations

Table 33 and Table 34 both show the tetrachoric correlations for Form A and Form B in the rubric-coded combined sub-area q-matrix design without guess or slip parameter constraints. Both tables indicate that all attributes are highly correlated, which may be considered evidence of unidimensionality. Indeed, the lowest correlation between attributes is attribute two for Analysis and Problem Solving (APS2) and attribute three for Writing Mechanics (WM3) in Form A. These results however are not unexpected since not only are the attribute scores highly correlated but the attributes also follow a linear attribute hierarchy as well.

Table 33. Form A Rubric Coded Combined Sub-Area Q-matrix Design Attribute Tetrachoric Correlations

Attribute	APS1	APS2	APS3	WE1	WE2	WM1	WM2	WM3
APS1	1.00	0.95	0.95	0.99	0.94	1.00	1.00	0.94
APS2	0.95	1.00	1.00	0.98	0.96	0.95	0.97	0.90
APS3	0.95	1.00	1.00	0.97	0.97	0.95	0.95	0.95
WE1	0.99	0.98	0.97	1.00	0.98	0.99	1.00	0.98
WE2	0.94	0.96	0.97	0.98	1.00	0.94	0.96	0.95
WM1	1.00	0.95	0.95	0.99	0.94	1.00	1.00	0.94
WM2	1.00	0.97	0.95	1.00	0.96	1.00	1.00	0.97
WM3	0.94	0.90	0.95	0.98	0.95	0.94	0.97	1.00

Table 34. Form B Rubric Coded Combined Sub-Area Q-matrix Design Attribute Tetrachoric Correlations

Attribute	APS1	APS2	APS3	WE1	WE2	WM1	WM2	WM3
APS1	1.00	0.97	0.94	1.00	0.97	1.00	1.00	0.97
APS2	0.97	1.00	1.00	0.99	1.00	0.97	0.97	0.94
APS3	0.94	1.00	1.00	0.98	1.00	0.94	0.96	0.92
WE1	1.00	0.99	0.98	1.00	0.99	1.00	0.93	0.99
WE2	0.97	1.00	1.00	0.99	1.00	0.97	0.98	0.94
WM1	1.00	0.97	0.94	1.00	0.97	1.00	1.00	0.97
WM2	1.00	0.97	0.96	0.93	0.98	1.00	1.00	0.97
WM3	0.97	0.94	0.92	0.99	0.94	0.97	0.97	1.00

4.2.5 Skill Classification Estimates

The latent class profile population membership probabilities in Figure 10 and Figure 11 show the estimated for each parameter constraint condition for each sub-area in the rubric coded single sub-area Q-matrix designs. In test Form A, for the Analysis and Problem Solving (APS) sub-area, the Guess < Slip parameter constrained DINA model estimates that approximately 80% of the population belongs to latent class profile $\{1,1,0\}$, indicating that they have mastered attributes one and two but not three. Similarly, the slip parameter constrained DINA model estimated that approximately 98% of the population

that took test Form A is a member of the latent class $\{1,1,0\}$ for Writing Mechanics (WM).

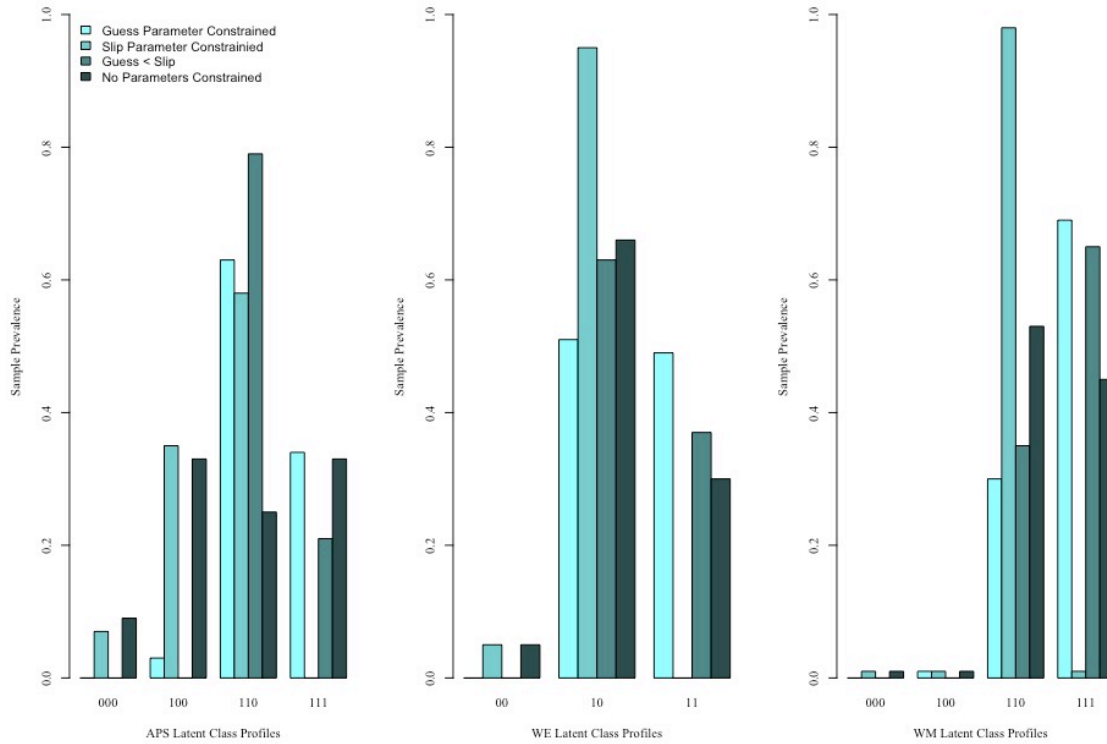


Figure 10. Form A Rubric Coded Single Sub-Area Latent Class Profile Population Membership Probabilities

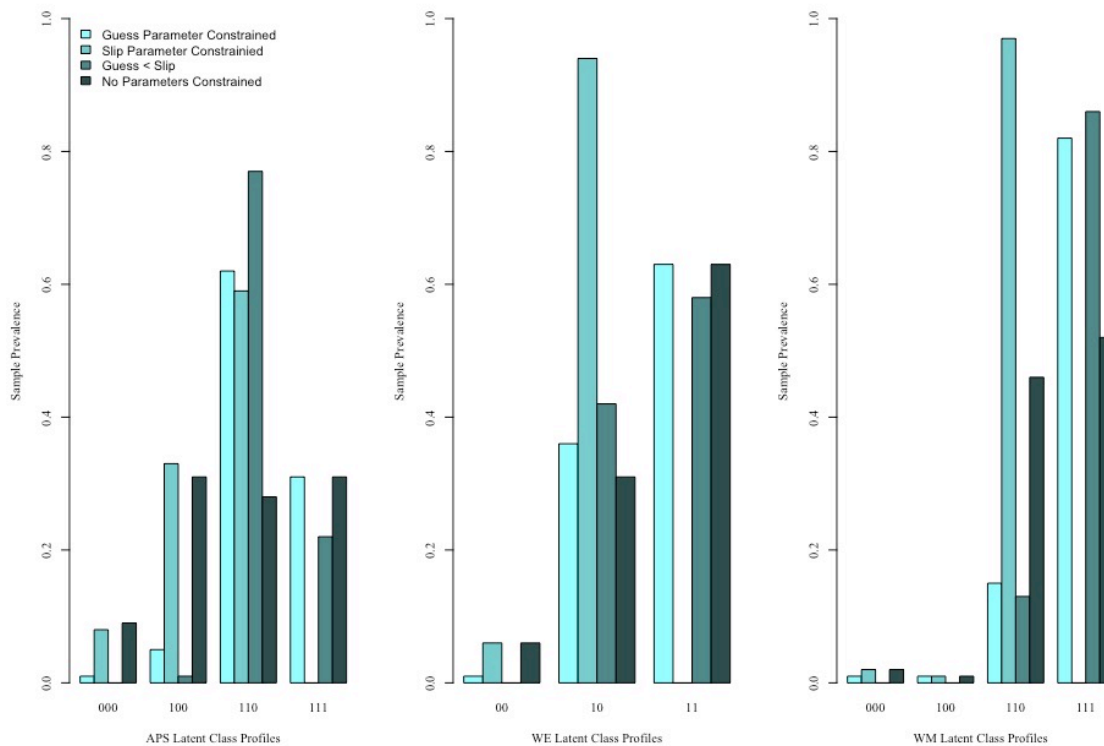


Figure 11. Form B Rubric Coded Single Sub-Area Latent Class Profile Population Membership Probabilities

Figure 11 shows that condition without parameter constraints DINA model estimates that in the population of test Form B, approximately 35% of the population belonging to latent class profiles $\{1,0,0\}$, $\{1,1,0\}$, and $\{1,1,1\}$ each. Likewise, approximately 50% of the population is a member of either latent class $\{1,1,0\}$ or $\{1,1,1\}$ for WM, according to the unconstrained DINA model. Next, the unconstrained DINA model estimated sample mastery profiles for the rubric coded combined Q-matrix design for Form A is shown in Table 35.

Table 35. Form A Unconstrained DINA Model Rubric Coded Combined Sub-Area
Sample Attribute Mastery Probability Profiles

Profile Number	APS1	APS2	APS3	WE1	WE2	WM1	WM2	WM3	Proportion
1	0	0	0	0	0	0	0	0	0.3%
2	1	0	0	0	0	1	0	0	1.3%
3	1	0	0	0	0	1	1	0	2.9%
4	1	0	0	1	0	1	1	0	23.1%
5	1	0	0	1	0	1	1	0.02	10.2%
6	1	0	0	1	0	1	1	1	2.3%
7	1	0	0	1	0.01	1	1	0.03	0.8%
8	1	0	0	1	0.47	1	1	1	0.7%
9	1	0	0	1	1	1	1	1	0.2%
10	1	1	0	1	0	1	1	0	2.6%
11	1	1	0	1	0	1	1	0.08	5.1%
12	1	1	0	1	0	1	1	1	3.1%
13	1	1	0	1	0.62	1	1	0	0.8%
14	1	1	0	1	0.7	1	1	0.26	4.7%
15	1	1	0	1	0.9	1	1	1	6.1%
16	1	1	0	1	1	1	1	0	0.1%
17	1	1	0	1	1	1	1	0.34	0.6%
18	1	1	0	1	1	1	1	1	1.7%
19	1	1	1	1	0	1	1	0	0.1%
20	1	1	1	1	0	1	1	0.05	0.4%
21	1	1	1	1	0	1	1	1	0.1%
22	1	1	1	1	0.98	1	1	0.96	1.0%
23	1	1	1	1	1	1	1	0	0.1%
24	1	1	1	1	1	1	1	0.98	0.9%
25	1	1	1	1	1	1	1	1	30.7%

N = 998

The above table represents for test Form A what proportion of the sample population was estimated to have as their most likely sample mastery probability profile. According to the results in test Form A, 30.7% of the population is most likely to have a 100% probability of mastering all of the required attributes of the CLA+ constructed-response section. Similarly, 10.2% of the population is estimated to have a 100% probability of having mastered attributes APS1, WE1, WM1, and WM2 as well as a 2% chance of being a master of WM 3.

4.3 G-PCM Analysis: Constructed-Response Section

This section presents the results of the Generalized Partial Credit Model (G-PCM) analysis, beginning with the typical descriptive statistics that would be presented in a polytomous-IRT model analysis, including CTT statistics such as test reliability. Next, tables and figures specific to G-PCM analysis are presented including item category response curves and difficulty step parameter estimates. Item information and test information figures and parameter estimates are then shown, providing insights as to at how much diagnostic information from the rubric is yielded from each rubric sub-area, as well as the entire rubric, on a latent trait continuum. Finally, the model fit of the G-PCM is compared to its Rasch model equivalent, the Partial Credit Model (PCM) in order to provide evidence that utilizing a discrimination parameter for each sub-area is appropriate for modeling rubric-scored data.

4.3.1 G-PCM Analysis

In Table 36 it is evident that all of the items in Form A have ordered step difficulties and that none of the steps are reversed, indicating that at some point on the latent trait continuum all category response options will be the most probable response option (Andrich, 1988). Writing Mechanics (WM) has the easiest transition possible, as it has the lowest step difficulty value (-3.093) for δ_{i1} , which is the transition of going from a score of zero to a score of one. Alternatively, it appears that Writing Effectiveness (WE) has the most difficult transition, with a step difficulty parameter value of 2.901 for

δ_{i6} , indicating that the most difficult transition to make on the assessment is the transition from getting a score of five to a score of six for WE. We can also see that WE also has the highest discrimination parameter, indicating that that sub-area has the highest degree of ability to discern how well an examinee will perform on the assessment, overall.

Table 37 indicates that the largest difference on the latent trait continuum between two step-difficulty parameters for Form A is between step-difficulty parameters two and three in Writing Mechanics (WM). This means that of all the items (which in our case are actually sub-areas) in the assessment, the transition from category response two to category response three requires the longest distance on the latent trait continuum. Alternatively, the shortest difference on the latent trait continuum between two step-difficulty parameters is between step difficulty parameters three and four in writing effectiveness. This means that of all the items in the assessment, the transition from category response three to category response four requires the least amount of distance on the latent trait continuum.

Table 36. Form A: Estimates of Coefficients from Fitting a G-PCM

Item	δ_{i1}	δ_{i2}	δ_{i3}	δ_{i4}	δ_{i5}	δ_{i6}	α_i
APS	-3.036	-1.871	-0.205	0.428	2.199	2.773	6.491
WE	-3.001	-1.729	0.028	0.536	2.034	2.901	8.340
WM	-3.093	-2.449	-0.463	0.074	2.143	2.608	3.523

Table 37. Form A: Latent Trait Ability Distances between G-PCM Difficulty Step Parameters

Item	Step 1 to 2	Step 2 to 3	Step 3 to 4	Step 4 to 5	Step 5 to 6
APS	1.165	1.666	0.632	1.771	0.573
WE	1.273	1.756	0.508	1.498	0.868
WM	0.644	1.986	0.537	2.069	0.465

Table 38 shows that all of the items in Form B have ordered step difficulties and none of the steps are reversed, which indicates that at some point on the latent trait continuum all category response variables will be the most probable response option. APS has the easiest transition possible, as it has the lowest step difficulty value (-2.428) for δ_{i1} , which is the transition of going from a score of zero to a score of one. Conversely, it appears that WM also has the most difficult transition, with a step difficulty parameter value of 2.970 for δ_{i6} , indicating that the most difficult transition to make on the assessment is the transition from getting a score of five to a score of six for WM. However, APS as well as WE have their most difficult parameters for δ_{i6} as well and the difference between their values are not more than 0.10. Again WE has the highest discrimination parameter, indicating that that sub-area has the highest degree of ability to discern how well an examinee will perform on the assessment, overall for Form B.

Table 39 indicates that the largest difference on the latent trait continuum between a pair of difficulty parameters for Form B is between step difficulty parameters four and five in Writing Mechanics (WM). This indicates that the transition from receiving a score of four to a score five for members of Form B requires the longest distance on the latent trait continuum. Alternatively, the shortest difference on the latent trait continuum between a pair of difficulty parameters is between step difficulty parameters one and two in WM. This means that the transition from receiving a score of three to a score of four in WM was easiest for the sample in test Form B.

Table 38. Form B: Estimates of Coefficients from Fitting a G-PCM

Item	δ_{i1}	δ_{i2}	δ_{i3}	δ_{i4}	δ_{i5}	δ_{i6}	α_i
APS	-2.428	-1.549	-0.233	0.502	2.114	2.952	6.246
WE	-2.375	-1.532	-0.311	0.402	2.068	2.871	13.312
WM	-2.308	-2.271	-0.980	-0.021	2.643	2.970	3.318

Table 39. Form B: Latent Trait Ability Distances between G-PCM Difficulty Step Parameters

Item	Step 1 to 2	Step 2 to 3	Step 3 to 4	Step 4 to 5	Step 5 to 6
APS	0.879	1.316	0.735	1.611	0.838
WE	0.844	1.220	0.713	1.666	0.804
WM	0.038	1.291	0.958	2.664	0.327

Comparing Figures 12, 13, and 14, it is evident that on the whole, item category response curves (ICRC's) 0,1,2,4, and 6 all tend to have maximum estimates of probability at or around 1.0, with the exception being ICRC 1 in writing mechanics (WM). This suggests that there are positions on the latent trait continuum wherein it is essentially 100% certain that a respondent will receive a score of zero, one, two, four, or six. Alternatively, the category response curves for 3 and 5, on the whole, appear to have maximums below 0.8, with the one exception being ICRC 5 for writing effectiveness (WE), which has a maximum of around 0.95. This suggests that for all sub-areas, there is never a 100% probability of receiving a score of three or five for any level on the latent trait continuum. Indeed, for WM it appears as though the maximum probability of receiving either a three or a five as a response on the latent trait continuum is approximately 0.6, which is substantially lower than the maximum probability estimates for all other ICRC's.

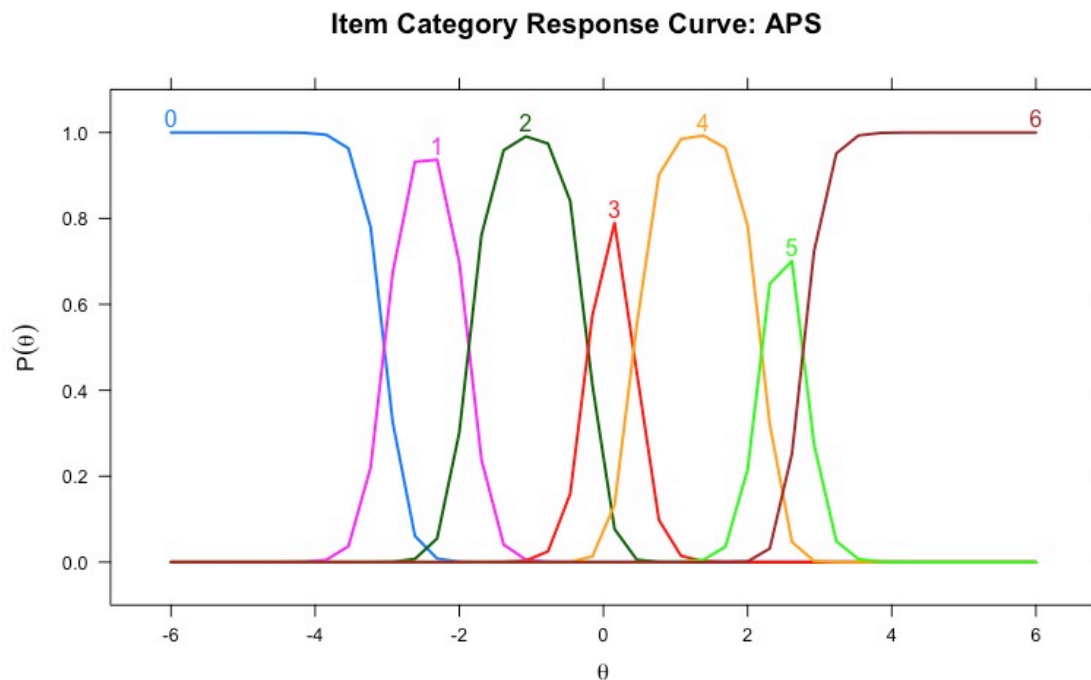


Figure 12. Form A: Item (sub-area) category response curves for analysis and problem solving (APS).

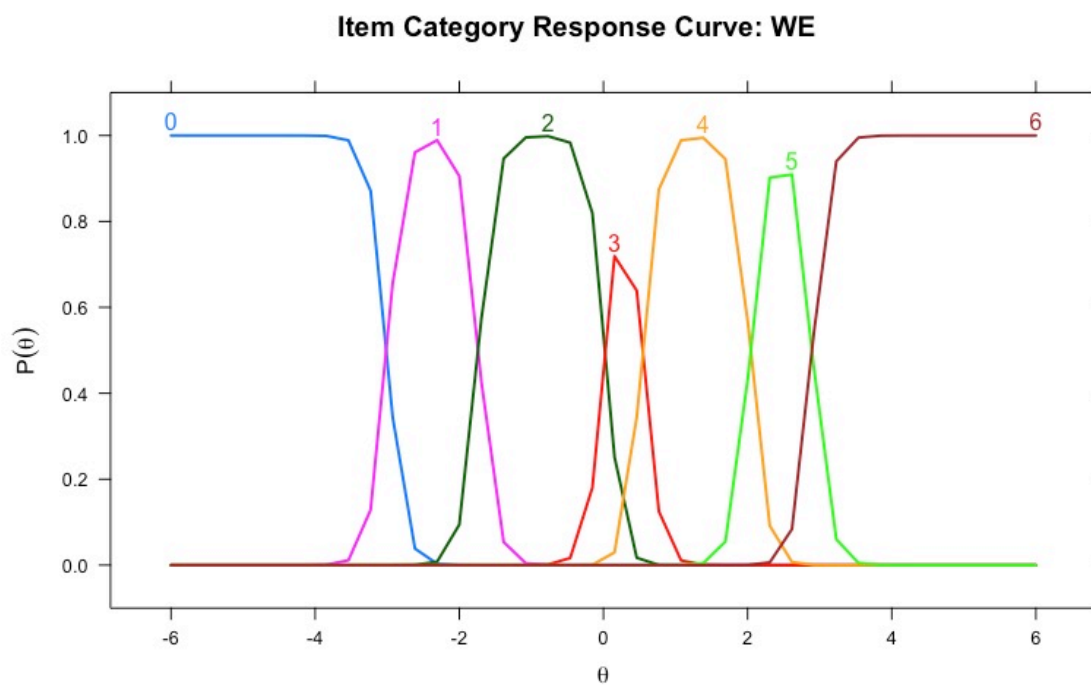


Figure 13. Form A: Item (sub-area) category response curves for writing effectiveness (WE).

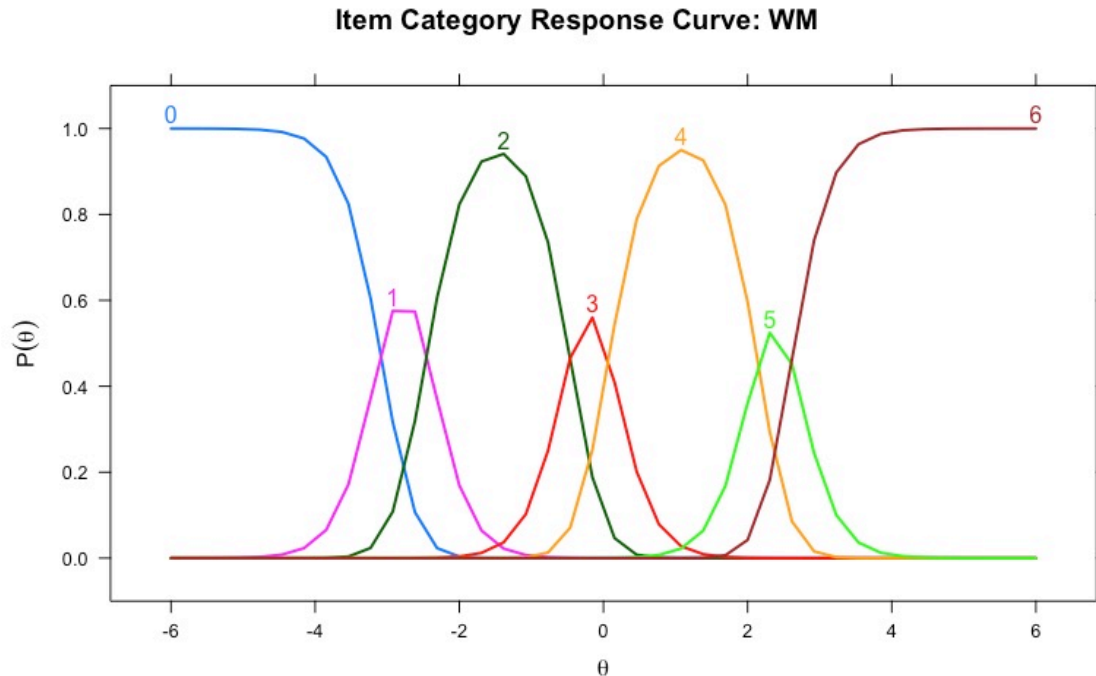


Figure 14. Form A: Item (sub-area) category response curves for writing mechanics (WM).

Figures 15, 16, and 17, show that in Form B, ICRCs 0,4, and 6 continue to have maximum estimates of probability at or around one. Still it appears as though the second sample contributes more evidence to support the conclusion that there are points on the latent trait continuum where it is very certain that a respondent will receive a score of zero, four, or six. However, the ICRCs for writing effectiveness (WE) seem to indicate that there are positions on the latent trait continuum where it is 100% probable for a response to receive a score for each level, although the distance that this is true is longer for some levels more than others, such as level four. Indeed, the ICRCs for WM for Form B show that level four remains the most probable response from an ability level of -0.021 to an level ability of 2.643 . This compared to the short amount of distance that ICRCs 1 and 5 are the most probable response levels as well as their significantly low maximum probability values of around 0.40, indicates that there may be evidence that the WM sub-area is being scored disproportionately.

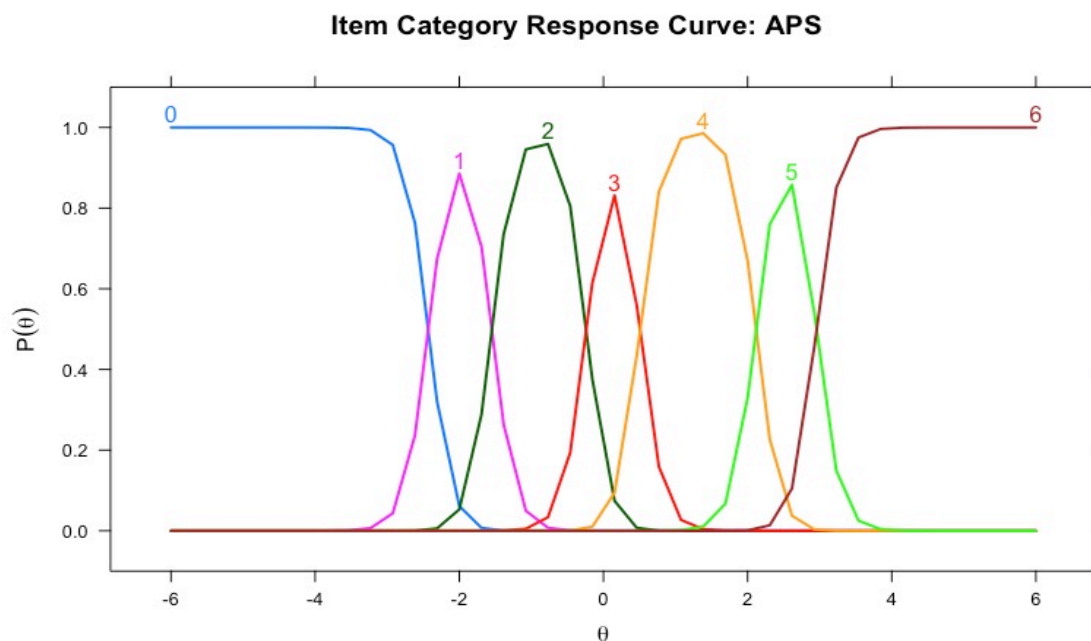


Figure 15. Form B: Item (sub-area) category response curves for analysis and problem solving (APS).

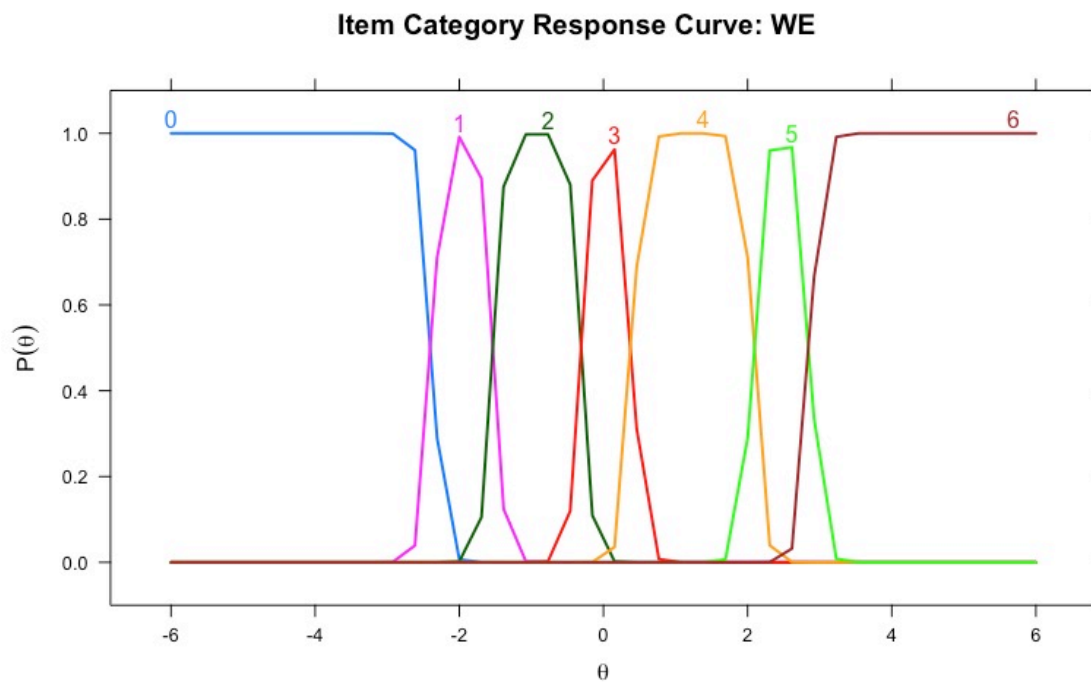


Figure 16. Form B: Item (sub-area) category response curves for writing effectiveness (WE).

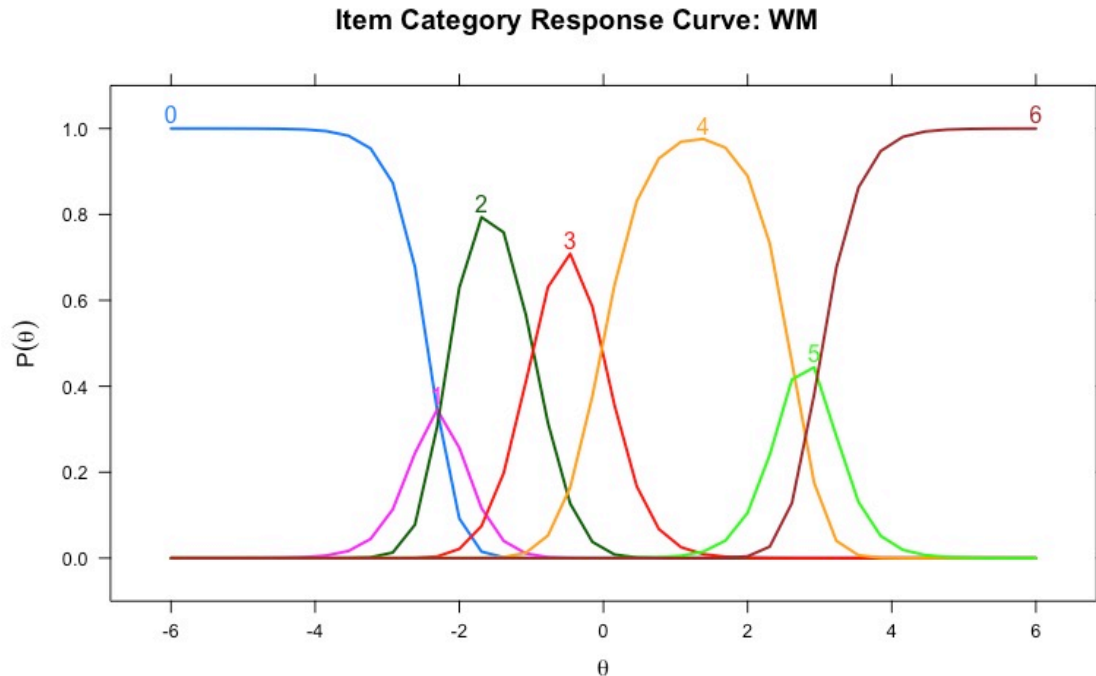


Figure 17. Form B: Item (sub-area) category response curves for writing mechanics (WM).

4.3.2 Item and Test Information

The amount of information from an item response category $I_{ij}(\theta)$ measures the degree of information or certainty that responses in that category provide regarding an examinee's ability (Baker & Kim, 2004). Higher values for information indicates where on the latent trait continuum there is lower uncertainty regarding the θ estimate and vice versa (Li & Baser, 2012). The area underneath the item information curve $I_{ij}(\theta)P_{ij}(\theta)$, known as the "item information share", represents the percentage of information contributed by each item to the overall test information (Baker & Kim, 2004). Observing Figures 18 and 19, we see that information values have a multimodal distribution along the latent trait continuum for all sub-areas. This indicates that there are sections on the latent trait

continuum, for all sub-areas, in which the degree of certainty that responses in those sub-areas estimate latent trait ability is low.

For example, the item information curves for test Form A, shown in Figure 18, the two common points on the latent trait continuum where information is lowest, for all sub-areas, also happen to be the points on the latent trait continuum where the ICRCs for 2 and 4 have their maximum probability estimates. Specifically, these maximums are approximately where θ equals -1.6 and 1.6 on the latent trait continuum in sub-areas APS and WE for categories P2 and P4, respectively. For sub-area WM, these maximums are approximately where θ equals -1.9 and 1.6 on the latent trait continuum.

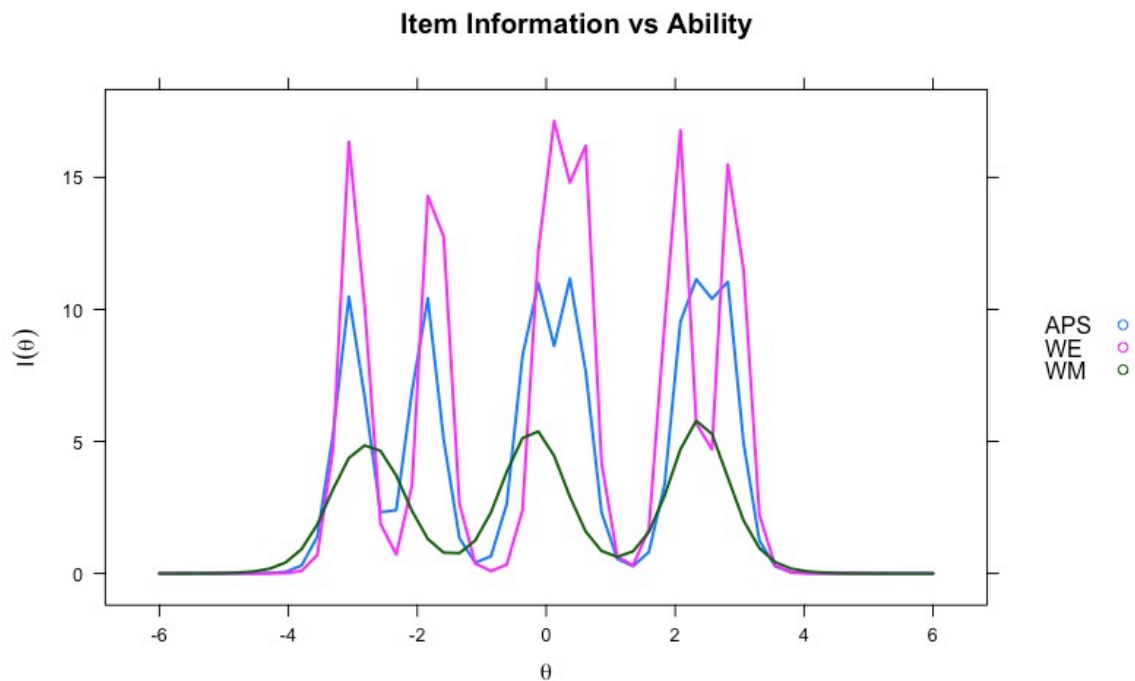


Figure 18. Form A: Item (sub-area) information curves.

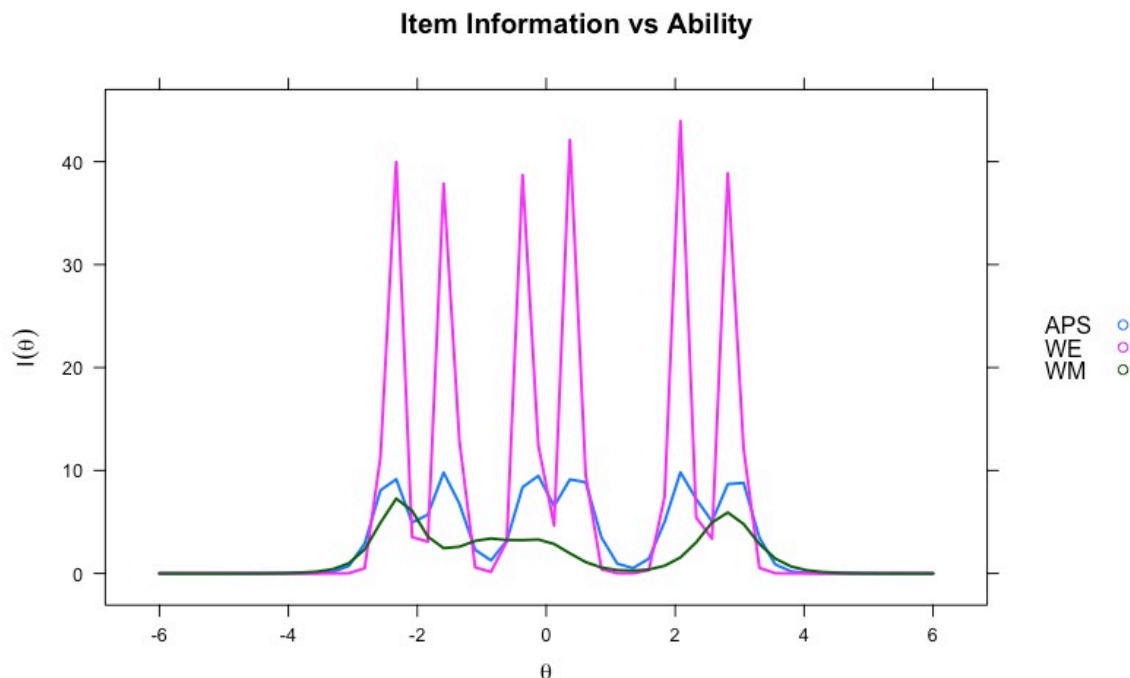


Figure 19. Form B: Item (sub-area) information curves.

As we can see in Table 40, Writing Mechanics (WM) contributes the least amount of information to the overall test information for both Form A and Form B. It is also evident that Writing Effectiveness (WE) contributes the most to the overall test information, which is interesting considering that WE also has the highest value for item discrimination (α_i). These findings corroborate the distribution of the ICRCs for WE in Figures 13 and 16, as well as the ICRCs for WM in Figures 14 and 17.

Table 40. Item (Sub-Area) Information Shares

	Form A		Form B	
	Information	Percentage	Information	Percentage
APS	38.95	35%	37.47	27%
WE	50.04	45%	79.87	58%
WM	21.14	19%	19.91	15%
Test Information:	110.13	100%	137.25	100%

4.3.3 Comparing G-PCM and PCM Model Fit

The results in Table 41 and Table 42 indicate that the AICc, SABIC, and log-Likelihood values are closer to zero for the G-PCM than for the PCM, indicating model fit. The results of the chi-squared test were also significant for both test forms, indicating that the G-PCM had significantly better fit than the PCM. These results suggest that for the data, an unconstrained discrimination parameter model fits the data better than a constrained Rasch model. This indicates that the inclusion of a slope parameter has a more parsimonious fit to the data than its exclusion, and so there is evidence that item discrimination is not equal between test items. Therefore, there is evidence to suggest that some sub-areas are more or less discriminating than others.

Table 41. Form A: Polytomous-IRT Model Fit Statistics and ANOVA Results

Model	AICc	SABIC	logLik	X2	df	p
PCM	5700.373	5732.460	-2830.798	-	-	-
G-PCM	5373.757	5409.134	-2665.405	330.786	2	0

Table 42. Form B: Polytomous-IRT Model Fit Statistics and ANOVA Results

Model	AICc	SABIC	logLik	X2	df	p
PCM	5418.623	5449.229	-2689.892	-	-	-
G-PCM	5021.771	5055.504	-2489.375	401.035	2	0

4.3.4 G-PCM Examinee Ability Estimates

Ability estimates were calculated using the expected a-posteriori (EAP) method (Embretson & Reise, 2000). In essence, theta values were estimated for every unique combination of Analysis and Problem Solving (APS), Writing Effectiveness (WE), and Writing Mechanics (WM) scores, of which there were 44 total for test Form A and 39 for test Form B. Figure 20 shows a comparison of the distribution of the ability estimates for respondents from tests Form A and Form B, respectively.

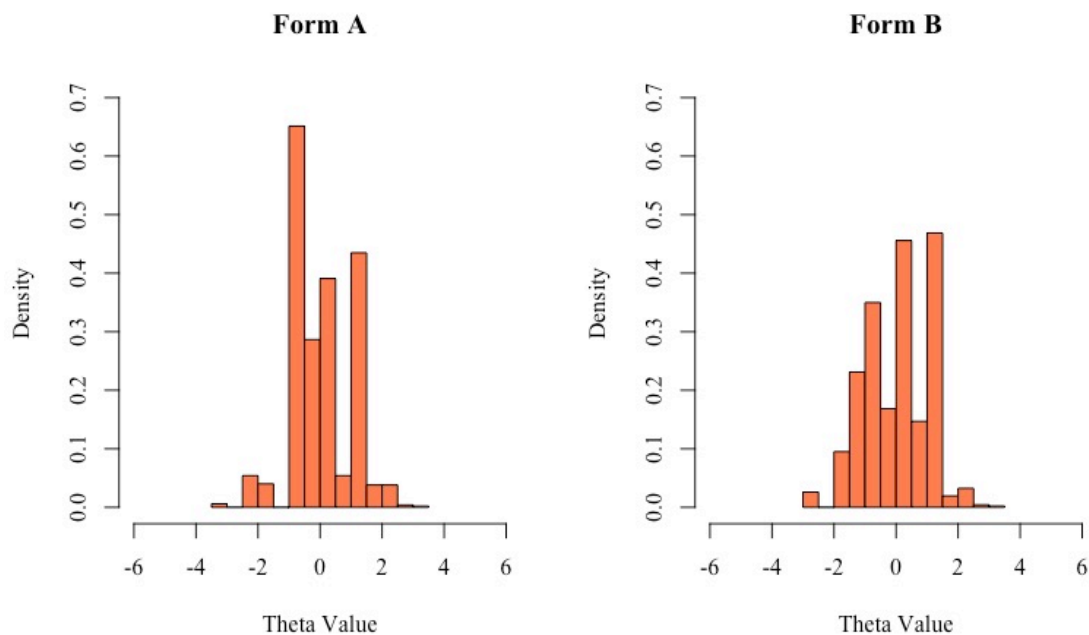


Figure 20. Form A distribution of ability estimates.

5 Discussion and Conclusions

This chapter will present a discussion of the results as they relate to the theoretical goals and contributions of the study. Specifically, the chapter will examine the efficacy of the proposed methods for retrofitting DCMs to rubric-scored constructed-response items. This will be followed by a discussion of limitations and ideas for future research.

5.1 Discussion of the Analyses

The descriptive statistics of both samples indicated that a disproportionate number of responses fell between scores of two and four for all three sub-areas of the rubric: Analysis and Problem Solving (APS), Writing Effectiveness (WE), and Writing Mechanics (WM) (see Table 19, 20). Values for Cronbach's Alpha indicated that, in both samples, the WE subsection had the greatest contribution to overall test reliability while the WM subsection had no substantive effect and, in the case of test Form B, a negative effect (see Table 21). These first two results indicated two findings with regard to the initial diagnostic capability of the rubric. The first finding was that some levels within the rubric might not be operating with equal diagnostic discrimination, and the second being that some sub-areas may also be operating with more diagnostic capability than others. However, despite these findings, sub-area score correlations appeared to indicate that the rubric scores were essentially unidimensional (see Table 22, 23). Thus, there was also evidence to suggest that sub-area scores were strongly associated, perhaps measuring the

same underlying latent trait. These results were further confirmed by the tetrachoric correlations shown in Table 33 and Table 34.

5.1.1 Model Fit Results

For the rubric-coded Q-matrix design, the relative model fit indices for both samples indicated that in most cases, the unconstrained parameters condition fit the data most parsimoniously for both single and combined sub-area analyses (see Table 24, 25). The exception to this finding was the single sub-area Writing Mechanics (WM) in test Form B, which had lower values for AICc and BIC when the guess parameter was constrained to zero. Within each parameter constraint condition, the single sub-area analyses had considerably better relative model fit than the combined Q-matrix analyses, although this finding is expected because AICc and BIC values penalize increasing model complexity, and by combining Q-matrices across sub-areas the number of parameters estimated by the DINA model increases (de la Torre & Douglas, 2008; Rupp, Templin, & Henson, 2010). In summary, the results of the relative model fit indices for rubric-coded Q-matrix designs suggest that an unconstrained DINA model is most appropriate for fitting rubric-scored data.

The absolute model fit indices of the rubric coded Q-matrices for both samples did not give definitive results regarding most parsimonious fit with the data (see Table 24,25). In test Form A, MADcor values for all three single sub-area Q-matrix designs in the unconstrained parameter condition and the WM single sub-area Q-matrix design in the Guess to Zero parameter condition indicated good model fit, with values less than

0.05. Similarly, for test Form B, MADcor values for the WM single sub-area Q-matrix design in the Guess to Zero parameter constraint condition indicated parsimonious model fit as well, although only the APS and WE single sub-area Q-matrix designs in the unconstrained parameter condition had values less than 0.05 in this sample. In both samples and across all conditions, the MADQ3 and SRMSR values were either equal to or above 0.05, indicating poor model fit for the rubric-coded Q-matrix designs.

Differences in absolute model fit statistics between study conditions are caused by differences in the correlation of observed responses of item pairs within Q-matrix design and guess/slip parameter estimates. For example, in test Form A, constraining the slip parameter to zero resulted in some items for the sub-areas WE and WM and combined Q-matrix designs to estimate an average difference of zero between observed and expected responses, resulting in non-convergence. The APS sub-area was still able to converge for MADQ3 in the Slip to Zero condition since the guess parameter values were large enough in all items to produce differences between examinees in observed and expected responses.

Based on the results of the absolute model fit indices, it is not justified to conclude that the DINA model fits parsimoniously with rubric-scored data when using rubric coded Q-matrix designs. Interpretively, the MADcor and SRMSR indices for the rubric coded Q-matrix designs indicate that there are large differences between pseudo-items in terms of the correlation of their observed values and the correlation of their expected values. Furthermore, the MADQ3 values for most parameter conditions indicate that there are large differences between observed and expected values for examinee responses as well.

Contrarily, for the full-score coded Q-matrix designs (see Table 26, 27) the relative model fit indices (AICc, BIC) indicated that in both samples setting constraining both the guess and slip parameters to zero yielded the best model fit for both single sub-area and combined Q-matrix designs. For the constrained Guess < Slip parameter condition, the relative model fit indices were noticeably higher than those in the Guess and Slip to Zero condition, but only slightly higher in the other conditions. Again, the combined Q-matrix designs produced significantly higher AICc and BIC values than the single sub-area Q-matrix designs.

In contrast to the results of the rubric coded Q-matrices, in both samples the full-score coded Q-matrix designs yielded almost uniformly perfect model fit for both MADcor and SRMSR values, with the exception of the Guess < Slip parameter constraint condition (see Tables 26, 27). These indices reflect the highly correlated observed and expected values for item pairs, indicating that there is little to no difference between the pairwise correlations of item observed values and the pairwise correlation of item expected values. In other words, the model perfectly reproduces the observed data, as expected.

The implications of this finding are that the latent response vectors are indeed mathematically identical to the observed data within the DINA model, and each level is free from guess or slip error (de la Torre, 2009). Model parsimony in this case is attributable to the polytomous score to dichotomous item response pattern conversion method as well. Given the guaranteed linear or “sequential” nature of the observed data, non-permissible observations are impossible, such as an examinee responding correctly to pseudo-item 4 but not to pseudo-item 3. If such a situation were to occur, the latent

response vectors would cease to reflect the observed item response patterns, thus introducing error and model divergence.

Moreover, while the combination of guess and slip parameters constrained to zero and a full-score Q-matrix design may allow the DINA model to fit rubric-scored data perfectly and parsimoniously, it implies the assumption that each score on the rubric represents the true score without error. As a result, unique insights regarding rubric performance are limited from retrofitting DCMs to rubric-scored data in this manner. In other words, simply being able to represent rubric-scored data and its scoring structure as a Q-matrix within the DINA model framework does not necessarily yield useful insights (except to offer a point of reference for the rubric-based Q-matrix coding).

It is also important to note that model fit indices are not and should not be definitive measures for determining the efficacy of retrofitting a DCM to an existing assessment (Thissen, 2016). Determining which model fit statistics is most suitable for analyzing DCMs is still an ongoing area of research, especially with regard to absolute model fit indices (Chen, de la Torre, & Zhang, 2013; Galeshi & Skaggs, 2014; Hu, Miller, Huggins-Manley, & Chen, 2016; Lei & Li, 2016). For example, the absolute model fit indices used in this study, which are commonly used and suggested in DCM research (George, Robitzsch, Kiefer, Groß, & Ünlü, 2016; Liu, Huggins-Manley, & Bulut, 2017), are calculated based upon item pairwise correlations that are in turn calculated using expected values estimated by the DINA model. Yet, in most DCM retrofitting cases, it is expected that the attributes measured by a diagnostic classification assessment be highly correlated (see Table 33, 34) since the process of retrofitting DCMs to standard assessments typically requires the identification of multidimensional traits

from unidimensional assessments (de la Torre & Karelitz, 2009; Sinharay, 2010, 2011, 2014). In other words, while it is known that some of the most commonly used model fit indices for DCMs are negatively affected by highly correlated attributes, much of the research on DCM analysis attempts to retrofit DCMs to existing assessments, a process which is known to produce highly correlated attributes. For reasons such as these, it is not usually expected that retrofitting DCMs to existing assessments will result in adequate absolute model fit (Gierl & Cui, 2008; Rupp & Templin, 2009).

Thissen (2016) argues that in general, IRT model fit indices should not be used as definitive criterion for decision-making with regard to assessment development, especially when applied to non-simulated data, since no model can ever perfectly fit a data set and no data set is ever perfectly unidimensional. Instead, rather than regarding relative and absolute model fit indices as verdicts for determining model functionality, they should be perceived as guidelines towards achieving the planned objective of the assessment (Thissen, 2016). Therefore, although the absolute fit indices of the rubric-coded model did not give clear indication of parsimonious model fit using the rubric coded Q-matrix design, it may still be preferable to use rubric coded Q-matrices over the full-score coded Q-matrices, as they can potentially provide useful insights regarding the rubric's measured attributes and the validity of the rubric-defined associations of specific skills with score levels.

5.1.2 Item Fit Results

Returning to the discussion regarding the poor model fit of the Guess < Slip parameter constraint condition, these findings were consistent across all forms of Q-matrix designs and across both samples (see Table 24, 25, 26, 27). The poor model fit of this condition was caused by two factors. The first factor is that, although the guess and slip parameters were constrained, there were as many parameters estimated in the Guess < Slip condition as there were in the No Constraints condition. The second factor that caused the Guess < Slip condition to have poor model fit for all Q-matrix designs was the minimum value of the slip parameter.

By restricting the slip parameter constraint to have a minimum value of 0.20, pseudo-items that would otherwise have slip parameter estimates of zero (had they been left unconstrained) could only have minimum slip parameter estimates of 0.20. This caused the DINA model to estimate expected values that were substantially different from the observed item response patterns, since the probability of a correct response in the DINA model is equal to the probability of a correct response for item j , given that a respondent has mastered all the required attributes i.e. $\{1 - s_j\}$ where s_j is the slip parameter for that item, or in this case, pseudo-item (de la Torre & Douglas, 2004). This creates larger differences between observed and expected values, which results in a lower maximum likelihood for all respondents, thus increasing the value of -2 log likelihood and AICc/BIC. In addition, larger differences between observed and expected values result in larger differences between item pairwise observed and expected value correlations, all of which increases the value of MADcor, MADQ3, and SRMSR.

The interpretation of the guess and slip parameters for rubric-scored data is perhaps unique compared to most other DCM analyses, in that the estimates for guess

and slip are not direct measurements of an examinee's possible response, since each rubric-score is given by (in the case of the CLA+ exam) the average of two or, in cases where a response is flagged as unusual, three ratings (Council for Aid to Education, 2017). Furthermore, the guess and slip parameters estimated here appear to be either extremely high or extremely low, although, like DCM model fit indices, there is not yet an established consensus for determining what is considered too high or low (Rupp, Templin, & Henson, 2010).

Note that in both samples, item parameter estimates either have a high estimate for the guess or slip parameter, but not both simultaneously. This may indicate that in a rubric-scored data context, the DINA model's estimates for the guess and slip parameters represent something other than the probability of a correct response given their attribute mastery profile. Rather, due to the linear/sequential nature of the pseudo-item response patterns as well as the fact that the pseudo-item response patterns themselves are not elicited by the respondents but by separate raters, the guess and slip parameters may either represent scorer error or levels in the rubric that are not performing as intended. However, because CLA+ constructed responses are double scored, all scorers undergo an extensive training process (Council for Aid to Education, 2017), and these results are consistent across both samples, makes it seem unlikely that high guess and slip parameter estimates are caused by scorer error.

For example, in Table 28 the single sub-area Q-matrix designs estimate for the guess parameter for sub-area WM levels 1 and 3 are 0.465 and 0.433 respectively, while the guess parameter estimate for level two is 0.004. Therefore, according to the DCM framework, WM level 1 and level 3 both have at least a 40% chance of being answered

correctly by an examinee who has not mastered the required attributes. Comparing these results to the Item Category Response Curves (ICRCs) shown in Figure 14, we also see that levels 1 and 3 for WM also have very low probabilities as well as very few values (compared to levels 0, 2, 4, and 6) for theta θ (which represents ability on the latent trait continuum) where either level is estimated to be the most probable response. Furthermore, we see in Table 36 that the distance from Step 1 to 2 and Step 3 to 4 are both extremely small compared to the distance from Step 2 to 3. This pattern suggests that large DINA model guess parameter estimates in a rubric-scored data context represent levels that are being consistently skipped or passed over by the raters/ graders.

Similarly, the same sub-area item parameter estimates in Table 28 for WM show that level 5 and level 6 have slip parameter estimates of 0.933 and 0.987, respectively. In the usual DCM framework context this would mean that there is at least a 93% chance that a respondent who has mastered the required attributes for level 5 and level 6 of WM will not be scored as such. If we look at the proportion of responses for scores of 5 and 6 for Form A in Table 19, we see that only 3% of all the scores for the sample are either a 5 or a 6. Furthermore, the ICRC for WM in Form A (see Figure 14) also shows that not only is the maximum probability of obtaining a score of 5 on WM estimated to be approximately 50%, but the width of the range on the latent trait continuum in which it is the most probable is small in comparison to the other levels as well. This finding suggests that large DINA model slip parameter estimates in a rubric-scored data context may represent levels that are too difficult.

5.1.3 Cognitive Diagnostic Indices

The cognitive diagnostic indices (CDIs) also appear to have the potential to identify levels that may be problematic. In Figure 8 and Figure 9 for example we can see that the some levels, consistent across Q-matrix designs, are not estimated to produce any diagnostic information. The values of the CDI indices also appear to be consistent between samples as well, which may be considered further evidence that the phenomenon that is being observed may be due to the design of the rubric rather than being sample specific. For example, according to the CDI indices, levels 5 and 6 do not produce any diagnostic information in any of the three sub-areas. In essence, these CDI indices can be understood much like the item parameter estimates in that they appear to identify potentially faulty rubric levels.

In contrast, the item attribute discrimination values presented in Table 32 show the estimated degree to which each level can discriminate masters and non-masters of each attribute. Comparing the results of the values in Table 32 to the single sub-area CDI values in Form A (see Figure 8) and Form B (see Figure 9), it is evident that total amount of diagnostic information each level produces for the test is the sum of its attribute discrimination values. Interestingly, it seems that in both samples, item discrimination for attribute 2 in APS increases dramatically at level 3 compared to level 2. Alternatively, in WM, the lowest score level at which attribute 2 is implicated (level 2), is the most discriminating between masters and non-masters. However, this result is most likely due to the fact that level 2 is the only level that measures only attributes 1 and 2 for WM -- the others either only measure attribute 1 or all three attributes at once.

5.1.3 Skill Classification Estimates

The skill classification estimates provided examples of the kinds of results that test takers and administrators could hope to see from applying DCMs to rubric-scored constructed-response data. Normally, test takers are given three different scores (one for each sub-area) on a scale of one to six. Theoretically, there exist 216 possible combinations of the three sub-area scores. As seen in Figure 10 and 11, a test taker can instead be classified into one of three or four latent class profiles for each sub-area, classifying them as either masters or non-masters of the attributes measured by each subsection. This process reduces the number of possible combinations of skill mastery to 48.

However, the sub-area scores tend to be highly correlated (see Table 22, 23). Furthermore the present results reframe the perception of successful test performance. Instead of focusing on averages and percentile ranks of the individual sub-area scores, test stakeholders should perhaps frame their results in terms of attribute mastery class probabilities instead. For example, Table 34 demonstrates how a test sample population can be broken down by DCMs into proportions of different attribute mastery classes. Each student is placed, based on the maximum likelihood estimated by the DINA model, into one of the 25 classes shown in the table.

Using these skill classifications, the language of the test results go from being summative and comparative (e.g. a score of five versus a score of four for APS) to being more formative and probabilistic (e.g., members of profile 13 have a 62% chance of exhibiting mastery of the second attribute of Writing Effectiveness, WE). Thus, DCMs may have potential for improving the diagnostic and formative capabilities of rubric-

scored tests. However, this is assuming that these test scores behave in a way that is in fact more formative and diagnostic in actual practice. Much research is still needed, not only in the context of this study that demonstrates the validity of the assumption that the results DCAs produce are in fact more formative and diagnostic than the results that typical unidimensional assessments produce.

Furthermore, compared to the G-PCM analysis, the results of the DINA model application to rubric-scored data are less straightforward in terms of interpretation. Unlike the DINA model, the G-PCM is able to identify, for exact locations on a theoretical latent trait continuum, the probability of being scored at one level in the rubric over another (see Table 35, 37). Like the guess and slip parameters estimated by the DINA model, these step difficulty parameters may provide evidence of inefficiencies in rubric sub-area scoring that may be occurring as a result of the rubric's design.

Additionally, treating each sub-area as its own item within the G-PCM also allows for the calculation of item information shares of a sub-area for the entire assessment, which highlights the sub-areas that contribute the most or the least towards overall test information. Model fit estimates and model fit tests between the G-PCM and PCM, in both samples, indicate that the G-PCM is more appropriate to be fit to the data than the PCM, which is evidence that the inclusion of an item discrimination parameter fits the data more parsimoniously (see Table 40, 41).

Lastly, the G-PCM is also able to provide ability estimates for individuals based on their combination of APS, WM, and WE scores (see Figure 20). These ability estimates are able to place each unique combination of test scores at its own position on the latent trait continuum. For example, normally a total constructed-response score of 10

on the CLA+ is equal to any other total score of 10, regardless of what combination of sub-areas constitute that score. According to the G-PCM ability estimates, each combination of sub-area scores is on a different location on the latent trait continuum, therefore an individual that has a score of three for APS, four for WE, and three for WM may be higher or lower on the latent trait continuum than an individual that has a score of five for APS, three for WE, and two for WM.

However, these kinds of G-PCM results do not fulfill the original objective of the present study, which was to propose and evaluate method to provide students as well as test administrators and developers with more formative and diagnostic feedback from rubric-score tests. If the ability estimates from the G-PCM are used merely to rank student performance, then the method does not change the unidimensional summative nature of the assessment. However, if the ability estimates were to be used as weights for sub-area scores or score combinations, or by test administrators to identify which sub-areas required more attention, then the G-PCM ability estimates could also be conceivably used to improve the diagnostic/formative ability of a rubric-scored exam as well.

5.2 Conclusions

The theoretical objective of this study was to propose methods for retrofitting DCMs to rubric-scored constructed-response items with three specific goals in mind: to examine whether or not a rubric-scored item can reasonably be conceived of as a DCA and describe the conditions under which a numeric or ordinal rubric score can be represented

by a Q-matrix, to propose specific methods for retrofitting DCMs to rubric-scored data, and to evaluate these methods and gain insights into their usefulness by applying them to a specific example of rubric-scored test, the CLA+. These results also provide an additional specific benefit, namely insights and guidance for future rubric design. With regard to the first goal, the study showed that a scoring rubric represents inherently hierarchical structures that can be represented in a mathematical space, such as a Q-matrix. These Q-matrices are unlike the Q-matrices originally designed for multiple-choice format DCMs in that each 'item' in the Q-matrix represents a level on the rubric. Put another way, the design of these Q-matrices must take into account the linear hierarchical structure of the attributes, because only certain attribute mastery classes will be permissible (Liu & Huggins-Manley, 2016; Liu, Huggins-Manley & Bradshaw, 2016; Madison & Bradshaw, 2015).

Regarding the second goal, the results showed that numeric, graded, or ordinal rubric scores can be represented by a Q-matrix when rubric sub-areas and sub-area attributes are identifiable and the graded nature of the rubric-scored data can also be represented in the same mathematical space. Within each sub-area, an attribute hierarchy as it relates to each level in the rubric must be delineated, preferably using research backed techniques such as those found in the Rule Space Method (Tatsuoka, 1983) or the Attribute Hierarchy Method (Leighton, Gierl, & Hunka, 2004). The present study proposed two methods for representing a Q-matrix in such conditions, the first being a rubric-based method, in which the language of the rubric criterion determined attribute loadings for each level and the second being a score based method, in which a Guttman scale (Guttman, 1944, 1950) was formed according to the magnitude of each sub-area's

highest score. The optimal way to identify sub-areas and sub-area attributes would be through collaboration with multiple subject matter experts, in order to ensure the validity of the sub-areas as well as the attributes being measured (Lee, Park, & Taylan, 2011).

In addition to identifying rubric sub-areas and sub-area attributes, the graded nature of the rubric data must also be represented within the same mathematical space as the Q-matrix. Unlike typical DCM analysis, the attribute hierarchies in rubric-scored data are already known, as well as which classes are permissible or impermissible as a result of the attribute hierarchy. The study satisfied this condition by a method essentially equivalent to Tutz's (1997) sequential response mechanism, that created pseudo-item response patterns wherein items that were marked correct unconditionally had all items that preceded it marked as correct as well. If the Q-matrix and the rubric-scored data can be represented within the same mathematical space, the study has shown that DCMs can be applied to rubric-scored data in a way that produces attribute mastery estimates derived from fine-grained skills either described in the criterion of the rubric or implied by its inherent linearly hierarchical structure.

With regard to the third theoretical contribution, the study showed that both the DCM and G-PCM analysis were able to identify signs of inconsistent or uneven performance for different levels within each sub-area, using the parameter estimates and indices presented in this study. For example, the DCM analysis showed that in both samples, the item parameter estimates for Writing Mechanics (WM) (see Table 28, 29) begin high at level 1 and level 2, decrease sharply at level 3 due to a sudden increase in guess parameter value, and then increase sharply at level 4. These results are corroborated by the cognitive diagnostic indices (CDIs) as well (see Figure 8, 9) where

the amount of diagnostic information from WM is noticeably lower between level two and level four. The G-PCM analysis then shows that the most difficult transition to make on the latent trait continuum, for both samples, is the transition from a score of four to a score of five in WM (see Table 36, 38). The item category response curves (ICRCs) of both samples (see Figure 14, 17) then clearly demonstrate the inordinate amount of space that level four in WM occupies in both samples. Furthermore, the item information shares in Table 39 show that WM contributes the least amount of information towards the overall test information. From these results, it may be possible to conclude that the levels within WM appear to not be functioning at their fully intended diagnostic capacity.

Assuming that the methods of rubric analysis demonstrated in this study are effective at identifying levels or sub-areas that are performing below the desired level of diagnostic ability, it may be possible to use such an analysis to make recommendations for future rubric design. In both samples, IDI, CDI, and step difficulty parameters show that levels 5 and 6 appear to offer little to no diagnostic value when using rubric-coded Q-matrices. Using these findings as evidence, one possible conclusion may be that six levels is unnecessary for the rubric, or alternatively, that the language of the rubric gradient needs revisions in order to produce less disproportionate score distributions.

Another possible conclusion is that it may not be necessary to measure certain attributes in every level of the rubric or that there may some attributes may need to be broken down into multiple attributes. For example, it appears that Attribute 1 in APS has high estimates for the guess parameter in level 1 (see Table 28, 29). This may be an effect of the CLA+ scoring method in which scorers are aware that in order for a response to be scored, it must meet the minimum requirements for a score of 1 on APS.

Also, one can argue that the language of the rubric for APS implies that a level score of 1 seems to imply non-mastery of Attribute 1 (see Table 9). If this Q-matrix entry is changed to ‘0’, then no skills are present for a score level of 1, which is not allowed in the DINA model. Therefore, if this alternate Q-matrix is used, it must be extended with an additional “universal” attribute, which is required by all skill levels. An example of the resulting revised Q-matrix is shown below in Table 43. Such revisions may also be applied to the WE and WM subsections as well, if appropriate.

Table 43. Example of a Revised Q-matrix for Analysis and Problem Solving (APS)

Pseudo-Item	Attribute 1	Attribute 2	Attribute 3	Attribute 4
1	1	0	0	0
2	1	1	1	0
3	1	1	1	0
4	1	1	1	1
5	1	1	1	1
6	1	1	1	1

Table 32 shows that the item attribute discrimination values identify which levels are the most discriminating for classifying masters and non-masters of each attribute. When values equal zero, this indicates that the level has no diagnostic ability to determine masters and non-masters of the attribute. In such cases, this may be considered evidence of levels on the rubric where those attributes no longer need to be measured. The implication is that, once a discovered threshold has been reached, attribute mastery has already occurred and so no subsequent levels can demonstrate further mastery. These attribute discrimination values then, combined with the rest of the DCM and G-PCM analysis can then guide the design of new Q-matrices by specifying new loadings for the attributes. Specifically, the results could be considered evidence to suggest reducing the number of pseudo-items in the Q-matrix. The descriptive statistics indicated that, in

all three sub-areas and across both samples, of the six levels measured by the rubric, at least 90% of the responses were scored as a 2, 3, or 4. The results of the DCM and G-PCM analysis also showed that levels 5 and 6 might have been too difficult. Based on these results, the number of pseudo-items could be reduced from six to five.

The results of the item parameter estimates and fit statistics (see Table 28, 29, 30, 31) also appear to show trends wherein lower levels in the rubric have higher estimates for guess parameters and higher levels in the rubric have higher estimates for slip parameters. Interpretively, in the case of applying DCMs to rubric-scored data (as opposed to directly measuring examinee responses) the guess parameters in this analysis represent the probability of being scored at a particular level or higher on the rubric without actually having mastered the required attributes. Similarly, the slip parameter in this analysis can be interpreted as the probability of not being scored at a particular level although the required attributes have been mastered. Therefore, in effect, the guess and slip parameters represent rater leniency and severity, respectively, when the DINA model is applied to rubric-scored data.

This interpretation of the guess and slip parameters is consistent with recent work by Tu, Zheng, Cai, Gao, & Wang (2017) wherein the researchers proposed a version of the DINA model for graded data (DINA-GD). In the DINA-GD model, the guess parameters are constrained so that respondents are more likely to guess on lower item scores i.e. the probability of guessing decreases as item scores increase. Similarly, the DINA-GD model constrains the slip parameter so that respondents are more likely to slip on higher item scores i.e. the probability of slipping increases as item scores increase. The main difference between the model proposed by Tu, Zheng, Cai, Gao, & Wang

(2017) and the approach to analyzing rubric-scored data in this study, is that in this study the graded nature of the data is inherently captured by the polytomous to dichotomous score conversion method, whereas in the DINA-GD model study the graded nature of the data is captured by the model parameters.

Finally, to frame the results of the study in terms of improving the formative and diagnostic ability of rubric performance, it is useful to consider how conducting a DINA model analysis of rubric-scored data as described in this study benefits test developers, administrators, and users. For test developers, the benefits to conducting a DINA model analysis of rubric-scored are centered on the identification of problematic or disproportionately scored rubric levels. A DINA model analysis of rubric-scored data is a method of supplying evidence for cases wherein rubric levels may be providing little to no diagnostic information. Specifically, item fit and parameter estimates, as well as cognitive diagnostic indices (CDI) showcase the probability that respondents will be scored at particular levels with respect to the attributes that they measure.

From these results, test developers may then be able to make recommendations to test administrators for potential changes that can be made to the rubric so that the rubric performs at an efficiency that is more optimal. If the revised rubric performs with greater efficiency as a result of the revisions made to the rubric based on the DINA model analysis of the rubric-scored data, then, within the context of the DCM framework, the rubric is classifying attribute masters and non-masters with greater efficiency as well. If the revised rubric is distinguishing between attribute masters and non-masters with greater efficiency, than the diagnostic of the rubric has been improved as a result of the DINA model analysis, and therefore its formative ability has been improved as well.

Lastly, the skill classification estimates provided by the DINA model, wherein respondents are given their most probable latent class profiles may be a more formative experience than summative test scores, which is one of the main objectives of the DCM framework in most educational contexts (Rupp, Templin, & Henson, 2010). Students may find that being classified as either masters or non-masters of the skills measured by a rubric, or their probability of having mastered particular skills or attributes, may be more conducive towards learning than graded numeric scores. However, this is the main theory that motivates research in DCMs within educational contexts, and has yet to be established or verified. Future research should attempt to validate the premise of research in DCM application within educational contexts that providing students with skill classification estimates and probabilities is more advantageous to learning development than numeric or summative test scores.

The disadvantages of the method proposed in this study for analyzing rubric-scored data using the DINA model lay within its difficulty of application, complexity in interpretation, and experimental nature. While the method may be statistically helpful, it is most likely not practically convenient for most test developers or administrators as its application requires a significant breadth of knowledge in psychometrics, specifically in DCMs, which is still a burgeoning topic in the field. Furthermore, the interpretation of the results from such an analysis is unlike those seen in most current applications of the DINA model, which are normally to dichotomous correct/incorrect multiple-choice data. Therefore, even those who are well versed on the topic of DCMs may still have difficulty in interpreting the meaning of their results as well.

Moreover, the G-PCM analysis appeared to provide a much more direct interpretation of the same kinds of anomalies found in the DINA model analysis. While being able to compare the results of the DINA model and G-PCM analysis proved to be effective at creating a more complete concept of what was occurring in the rubric-scored data statistically, the main difference between the DINA model and G-PCM analysis was that the DINA model was able to provide insights as to how the performance of the rubric related to the measured attributes. The G-PCM was also able to provide estimates for ability based on the combination of rubric sub-area scores (see Figure 20), in effect demonstrating that different combinations of sub-area scores represent different locations on the theoretical latent trait continuum. In fact, due to differences in item (which in this case are sub-areas on the rubric) discrimination, some combinations of sub-area scores may have higher ability level estimates even though their summed total is less than another combination's summed total.

For example, two respondents may both have total scores of 17, while one respondent has a score of 6 for APS, 5 for WE, and 6 for WM and the other has a score of 6 for APS, 6 for WE, and 5 for WM. In this case, the second respondent will have a higher estimate for ability on the latent trait continuum due to the fact that the WE sub-area has a higher value for discrimination than WM. These results then provide evidence to test developers and administrators of differences in performance between rubric levels. However, much like how research has not yet determined whether or not the effect of providing respondents with DINA model skill mastery classification or probability estimates is more formative for students than typical summative test scores, the same can also be said for G-PCM theta ability estimates. Future research should also examine the

effect that providing students with theta ability estimates have on improving test formative and diagnostic ability, especially since theta ability estimates in effect rank student scores even further.

5.3 Limitations and Future Research

One of the major limitations of the study was that the design of the Q-matrix was based on the subjective interpretation of the researcher, guided by the language of the test's scoring rubric. Ideally, a subject matter expert would have been consulted to help identify the number of attributes in each sub-area of the rubric. Because Q-matrix misspecification has a significant influence on the outcome of a DCM analysis, it is integral that the Q-matrices be specified correctly (Chen, Liu, Xu, & Ying, 2015; Chiu, 2013). Future research should examine the effects of differently rubric-coded Q-matrix designs in order to determine the root cause of changes in the results.

Another limitation of the study is the potential for sample-specific results. Although the sample population used in this study was not small, similar studies should use both alternative samples and simulation studies in order to establish maximally generalizable results. Rubrics from other assessments should also be analyzed, with varying lengths and numbers of attributes, in order to better clarify the causality of different effects. Subsequent research should also investigate simulations that manipulate variables that have the potential to effect the results, such as Q-matrix design, model parameter constraints, the DCM condensation rule, sample size, test length, or score distribution. A follow up simulation study, for example, may choose random initial

values for the DINA model guess and slip parameters and average the results over many iterations in order to determine their effect on model convergence and parameter estimation. By generalizing the results over multiple populations and measurement tools, it may be possible to establish statistically based protocols for improving the diagnostic ability of rubrics and rubric based assessments. If the reliability

In conclusion, the comparison of efficacy between DCMs and traditional IRT models should be further explored. Although DCMs have emerged as a potential for providing test takers and administrators with more formative feedback, their complexity may sometimes negate their advantages over simpler mathematical models that can accomplish the same goals (Gorin, 2009; von Davier & Haberman, 2014). The results of this study demonstrate the value of comparing the present proposed retrofitting methods for DCMs with more established methods of psychometric research. Recent research has also proposed using a combination of DCM and IRT methods for analyzing data as well (Bradshaw & Templin, 2014). Researchers should continue to compare and contrast the difficulties and benefits of DCMs and more traditional mathematical models such as IRT, applied to a variety of test formats and application areas.

References

- Abdi, H. (2010). Guttman scaling. In N. Salkind (Ed.), *Encyclopedia of research design* (pp. 1-5). Thousand Oaks, CA: Sage.
- Adams, R. J. (1988). Applying the partial credit model to educational diagnosis. *Applied Measurement in Education, 1*(4), 347-361.
- 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>
- Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. *Sociological Methodology, 15*, 33-80.
- Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). University of Maryland College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.
- Baker, F. B., & Kim, S.-H. (2004). *Item response theory parameter estimation techniques* (Second, Rev. and Expanded ed.). Boca Raton, FL: Taylor & Francis Group.
- Baker, J. G., Rounds, J. B., & Zvon, M. A. (2000). A comparison of graded response and Rasch partial credit models with subjective well-being. *Journal of Educational and Behavioral Statistics, 25*(3), 253-270.
- Banerjee, J., Yan, X., Chapman, M., & Elliott, H. (2015). Keeping up with the times: Revising and refreshing a ratings scale. *Assessing Writing, 26*, 5-19.
<https://doi.org/10.1016/j.asw.2015.07.0011075-2935/>

- Bazaldúa, D. A. L. (2016). *Exploring skill condensation rules for cognitive diagnostic models in a bayesian framework* (Unpublished doctoral dissertation). Columbia University, New York, NY.
- Benjamin, R. (2014). Two questions about critical-thinking tests in higher education. *Change: The Magazine of Higher Learning*, 46(2), 24-31.
<https://doi.org/10.1080/00091383.2014.897179>
- Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education*, 39, 370-407.
<http://dx.doi.org/10.3102/0091732X14554179>
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443-459.
- Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, 35, 179-197.
- Bradshaw, L., & Templin, J. (2014). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika*, 79(3), 403-425. <https://doi.org/10.1007/S11336-013-9350-4>
- Buck, G., & Tatsuoka, K. K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing*, 15, 119-157. <https://doi.org/10.1177/026553229801500201>
- Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury.

- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6).
- Chappuis, S., & Chappuis, J. (2008). The best value in formative assessment. *Informative Assessment*, *65*(4), 14-19.
- Chen, H., & Chen, J. (2016). Retrofitting non-cognitive diagnostic reading assessment under the generalized DINA model framework. *Language Assessment Quarterly*, *13*(3), 218-230. <https://doi.org/10.1080/15434303.2016.1210610>
- Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement*, *37*, 419-437.
- Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*(2), 123-140.
- 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.
- 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>
- 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>

- Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*(4), 633-665.
<https://doi.org/10.1007/S11336-009-9125-0>
- Choi, K. M., Lee, Y.-S., & Park, Y. S. (2014). What CDM can tell about what students have learned: An analysis of TIMSS eighth grade mathematics. *Eurasia Journal of Mathematics, Science & Technology Education*, *11*(6), 1563-1577.
- Clauser, B. E. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement*, *24*(4), 310-324.
- Close, C. N. (2012). *An exploratory technique for finding the Q-matrix for the DINA model in cognitive diagnostic assessment: Combining theory with data* (Unpublished doctoral dissertation). University of Minnesota, St. Paul, MN.
- Council for Aid to Education. (2017). *CLA+ technical FAQs*. New York, NY: Council for Aid to Education.
- Council for Aid to Education. (2018). Council for aid to education. Retrieved April 23, 2018, from <http://CAE.org>
- Cui, Y., Gierl, M. J., & Chang, H.-H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement*, *49*(1), 19-38.
- de Ayala, R.J. (2009). *Methodology in the Social Sciences: The theory and practice of item response theory*. New York, NY: The Guilford Press.
- DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement*, *35*(1), 8-26. <https://doi.org/10.1177/0146621610377081>

- 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.
- de la Torre, J. (2009). DINA Model and Parameter Estimation: A Didactic. *Journal of Educational and Behavioral Statistics*, 34(1), 115-130.
<https://doi.org/10.3102/1076998607309474>
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76(2), 179-199. <https://doi.org/10.1007/s11336-011-9207-7>
- de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69, 333-353. <https://doi.org/10.1007/BF02295640>
- de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, 73(4), 595-624. <https://doi.org/10.1007/S11336-008-9063-2>
- de la Torre, J., & Karelitz, T. M. (2009). Impact of diagnosticity on the adequacy of models for cognitive diagnosis under a linear attribute structure: A simulation study. *Journal of Educational Measurement*, 46(4), 450-469.
- de la Torre, J., & Minchen, N. (2014). Cognitive diagnostic assessments and the cognitive diagnosis model framework. *Psicologia Educativa*, 20, 89-97.
<http://dx.doi.org/10.1016/j.pse.2014.11.001>
- DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. *Handbook of Statistics*, 26, 979-1030. [https://doi.org/10.1016/S0169-7161\(06\)26031-0](https://doi.org/10.1016/S0169-7161(06)26031-0)

- DiBello, L. V., Stout, W. F., & Roussos, L. (1995). Unified cognitive psychometric assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361-390). Hillsdale, NJ: Erlbaum.
- Donoghue, J. R. (1993, March). *An empirical examination of the IRT information in polytomously scored reading items* (Research Report No. RR-93-12). Princeton, NJ: Educational Testing Service.
- Eggert, S., & Bögeholz, S. (2009). Students' use of decision-making strategies with regard to socioscientific issues: An application of the Rasch partial credit model. *Science Education*.
- Embretson, S. E. (1994). Applications of cognitive design systems to test development. In C. R. Reynolds (Ed.), *Cognitive assessment: A multidisciplinary perspective* (pp. 107-135). New York: Plenum Press.
- Embretson, S. E. (1998). A cognitive design system approach to generating valid tests. *Psychological Methods*, 3, 380-396.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Foster, G. C., Min, H., & Zickar, M. J. (2017). Review of item response theory practices in organizational research: Lessons learned and paths forward. *Organizational Research Methods*, 20(3), 465-486.
- Fulcher, G. (1987). Tests for oral performance: the need for data-based criteria. *ELT Journal*, 41(4), 287-291.

- Fulcher, G. (1987). Tests of oral performance: The need for data-based criteria. *ELT Journal*, 41(4), 287-291. <https://doi-org.ezproxy.cul.columbia.edu/10.1093/elt/41.4.287>
- Galeshi, R., & Skaggs, G. (2014). Traditional fit indices utility in new psychometric model: Cognitive diagnostic model. *International Journal of Quantitative Research in Education*, 2(2), 113-132.
- Gardner, D. P. (1983, April). *A nation at risk: The imperative for educational reform. An open letter to the American people. A report to the nation and the secretary of education*. Washington, DC: Department of Education.
- Garrison, C., & Ehrlinghaus, M. (2010). *Formative and summative assessments in the classroom*. Unpublished manuscript, National Middle School Association, Westerville, OH.
- George, A. C. (2013). *Investigating CDMs: Blending theory with practicality* (Unpublished doctoral dissertation). TU Dortmund University, Dortmund, Germany.
- George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software*, 74(2), 1-24. <https://doi.org/10.18637/jss.v074.102>
- Gibbons, R. D., Bock, D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., . . . Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, 31(1), 4-19.
- Gierl, M. J., & Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment.

- Measurement: Interdisciplinary Research and Perspectives*, 6(4), 263-268.
<https://doi.org/10.1080/15366360802497762>
- Gierl, M. J., Leighton, J. P., & Hunka, S. M. (2000). An NCME instructional module on exploring the logic of Tatsuoka's rule-space model for test development and analysis. *Educational Measurement: Issues and Practice*, 19, 34-44.
<https://doi.org/10.1111/j.1745-3992.2000.tb00036.x>
- Gierl, M. J., Leighton, J. P., & Hunka, S. M. (2007). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 242-274). New York, NY: Cambridge University Press.
- Gierl, M. J., Roberts, M., Alves, C., & Gotzmann, A. (2009, April). *Using judgments from content specialists to develop cognitive models for diagnostic assessments*. Paper presented at the annual meeting of National Council on Measurement in Education, San Diego, CA.
- Gomez, R. (2008). Parent ratings of the ADHD items of the disruptive behavior rating scale: Analyses of their IRT properties based on the generalized partial credit model. *Personality and Individual Differences*, 45, 181-186.
- Gorin, J. S. (2009). Commentaries on issue 6(4): "Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art," Rupp & Templin. *Measurement*, 7(30-34), 31-33.
<https://doi.org/10.1080/15366360802715387>
- Groß, J., & George, A. C. (2014). On permissible attribute classes in noncompensatory cognitive diagnosis models. *Methodology: European Journal of Research*

- Methods for the Behavioral and Social Sciences*, 1-8.
<https://doi.org/10.1027/1614-2241/a000079>
- Guest, G. (2000). Using Guttman scaling to rank wealth: Integrating quantitative and qualitative data. *Field Methods*, 12(4), 346-357.
<https://doi.org/10.1177/1525822X0001200406>
- Guttman, L. A. (1944). A basis for scaling qualitative data. *American Sociological Review*, 91, 139-150.
- Guttman, L. A. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. A. Guttman, & E. A. Schuman (Authors), *Studies in social psychology in world war II: Measurement and prediction* (Vol. 4). Princeton, NJ: Princeton University Press.
- Haberman, S. J., von Davier, M., & Lee, Y.-H. (2008, August). *Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous ability distributions* (Research Report No. ETS RR-08-45). Princeton, NJ: Educational Testing Service.
- Hafner, J., & Hafner, P. (2010). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International Journal of Science Education*, 25(12), 1509-1528. <https://doi.org/10.1080/0950069022000038268>
- Harsch, C., & Martin, G. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing*, 17, 228-250.

- Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign, Urbana-Champaign, IL.
- Hawkey, R., & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing, 9*, 122-159.
- Heartel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*(4), 301-321.
- Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement, 29*, 262-277.
- Henson, R., Templin, J., Douglas, J., & He, X. (2008). Cognitive diagnostic attribute-level discrimination indices. *Applied Psychological Measurement, 32*, 275-288.
- Hoffman, L., Xiangdong, Y., Bovaird, J. A., & Embretson, S. E. (2006). Measuring attentional ability in older adults: Development and psychometric evaluation of DriverScan. *Faculty Publications, Department of Psychology, University of Nebraska - Lincoln, 437*, 984-1000.
- Hofmann, R. J. (1979). On testing a Guttman scale for significance. *Educational and Psychological Measurement, 39*, 297-301.
- Hong, H., Wang, C., Lim, Y. S., & Douglas, J. (2015). Efficient models for cognitive diagnosis with continuous and mixed-type latent variables. *Applied Psychological Measurement, 39*(1), 31-43. <https://doi.org/10.1177/0146621614524981>
- Hu, J., Miller, M. D., Huggins-Manley, A. C., & Chen, Y.-H. (2016). Evaluation of model fit in cognitive diagnosis models. *International Journal of Testing, 16*(2), 119-141. <https://doi.org/10.1080/15305058.2015.1133627>

- Huff, K., & Goodman, D. P. (2007). The demand for cognitive assessment. In J. P. Leighton (Ed.), *Cognitive diagnostic assessment for education* (pp. 19-60). New York, NY: Cambridge University Press.
- Immekus, J. C., & Imbrie, P. K. (2008). Dimensionality assessment using the full-information item bifactor analysis for graded response data. *Educational and Psychological Measurement, 68*(4), 695-709.
- Jang, E. E., Dunlop, M., Wagner, M., Kim, Y.-H., & Gu, Z. (2013). Elementary school ELLs' reading skill profiles using cognitive diagnosis modeling: Roles of length of residence and home language environment. *Language Learning, 63*(3), 400-436. <https://doi.org/10.1111/lang.12016>
- Janssen, G., Meier, V., & Trace, J. (2015). Building a better rubric: Mixed methods rubric revision. *Assessing Writing, 26*, 51-66. <https://doi.org/10.1016/j.asw.2015.07.002>
- Junker, B. (1999, November 30). *Some statistical models and computational methods that may be useful for cognitively-relevant assessment*. Unpublished manuscript, Carnegie Mellon University, Pittsburgh, PA.
- 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.
- Jurich, D. P., & Bradshaw, L. P. (2014). An illustration of diagnostic classification modeling in student learning outcomes assessment. *International Journal of Testing, 14*, 49-72. <https://doi.org/10.1080/15305058.2013.835728>

- Kim, A.-Y. (2014). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing, 32*(2), 227-258. <https://doi.org/10.1177/0265532214558457>
- Klein, R. (2015, September 3). More students are taking the SAT, Even as scores fail to improve [Newsgroup post]. Retrieved from The Huffington Post website: http://www.huffingtonpost.com/entry/2015-sat-results_us_55e751c6e4b0c818f61a56ce
- Kolanowski, A., Hoffman, L., & Hofer, S. M. (2007). Concordance of self-report and informant assessment of emotional well-being in nursing home residents with dementia. *Journal of Gerontology, 62B*(1), 20-27.
- Koretz, D. (1988). Arriving at Lake Wobegon: Are standardized tests exaggerating achievement and distorting instruction? *American Educator, 12*(2), 8-15-46-52.
- Kullback, S., & Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*(1), 79-86.
- Kunina-Habenicht, O., Rupp, 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*, 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>

- Kuo, B.-C., Chen, C.-H., Yang, C.-W., & Mok, M. M. C. (2016). Cognitive diagnostic models for tests with multi-choice and constructed-response items. *Educational Psychology, 36*(6), 1115-1133. <https://doi.org/10.1080/01443410.2016.1166176>
- Kuo, B.-C., Pai, H.-S., & de la Torre, J. (2016). Modified cognitive diagnostic index and modified attribute level discrimination index for test construction. *Applied Psychological Measurement, 40*(5), 315-330. <https://doi.org/10.1177/0146621616638643>
- Lee, J., & Corter, J. E. (2011). Diagnosis of subtraction bugs using Bayesian networks. *Applied Psychological Measurement, 31*(1), 27-47. <http://dx.doi.org/10.1177/0146621610377079>
- Lee, Y.-S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. National Sample Using the TIMSS 2007. *International Journal of Testing, 11*, 144-177. <https://doi.org/10.1080/15305058.2010.534571>
- Lehman, A. F. (1988). A quality of life interview for the chronically mentally ill. *Evaluation and Program Planning, 11*, 51-62.
- Lehrfeld, J. M., Muller, E., & Zahner, D. (2017). *Value-added modeling without SAT/ACT scores*. Paper presented at National Council on Measurement in Education Annual Meeting, San Antonio, TX.
- 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. <http://dx.doi.org/10.1177/0146621616647954>

- Leighton, J. P., & Gierl, M. J. (Eds.). (2007a). *Cognitive diagnostic assessment for education: Theory and applications*. New York, NY: Cambridge University Press.
- Leighton, J. P., & Gierl, M. J. (2007b). Verbal reports as data for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Authors), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 146-172). New York, NY: Cambridge University Press.
- Leighton, J. P., & Gierl, M. J. (2007c). Why cognitive diagnostic assessment? In J. P. Leighton & M. J. Gierl (Authors), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 3-18). New York, NY: Cambridge University Press.
- Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*, 41(3), 205-237.
- Leighton, J. P., Gokiert, R. J., Cor, M. K., & Heffernan, C. (2010). Teacher beliefs about the cognitive diagnostic information of classroom - versus large - scale tests: Implications for assessment literacy. *Assessment in Education: Principles, Policy & Practice*, 17(1), 7-21. <https://doi.org/10.1080/09695940903565362>
- Li, H., Hunter, C. V., & Lei, P.-W. (2015). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing*, 33(3), 391-409. <https://doi.org/10.1177/0265532215590848>
- Li, H., & Suen, H. K. (2013). Constructing and validating a Q-matrix for cognitive diagnostic analyses of a reading test. *Educational Assessment*, 18, 1-25. <https://doi.org/10.1080/10627197.2013.761522>

- Li, Y., & Baser, R. (2012). Using R and WinBUGS to fit a generalized partial credit model for developing and evaluating patient-reported outcomes assessments. *Statistics in Medicine, 31*, 2010-2026.
- Liang, T., & Wells, C. S. (2009). A model fit statistic for generalized partial credit model. *Educational and Psychological Measurement, 69*(6), 913-928.
- Linacre, J. (2010). FACETS [Computer software on CD-ROM]. Chicago, IL: MESA Press.
- Linacre, J. M. (1989). *Many-facet rasch measurement* (2nd ed.). Chicago, IL: MESA Press.
- Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.
- Liu, R., & Huggins-Manley, A. C. (2016). The specification of attribute structures and its effects on classification accuracy in diagnostic test design. In L. A. Van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & S.-M. Chow (Eds.), *Quantitative psychology research* (pp. 243-254). New York, NY: Springer.
https://doi.org/10.1007/978-3-319-38759-8_18
- Liu, R., Huggins-Manley, A. C., & Bradshaw, L. (2016). The impact of Q-matrix designs on diagnostic classification accuracy in the presence of attribute hierarchies. *Educational and Psychological Measurement, 1-21*.
<http://dx.doi.org/10.1177/0013164416645636>
- Liu, R., Huggins-Manley, A. C., & Bulut, O. (2017). Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educational and Psychological Measurement, 1-27*.

- MacReady, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, 2(2), 99-120.
- 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>
- Maris, E. (1995). Psychometrik latent response models. *Psychometrika*, 60(4), 523-547.
- Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, 64(2), 187-212.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174.
- Masters, G. N. (1988). The analysis of partial credit scoring. *Applied Measurement in Education*, 1(3), 279-297.
- Mathison, S., & Ross, E.W. (2004). *Defending public schools: The nature and limits of standards-based reform and assessment*. Westport, CT: Praeger.
- McDonald, R. P., & Mok, M. M.-C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, 30(1), 23-40.
- McNamara, T. F. (1995). *Measuring second language performance*. Harlow, UK: Addison Wesley Longman.
- Meier, V. (2012). Evaluating rater and rubric performance on a writing placement exam. *Second Language Studies*, 31(1), 47-101.
- Mislevy, R. J. (1994, March). *Probability-based inference in cognitive diagnosis*. Princeton, NJ: Educational Testing Service.

- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-62.
- Mislevy, R. J., Steinberg, L. S., Almond, R. G., & Lukas, J. F. (2006). Concepts, terminology, and basic models of evidence-centered design. In D. M. Williamson, I. I. Bejar, & R. J. Mislevy (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 15-48). Mahwah, NJ: Erlbaum.
- Muraki, E. (1992, January). *A generalized partial credit model: Application of an EM algorithm* (Research Report No. RR-92-6). Princeton, NJ: Educational Testing Service.
- Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement, 17*(4), 351-363.
- Muraki, E., & Bock, R. D. (2003). Parscale (Version 4) [Computer software on CD-ROM]. Chicago, IL: Scientific Software International.
- Nering, M. L., & Ostini, R. (Eds.). (2010). *Handbook of polytomous item response theory models*. New York, NY: Routledge.
- Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research, 64*, 575-603.
- Nichols, P. D., Kobrin, J. L., Lai, M., & Koepfler, J. (2016). [The role of theories of learning and cognition in assessment design and development]. In A. A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 297-327). Chichester, England: Wiley-Blackwell.

- Organization for Economic Co-operation and Development (OECD). (2004). *Learning for tomorrow's world: First results from PISA 2003*. Paris, France: Author.
- Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9*, 129-144.
- Penfield, R. D., Myers, N. D., & Wolfe, E. W. (2008). Methods for assessing item, step, and threshold invariance in polytomous items following the partial credit model. *Educational and Psychological Measurement, 68*(5), 717-733.
- Peregrine, P. N., Ember, C. R., & Ember, M. (2004). Universal patterns in cultural evolution: An empirical analysis using Guttman scaling. *American Anthropologist, 106*(1), 145-149.
- Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment, 34*(8), 782-799.
<https://doi.org/10.1177/0734282915623053>
- Ravand, H., Barati, H., & Widhiarso, W. (2013). Exploring diagnostic capacity of a high stakes reading comprehension test: A pedagogical demonstration. *Iranian Journal of Language Testing, 3*(1), 11-37.
- Ravand, H., & Robitzsch, A. (2015). Cognitive diagnostic modeling using R. *Practical Assessment, Research & Evaluation, 20*(11), 1-12.
- R Core Team. (2016). R: A language and environment for statistical computing [Digital File]. Vienna, Austria: R Foundation for Statistical Computing.
- Reddy, Y. M., & Andrade, H. (2010). Developing the theory of formative assessment. *Assessment & Evaluation in Higher Education, 35*(4), 435-448.
<https://doi.org/10.1080/02602930902862859>

- Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analysis. *Journal of Statistical Software*, 17(5), 1-25.
- Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2017). CDM: Cognitive diagnosis modeling (Version 5.5-21) [R package].
- Rupp, A. A., & Templin, J. (2008a). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Education and Psychological Measurement*, 68, 78-96.
- Rupp, A. A., & Templin, J. (2008b). Unique characteristics of cognitive diagnosis models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives*, 6, 219-262.
- Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement*. New York, NY: The Guilford Press.
- Rupp, A. A., & Templin, J. L. (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>
- Samejima, F. (1969). *Psychometric Monograph: Vol. 17. Estimation of latent ability using a response pattern of graded scores*. Richmond, VA: Psychometric Society.
- Schultz, D. G., & Siegel, A. I. (1961). Generalized Thurstone and Guttman scales for measuring technical skills in job performance. *Journal of Applied Psychology*, 45(3), 137-142.
- Schwarzer, G. (1976). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.

- Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Newbury Park, CA: SAGE Publications.
- Sengupta, P., & Chaudhuri, H. R. (2017). An alternative analysis of scale data: A marketing application. *Global Business Review, 18*(1), 163-180.
- Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement, 47*(2), 150-174.
- Sinharay, S. (2014). Analysis of added value of subscores with respect to classification. *Journal of Educational Measurement, 51*(2), 212-222.
<https://doi.org/10.1111/jedm.12043>
- Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic models. *Educational and Psychological Measurement, 67*(2), 239-257.
- Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice, 30*(3), 29-40.
- Steedle, J. T. (2012). Selecting value-added models for postsecondary institutional assessment. *Assessment & Evaluation in Higher Education, 37*(6), 637-652.
<https://doi.org/10.1080/02602938.2011.560720>
- Stufflebeam, D. L., & Coryn, C. L.S. (2014). *Evaluation theory, models, and applications*. San Francisco, CA: Jossey-Bass.
- Svetina, D., Dai, S., & Wang, X. (2017). Use of cognitive diagnostic model to study differential item functioning in accommodations. *Behaviormetrika, 44*(2), 313-349.

- Svetina, D., Gorin, J. S., & Tatsuoka, K. K. (2011). Defining and comparing the reading comprehension construct: A cognitive-psychometric approach. *International Journal of Testing, 11*(1), 1-23.
- Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society Applied Statistics Series C, 51*(3), 337-350.
- 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.
- Tatsuoka, K. K. (1991). *Boolean algebra applied to determination of universal set of knowledge states* (Research Report No. RR-91-44-ONR). Princeton, NJ: Educational Testing Service.
- Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327-360). Mahwah, NJ: Erlbaum.
- Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. New York, NY: Routledge.
- Tatsuoka, K. K., Corter, J. E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal, 41*(4), 901-926.
- Tatsuoka, K. K., & Tatsuoka, M. M. (1989). Rule space. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (pp. 217-220). New York: Wiley.
- Templin, J. (2006). *CDM user's guide*. Unpublished manuscript.

- Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79*(2), 317-339. <https://doi.org/10.1007/s11336-013-9362-0>
- Templin, J. L., & 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. L., Henson, R. A., Templin, S. E., & Roussos, L. (2008). Robustness of hierarchical modeling of skill association in cognitive diagnosis models. *Applied Psychological Measurement*, *32*(7), 559-574. <https://doi.org/10.1177/0146621607300286>
- Thissen, D. (2003). Multilog (Version 7) [Computer software on CD-ROM]. Chicago, IL: Scientific Software International.
- Thissen, D. (2016). Bad questions: An essay involving item response theory. *Journal of Educational and Behavioral Statistics*, *41*(1), 81-89. <https://doi.org/10.3102/1076998615621300>
- Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*(4), 567-577.
- Tu, D., Zheng, C., Cai, Y., Gao, X., & Wang, D. (2017). A polytomous model of cognitive diagnostic assessment for graded data. *International Journal of Testing*, 1-22. <https://doi.org/10.1080/15305058.2017.1396465>
- Tutz, G. (1997). [Sequential models for ordered responses]. In W. J. Van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 139-152). New York, NY: Springer.

U.S. Department of Education. (2004, September 16). NCLB: Stronger accountability: Testing for results: Helping families, schools and communities understand and improve student achievement. Retrieved March 2, 2018, from U.S. Department of Education website:

<https://www2.ed.gov/nclb/accountability/ayp/testingforresults.html>

Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills. (2007). In M. J. Gierl, J. P. Leighton, & S. M. Hunka (Authors), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 242-274). New York, NY: Cambridge University Press. (Excerpted from *Cognitive diagnostic assessment for education: Theory and applications*, pp. 242-274, by J. P. Leighton & M. J. Gierl, Ed., 2007, New York, NY: Cambridge University Press)

van der Linden, W. J. (2012). On compensation in multidimensional response modeling. *Psychometrika*, 77(1), 21-30. <https://doi.org/10.1007/S11336-011-9237-1>

VanderVeen, A., Huff, K., Gierl, M., McNamara, D. S., Louwerse, M., & Graesser, A. (2007). Developing and validating instructionally relevant reading competency profiles measured by the critical reading section of the SAT Reasoning Test™. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 137-172). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Vimalraj Kumar, N., Mathialagan, P., & Sabarathnam, V. (2016). Developing a Guttman scale for measuring the degree of empowerment of rural women. *International Journal of Applied Research*, 2(3), 195-201.

- von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Research Report No. RR-05-16). Princeton, NJ: Educational Testing Service.
- von Davier, M. (2009). Some notes on the reinvention of latent structure models as diagnostic classification models. *Measurement: Interdisciplinary Research and Perspectives*, 7(1), 67-74. <https://doi.org/10.1080/15366360902799851>
- von Davier, M. (2014). The DINA model as a constrained general diagnostic model: Two variants of a model equivalency. *British Journal of Mathematical and Statistical Psychology*, 67, 49-71.
- von Davier, M., & Haberman, S. J. (2014). Hierarchical diagnostic classification models morphing into unidimensional 'diagnostic' classification models - A commentary. *Psychometrika*, 79(2), 340-346. <https://doi.org/10.1007/S11336-013-9363-Z>
- von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial-credit model. *Applied Psychological Measurement*, 28(6), 389-406.
- Wilson, M., & Iventosch, L. (1988). Using the partial credit model to investigate responses to structured subtests. *Applied Measurement in Education*, 1(4), 319-334.
- Wolf, K. (2007). The role of rubrics in advancing and assessing student learning. *The Journal of Effective Teaching*, 7(1), 3-14.
- Xu, G., & Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika*, 81(3), 625-649. <https://doi.org/10.1007/s11336-015-9471-z>
- Yang, F. M., & Kao, S. T. (2014). Item response theory for measurement validity. *Shanghai Archives of Psychiatry*, 26(3), 171-177.

- 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.
- Zahner, D. (2013). *Reliability and validity of the CLA+*. New York, NY: Council for Aid to Education.
- Zahner, D. (2014, December). *Standard-setting study final report*. New York, NY: Council for Aid to Education.
- Zahner, D., & James, J. K. (2016). *Predictive validity of a critical thinking assessment for post-college outcomes*. Paper presented at Annual Meeting of the American Educational Research Association, Washington, DC.
- Zahner, D., Ramsaran, L. M., & Steedle, J. T. (2012). *Comparing alternatives in the prediction of college success*. Paper presented at Annual Meeting of the American Educational Research Association, Vancouver, Canada.

Appendix I. The CLA+ Scoring Rubric

CLA+ Scoring Rubric	1	2	3	4	5	6
<p>Analysis and Problem Solving</p> <p>Making a logical decision or conclusion (or taking a position) and supporting it by utilizing appropriate information (facts, ideas, computed values, or salient features) from the Document Library</p>	<p>May state or imply a decision/conclusion/ position</p> <p>Provides minimal analysis as support (e.g., briefly addresses only one idea from one document) or analysis is entirely inaccurate, illogical, unreliable, or unconnected to the decision/conclusion/ position</p>	<p>States or implies a decision/conclusion/ position</p> <p>Provides analysis that addresses a few ideas as support, some of which is inaccurate, illogical, unreliable, or unconnected to the decision/conclusion/ position</p>	<p>States or implies a decision/conclusion/ position</p> <p>Provides some valid support, but omits or misrepresents critical information, suggesting only superficial analysis and partial comprehension of the documents</p> <p>May not account for contradictory information (if applicable)</p>	<p>States an explicit decision/conclusion/ position</p> <p>Provides valid support that addresses multiple pieces of relevant and credible information in a manner that demonstrates adequate analysis and comprehension of the documents; some information is omitted</p> <p>May attempt to address contradictory information or alternative decisions/ conclusions/ positions (if applicable)</p>	<p>States an explicit decision/conclusion/ position</p> <p>Provides strong support that addresses much of the relevant and credible information in a manner that demonstrates very good analysis and comprehension of the documents</p> <p>Reduces contradictory information or alternative decisions/ conclusions/ positions (if applicable)</p>	<p>States an explicit decision/conclusion/ position</p> <p>Provides comprehensive support, including nearly all of the relevant and credible information, in a manner that demonstrates outstanding analysis and comprehension of the documents</p> <p>Thoroughly refutes contradictory evidence or alternative decisions/ conclusions/ positions (if applicable)</p>
<p>Writing Effectiveness</p> <p>Constructing organized and logically cohesive arguments. Strengthening the writer's position by providing elaboration on facts or ideas (e.g., explaining how evidence bears on the problem, providing examples, and emphasizing especially convincing evidence)</p>	<p>Does not develop convincing arguments; writing may be disorganized and confusing</p> <p>Does not provide elaboration on facts or ideas</p>	<p>Provides limited, invalid, over-stated, or very unclear arguments; may present information in a disorganized fashion or undermine own points</p> <p>Any elaboration on facts or ideas tends to be vague, irrelevant, inaccurate, or unreliable (e.g., based entirely on writer's opinion); sources of information are often unclear</p>	<p>Provides limited or somewhat unclear arguments. Presents relevant information in each response, but that information is not woven into arguments</p> <p>Provides elaboration on facts or ideas a few times, some of which is valid; sources of information are sometimes unclear</p>	<p>Organizes response in a way that makes the writer's arguments and logic of those arguments apparent but not obvious</p> <p>Provides valid elaboration on facts or ideas several times and cites sources of information</p>	<p>Organizes response in a logically cohesive way that makes it fairly easy to follow the writer's arguments</p> <p>Provides valid elaboration on facts or ideas related to each argument and cites sources of information</p>	<p>Organizes response in a logically cohesive way that makes it very easy to follow the writer's arguments</p> <p>Provides valid and comprehensive elaboration on facts or ideas related to each argument and clearly cites sources of information</p>
<p>Writing Mechanics</p> <p>Demonstrating facility with the conventions of standard written English (agreement, tense, capitalization, punctuation, and spelling) and control of the English language, including syntax (sentence structure) and diction (word choice and usage)</p>	<p>Demonstrates minimal control of grammatical conventions with many errors that make the response difficult to read or provides insufficient evidence to judge</p> <p>Writes sentences that are repetitive or incomplete, and some are difficult to understand</p> <p>Uses simple vocabulary, and some vocabulary is used inaccurately or in a way that makes meaning unclear</p>	<p>Demonstrates poor control of grammatical conventions with frequent minor errors and some severe errors</p> <p>Consistently writes sentences with similar structure and length with some may be difficult to understand</p> <p>Uses simple vocabulary, and some vocabulary may be used inaccurately or in a way that makes meaning unclear</p>	<p>Demonstrates fair control of grammatical conventions with frequent minor errors</p> <p>Writes sentences that read naturally but tend to have similar structure and length</p> <p>Uses vocabulary that communicates ideas adequately but lacks variety</p>	<p>Demonstrates good control of grammatical conventions with few errors</p> <p>Writes well-constructed sentences with some varied structure and length</p> <p>Uses vocabulary that clearly communicates ideas but lacks variety</p>	<p>Demonstrates very good control of grammatical conventions</p> <p>Consistently writes well-constructed sentences with varied structure and length</p> <p>Uses varied and sometimes advanced vocabulary that effectively communicates ideas</p>	<p>Demonstrates outstanding control of grammatical conventions</p> <p>Consistently writes well-constructed complex sentences with varied structure and length</p> <p>Displays adept use of vocabulary that is precise, advanced and varied</p>

Appendix II. Sample CLA+ Assessment



College Sample Assessment

OVERVIEW

CLA+ comprises a Performance Task (PT) and a Selected-Response Question (SRQ) section. There are three types of questions in the SRQ section: Scientific and Quantitative Reasoning (SQR), Critical Reading and Evaluation (CRE), and Critique-an-Argument (CA). CLA+ is administered online. The PT contains an open-ended prompt that requires written responses. The SRQs ask the student to choose the best response based on the Document Library provided.

CLA+ tasks are designed to assess students' general critical-thinking and written-communication skills, regardless of their academic concentrations. These skills include scientific and quantitative reasoning, analytic reasoning and evaluation of information, problem solving, writing effectiveness, and writing mechanics. These skills are necessary, not only for success in high school and college; they are important for success in the workplace and other aspects of life outside the classroom. No prior knowledge of any particular field is necessary in order to perform well.

What is presented in the practice example is an abbreviated version of a PT and of SRQs. Nevertheless, please familiarize yourself with how the assessment includes real-world scenarios and a series of documents that reflect an authentic situation.

This example is also intended to demonstrate what is expected in a high-quality response. The sample response demonstrates the student's critical-thinking and written-communication skills.

Table of Contents

- 1 Overview
- 2 Performance Task
- 3 Document 1 | SportsCo Profile
- 4 Document 2 | Daily News Story
- 5 Document 3 | Incident Report
- 6 Document 4 | Interview Transcript
- 8 Document 5 | NCSA Bulletin
- 9 Document 6 | Advertising Storyboard
- 10 Document 7 | Blog Post
- 11 Answer Sheet
- 12 Sample Response 1
- 13 Sample Response 2
- 14 Sample Response 3
- 15 CLA+ Scoring Rubric
- 16 SQR Document
- 17 SQR Questions
- 18 CRE Document 1
- 19 CRE Document 2
- 20 CRE Questions
- 21 CA Document
- 21 CA Questions

Additional Information

The CLA+ is an online assessment. For more information about the CLA+, please visit cae.org/cla.

You may also email the CLA+ Team at clateam@cae.org.

PERFORMANCE TASK: SKATING ACCIDENT

INSTRUCTIONS

This is an example of a CLA+ Performance Task. In the course of this practice Performance Task, you will prepare a written response to a hypothetical but realistic situation. The Performance Task is made up of an introductory scenario, a question, and seven documents/information sources. You will use information from the Document Library in carrying out the task.

While your personal values and experiences are important, you should base your response solely on the evidence provided in the documents.

SCENARIO

You are the chief marketing officer of SportsCo, an athletic equipment company. The most profitable sector of the company is its new line of inline skates called HotSkates. Given the success of the current HotSkates advertising campaign, the company plans to continue with it for the next three months. However, after a recent skating accident in which a teenager was seriously injured, SportsCo is now receiving negative press relating to possible safety hazards associated with its products. Critics are saying that the HotSkates advertisements do not adequately convey the advanced skill level necessary to safely perform tricks on the skates. If SportsCo continues with the current campaign, it risks facing lawsuits as well as increasing negative public opinion of the company's ethical standards. However, instating a new advertising campaign will require a great deal of time and money, and the new campaign may not be as successful as the present one. **It is your job to decide whether to continue with the present ad campaign.** You have 60 minutes to complete this task.

PROMPT

Your task is to write a report for your marketing team that explains your decision on whether to continue the present campaign. You should support your position with ideas and evidence found in the documents and address potential counterarguments in your recommendation. If you choose not to continue with the current campaign, you should include recommendations for an alternative campaign. There is no "correct" answer. Your report should clearly describe all the details necessary to support your position. Your answers will be judged not only on the accuracy of the information you provide but also on how clearly the ideas are presented, how thoroughly the information is covered, how effectively the ideas are organized, and how well your writing reflects the conventions of standard written English.

While your personal values and experiences are important, please answer the question in the this task solely on the basis of the information provided above and in the Document Library.

DOCUMENT LIBRARY

- Document 1** - SportsCo Profile
- Document 2** - Daily News Story
- Document 3** - Incident Report
- Document 4** - Interview Transcript
- Document 5** - NCSA Bulletin
- Document 6** - Advertising Storyboard
- Document 7** - Blog Post

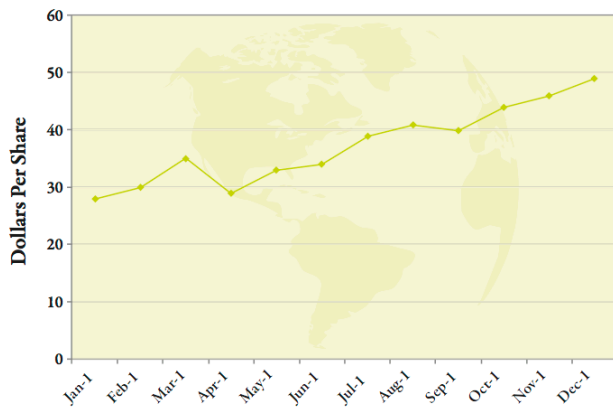


Company Profile-2014

SportsCo Manufacturing

Description: SportsCo is a diversified sporting equipment and leisure company that has grown significantly over the past decade. Founded in 1999 to produce wheels and wheel parts for the secondary bicycle market, SportsCo experienced rapid growth when the wheeled vehicle market grew in the 2000s. It broadened its product line significantly in 2007, with the acquisition of Fantam Sports. It has had its greatest success in the area of inline skating, where it holds a dominant share of the domestic market. Recent expansion into the apparel and leisure markets has netted solid returns. Investors have bid up the SportsCo share price by almost 80% during the past year.

SportsCo Share Price



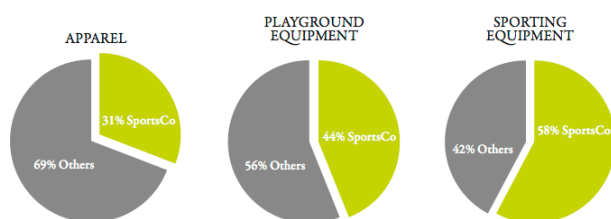
Share price increase
of almost
80%
during the past year

Leadership: SportsCo was founded by two brothers, Kyle and David Foster, who shared executive responsibility for the firm during its first few years. The Foster brothers were equally concerned about both community development and business success, and they devoted considerable effort to building a positive local environment by contributing to community projects. With the acquisition of Fantam Sports, the Foster brothers sought more experienced leadership and brought in Mitch Hennessey as chief executive officer (CEO). Hennessey has guided SportsCo to its current

GLOBAL FINANCIAL SOLUTIONS

success through conservative management coupled with aggressive, creative marketing. The Foster brothers' commitment to community continues in SportsCo's active involvement in community organizations. SportsCo is a major sponsor of the Junior Special Olympics for children with disabilities, and the company donates sporting goods equipment to inner-city schools throughout the country. CEO Hennessey serves on the board of "All Children Matter," an organization concerned with children from abusive homes.

Business Units: SportsCo has three major divisions: apparel, sporting equipment, and playground equipment. Each of the three divisions has a substantial share of the domestic market, but the sporting equipment unit remains the company's largest in terms of market share and total revenue.



Sporting equipment generated more than half of the company's revenue in 2013. The division has six operating units that focus on specific sectors of the U.S. market. SportsCo is the dominant manufacturer of skating equipment in the US, and they are among the largest firms in the market for fishing and boating and competitive team sports equipment. Their newer units have been solid performers but are still focused on niche markets.

Sporting Equipment Sector

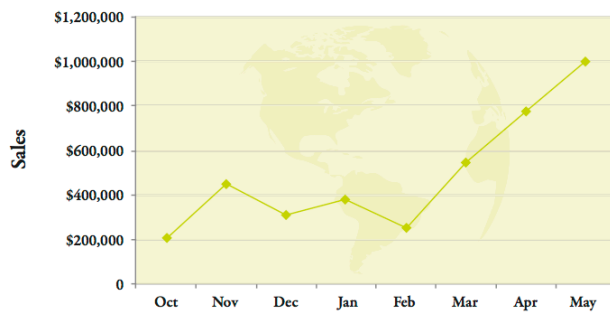
Team Sports	27%
Fishing and Boating	37%
Skating	52%
Bicycling	16%
Exercise Fitness	9%
Skiing	17%

GLOBAL FINANCIAL SOLUTIONS

Overall Growth: SportsCo is well managed, well positioned, and analysts predict continued growth.

Recent Developments: In 2013, SportsCo began manufacturing a new line of high-performance skates called HotSkates to market alongside its more traditional inline skates, StreetSkates. After releasing a new advertising campaign in March 2014, SportsCo saw a significant increase in the sales of HotSkates. This successful new campaign features exciting commercials targeted at children and young teenagers. Given the increase in sales following the launch of this campaign, the company plans to continue producing similar commercials with the same marketing platform going forward.

HotSkates Sales 2013-2014



NEW
PRODUCT

KENSINGTON DAILY NEWS

Kensington, Missouri

Friday, May 16

Local High School Student Paralyzed in Skating Accident

Fourteen-year-old Kyle Clester was paralyzed on Tuesday in a skating accident at Burroughs High School. Clester was found lying on the ground at about 4:00 p.m. by a school custodian. He had apparently been skating on school grounds when the accident occurred. Custodian Brad Steffens, who has worked at the school for five years, said he found Clester at the bottom of a flight of stairs leading to the football field. His helmet was lying next to him on the ground. According to Steffens, the boy was wearing HotSkates, which are the latest craze in inline skating. The custodian said he thought the boy must have fallen while trying to jump the stairs.

Clester was taken to Memorial Hospital where his condition is listed as serious. Hospital sources said the boy appears to be paralyzed, but it is too soon to tell whether the condition is permanent. Clester's parents said they continually warned Kyle to be careful when using the skates. They bought him elbow and wrist guards and required him to wear his helmet whenever he skated. His mother said, "I've heard about so many kids who have had accidents while wearing HotSkates that I didn't want to buy him a pair. But



source: http://commons.wikimedia.org/wiki/File:Roxa_Xtreme.jpg

he kept saying that HotSkates were what he wanted for his birthday, so eventually caved. Even so, I was scared every time he used them." Kyle turned 14 last month.

A nurse in the emergency room told reporters that the number of skating accidents has increased over the past few months, and more often the injuries involve HotSkates. "I'd say that half of the skating accidents we've seen involved these new HotSkates," the

nurse told reporters.

Shelly Banks, spokesperson for SportsCo, which manufactures HotSkates, refused to comment on this incident. "We stand behind the safety of our products," she said. "Our safety precautions exceed all industry standards."

One of Clester's friends said the boy had been practicing extra hard to perfect a trick he saw in a commercial for HotSkates. ■



INCIDENT REPORT

NAME: Brad Steffens	DATE OF REPORT: May 15, 2014
DATE OF INCIDENT: May 14, 2014	TIME OF INCIDENT: 4:30 pm

LOCATION: Steps to football field

DESCRIBE THE INCIDENT:

I saw a kid lying at the bottom of the steps to the field. I ran outside and asked him what was wrong. He said he couldn't move or feel anything from the neck down. Nobody else was around. I ran back inside and called 911. The paramedics came right away. I gave them the helmet that was on the grass near the steps. After a while, they took the kid away.

IF ANYONE WAS INJURED, DESCRIBE WHO IT WAS AND WHAT THEY WERE DOING:

I don't know the kid who got hurt. He had one of those crazy haircuts and was wearing skates and a T-shirt with somebody's face on it. I don't know how he got hurt. Probably from skating.

DESCRIBE WHAT WERE YOU DOING PRIOR TO THE INCIDENT:

Checking that all the doors and windows in the main building were closed.

ADDITIONAL INFORMATION:

Around 3 o'clock, I saw some kids skating on the handrails and benches, so I told them to leave like I always do. I usually let kids skate on the cement patio if they want to because it's wide open, but they have to wear a helmet. I never let them on the handrails or benches. I think the kid who got hurt was one of the ones I saw earlier in the afternoon. Not sure.

Transcript of Interview with Heather McKinley, May 24
 Institute for Consumer Protection



TN: We are talking today with Heather McKinley, research director of the Institute for Consumer Protection, about roller skates. Welcome to the show, Heather.



HM: Thank you. It is a pleasure to be here.



TN: When I was growing up, skates had side-by-side wheels; now the wheels are one behind the other. Is this better?



HM: It depends on what you mean by better. Once you learn to use them, the new inline skates are faster and more maneuverable than the side-by-side, four-wheel roller skates you grew up with. But they are harder to learn.



TN: Does that mean more accidents?



HM: We are certainly seeing more skating injuries every year, but we are also seeing much more skating.



TN: Are these inline skates dangerous?



HM: Definitely. One of the advantages of side-by-side four-wheel skates is that they give you stable contact with the ground. There is less lateral pressure on your feet and ankles.



TN: In simpler terms, please.



HM: Inline skates tip from side to side. Roller skates don't. The only thing that keeps inline skates upright is your balance and the strength of your ankles.



TN: Are there more ankle injuries with inline skates?



HM: Definitely. Doctors call them "the orthopedic surgeon's friend" because they are associated with so many broken ankles, wrists, and arms.



TN: Why wrists and arms?



HM: Because people skate so fast that they can't keep their balance. And when they fall, they reach out to protect themselves and end up breaking an arm or wrist.



TN: So speed is part of the problem.



HM: Absolutely. Speed and stability.



TN: What about the new generation of inline skates, such as HotSkates?



HM: These skates are faster and narrower than earlier versions, and they are more dangerous as a result.



TN: How do they do that? Do they use only one wheel?



HM: No, they employ new space-age bearings that have less friction, allowing the wheels to turn faster. Also, they use new synthetic materials that permit narrower wheels for more maneuverability but less stability.



TN: So, are they safe?



HM: Not for beginners. You go faster, so if you do fall, you are likely to have a more serious injury. I inline skate myself, but I know my limits. HotSkates are too fast for me, and, I suspect, for most children.

Skaters Beware: Serious Injury Rates with Inline Skates

Sporting King, the nation's leading retailer of inline skates, provided the National Consumer Safety Association (NCSA) with sales figures and customer lists for all the inline skates it sold between October 31, 2013 and April 30, 2014. These data indicate that 60% of the skates Sporting King sold during this period were manufactured by SportsCo, and the remaining 40% were manufactured by AXM.

Both SportsCo and AXM make traditional inline skates as well as newer high-performance inline skates that are faster and more expensive. SportsCo and AXM are currently the only two manufacturers of these new high-performance skates.

We surveyed a stratified random sample of 8,200 Sporting King customers who purchased SportsCo and AXM inline skates during the October 31, 2013 and April 30, 2014 period. The survey included questions about the skaters' experience and skill level, the frequency of skate use, and the frequency of accidents and injuries. Skate purchasers who returned completed surveys to NCSA by May 15, 2014 received a store gift certificate worth \$15.

This report is based on the 3,884 completed surveys (47.4%) that NCSA received by May 15, 2014. SportsCo and AXM inline skate purchasers had nearly identical response rates (47.3% and 47.4% respectively).

Table 1 shows the number of Sporting King customers that were surveyed compared to the total customers who purchased SportsCo and AXM inline skates.

Table 1. Number of Sporting King customers completing the survey who purchased SportsCo and AXM skates.

Company	SportsCo		AXM	
Model Name	StreetSkates	HotSkates	Inlyne	Inlyne Pro
Model Type	Regular	High Performance	Regular	High Performance
Responding	1613	716	1083	472

Many survey respondents reported that the person using the skates suffered one or more skating related injuries during the preceding three-month period. Injuries included abrasions and cuts, muscle strains and tears, and broken bones. This report does not include less severe injuries and, instead, focuses on the strains, tears, and breaks that required medical treatment by a physician. The numbers below are based on skaters who suffered one or more of these more serious injuries.¹

¹ No questions were asked about what caused the injury. There were too few fatalities to report reliable results by manufacturer, skate type, or experience level.

The breakdown of skaters by self-reported level of experience is shown in Table 2.

Table 2. Number of skaters at each level of experience using each type of skate.

Experience Level	SportsCo		AXM		All Skates
	Regular (StreetSkates)	High Performance (HotSkates)	Regular	High Performance	
Beginner	600	198	412	132	1342
Intermediate	527	238	350	157	1272
Advanced	486	280	321	183	1270
All Levels	1613	716	1083	472	3884

Table 3 shows the number of skaters, among the 3,884 questionnaires returned, who suffered one or more serious injuries (as defined above).

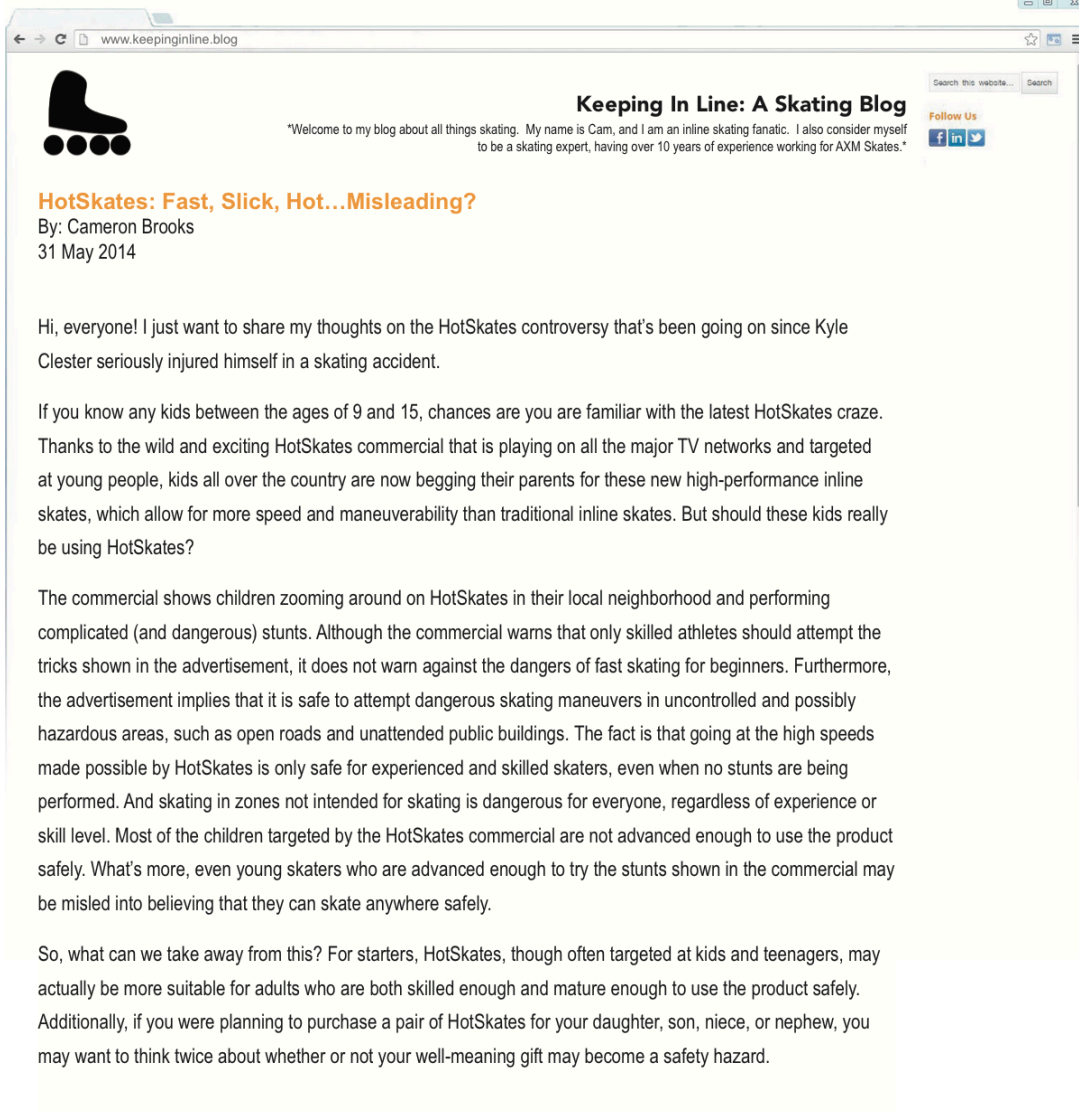
Table 3. Number of skaters with serious injuries by type of skate, experience of user, and manufacturer.

Experience Level	SportsCo		AXM		All Skates
	Regular (StreetSkates)	High Performance (HotSkates)	Regular	High Performance	
Beginner	81	38	52	26	197
Intermediate	52	24	36	16	128
Advanced	25	13	13	6	57
All Levels	158	75	101	48	382

Table 4 shows the percentage of skaters in each combination of skate type and experience level that suffered at least one "serious" injury (as defined above) during the preceding three months. For example, of the 600 beginning skaters who used StreetSkates (i.e. SportsCo's regular inline skate), 81 (13.5%) had at least one serious injury.

Table 4. Percentage of skaters with serious injuries by type of skate, experience of user, and manufacturer.

Experience Level	SportsCo		AXM		Overall
	Regular (StreetSkates)	High Performance (HotSkates)	Regular	High Performance	
Beginner	13.5%	19.2%	12.6%	19.7%	14.7%
Intermediate	9.9%	10.1%	10.3%	10.2%	10.1%
Advanced	5.1%	4.6%	4.0%	3.3%	4.5%
All Levels	9.8%	10.5%	9.3%	10.2%	9.8%



The screenshot shows a web browser window with the address bar displaying "www.keepinginline.blog". The page features a logo of a black inline skate with four wheels. The main heading is "Keeping In Line: A Skating Blog" with a sub-headline: "Welcome to my blog about all things skating. My name is Cam, and I am an inline skating fanatic. I also consider myself to be a skating expert, having over 10 years of experience working for AXM Skates." There is a search bar and social media links for Facebook, LinkedIn, and Twitter. The article title is "HotSkates: Fast, Slick, Hot...Misleading?" by Cameron Brooks, dated 31 May 2014.

HotSkates: Fast, Slick, Hot...Misleading?
By: Cameron Brooks
31 May 2014

Hi, everyone! I just want to share my thoughts on the HotSkates controversy that's been going on since Kyle Clester seriously injured himself in a skating accident.

If you know any kids between the ages of 9 and 15, chances are you are familiar with the latest HotSkates craze. Thanks to the wild and exciting HotSkates commercial that is playing on all the major TV networks and targeted at young people, kids all over the country are now begging their parents for these new high-performance inline skates, which allow for more speed and maneuverability than traditional inline skates. But should these kids really be using HotSkates?

The commercial shows children zooming around on HotSkates in their local neighborhood and performing complicated (and dangerous) stunts. Although the commercial warns that only skilled athletes should attempt the tricks shown in the advertisement, it does not warn against the dangers of fast skating for beginners. Furthermore, the advertisement implies that it is safe to attempt dangerous skating maneuvers in uncontrolled and possibly hazardous areas, such as open roads and unattended public buildings. The fact is that going at the high speeds made possible by HotSkates is only safe for experienced and skilled skaters, even when no stunts are being performed. And skating in zones not intended for skating is dangerous for everyone, regardless of experience or skill level. Most of the children targeted by the HotSkates commercial are not advanced enough to use the product safely. What's more, even young skaters who are advanced enough to try the stunts shown in the commercial may be misled into believing that they can skate anywhere safely.

So, what can we take away from this? For starters, HotSkates, though often targeted at kids and teenagers, may actually be more suitable for adults who are both skilled enough and mature enough to use the product safely. Additionally, if you were planning to purchase a pair of HotSkates for your daughter, son, niece, or nephew, you may want to think twice about whether or not your well-meaning gift may become a safety hazard.

SAMPLE RESPONSE 1

As you all are well aware, HotSkates, our highest-grossing product, has come under public attack in recent months. The high incidence of accidents and injuries in young HotSkates users has attracted this negative media attention. Because this negative publicity could very likely damage our reputation as a company and hurt sales, I have made the decision to instate a new advertising campaign to protect the reputation of our company and our products.

This decision was made based on a thorough analysis of media coverage and financial reports that reveal the need to reassess and redesign our advertising approach to marketing HotSkates. HotSkates sales have jumped nearly 1 million dollars from February to May of 2014, providing us with the revenue necessary to make the marketing changes I propose. While these changes may lead to an immediate decrease in HotSkates sales, it will allow SportsCo to maintain our reputation—which we have worked for the past 15 years to build—as a reliable and high-quality sports equipment company.

An examination of recent media discussions of HotSkates will explain the necessity of a campaign adjustment. The incident of a fourteen-year old boy paralyzed in a skating accident involving HotSkates has received considerable press, even though it's unclear whether the boy was an experienced skater, given that—according to the Daily News article—Kyle Clester had only recently received his HotSkates. He may simply not have been skating appropriately for his skill level; a beginner should not have been attempting advanced skating tricks, as was suggested by the school custodian. In the new story, however, his mother is quoted describing HotSkates as the cause of the accident. The article even cites our popular HotSkates commercial as an influence in the boy's risky skating behavior. If this were an isolated incident, it might not be cause for alarm, but this sentiment has become part of the national attitude towards our product. In a recent television interview, Heather McKinley, the research director for Consumer Protection, called our skates "dangerous" and even announced that she would not use them or recommend them for children.

The National Consumer Safety Association has released a study showing that our high-performance skates do not have a significantly higher rate of serious injury than our major competitor's skates at any experience level (a 10.5% overall serious injury rate for HotSkates, compared to 10.2% for AXM's high-performance skates). Though this report was published later than some of the other documents I have compiled on the subject, we do not know whether these statistics will affect broader public opinion. People may instead focus on the fact that our advertisement targets younger skaters, when we know from the NCSA report inexperienced skaters have much higher risk of serious injury than experienced skaters when using high-performance skates; 19% of all beginner skaters who use HotSkates experience serious injuries.

Despite the fact that we provide a disclaimer at the end of our commercial, all the negative attention portrays SportsCo as an untrustworthy or even a dishonest company. In order to keep our sales steady and rising for decades to come, it is important that we maintain the trust of our customers. To continue with our campaign might save us money in the short-term but it will not be good for the long-term public opinion of SportsCo. With the recent spike in sales, we can afford to alter our HotSkates marketing campaign. Skating equipment is the largest unit within our Sports Equipment sector and HotSkates sales have been astronomical since the launch of the ad campaign, but the negative press could seriously impact our long-term success. I propose we work on marketing HotSkates to an older, semi-professional demographic and work with the development team to produce a new skating product that is safe for beginners who want to try low-level skating tricks. It is our responsibility to our consumers and to the future of our company.

Analysis and Problem Solving

Subscore: 6

- States an explicit decision/conclusion/position
- Provides comprehensive support, including nearly all of the relevant and credible information, in a manner that demonstrates outstanding analysis and comprehension of the documents
- Thoroughly refutes contradictory evidence or alternative decisions/conclusions/positions (if applicable)

Writing Effectiveness

Subscore: 6

- Organizes response in a logically cohesive way that makes it very easy to follow the writer's arguments
- Provides valid and comprehensive elaboration on facts or ideas related to each argument and clearly cites sources of information

Writing Mechanics

Subscore: 6

- Demonstrates outstanding control of grammatical conventions
- Consistently writes well-constructed complex sentences with varied structure and length
- Displays adept use of vocabulary that is precise, advanced, and varied

SAMPLE RESPONSE 2

Dear Marketing Team,

Our product, HotSkates has brought in 1 million dollars in recent months. It is an essential product to our company and it is selling so well because kids love the fancy tricks we portray in the commercial. It is enough to protect us that we put a disclaimer at the end saying that this product is for professional experts. I have decided we should not change our campaign just because some people complain. The issue is that they dont understand safety precautions or proper training. They should learn that and read our disclaimer.

SportsCo is not the only company making High Performance inline skates that lead to injury. AXM also does. In fact we have almost the same amount of injuries. Even they have more beginners with a percentage of injuries.

Even the experts who say that this is not for kids understand that you need special skills, and protection to safely use HotSkates. In the incident report it is clear, that the boy who became paralyzed was acting unsafely. This is not the fault of our company. We have other kinds of equipment for sale for beginners. Kids could also use our regular skates called StreetSkates. The most important point of course is that HotSkates make an enormous amount of money for our company. Skating is 52% of all the equipment we sell. We can't afford to drop this campaign. It's what the kids want.

Analysis and Problem Solving

Subscore: 4

- States an explicit decision/conclusion/position
- Provides valid support that addresses multiple pieces of relevant and credible information in a manner that demonstrates adequate analysis and comprehension of the documents; some information is omitted
- May attempt to address contradictory information or alternative decisions/conclusions/positions (if applicable)

Writing Effectiveness

Subscore: 3

- Provides limited or somewhat unclear arguments. Presents relevant information in each response, but that information is not woven into arguments
- Provides elaboration on facts or ideas a few times, some of which is valid; sources of information are sometimes unclear

Writing Mechanics

Subscore: 3

- Demonstrates fair control of grammatical conventions with frequent minor errors
- Writes sentences that read naturally but tend to have similar structure and length
- Uses vocabulary that communicates ideas adequately but lacks variety

SAMPLE RESPONSE 3

I have decided to stop the campaign ads for HotSkates. We owe it to, the kids and the mother of the kid who was paralyzed to advertise in a different way and even the experts agree that inline skating leads to serious injury like it says in the report about SportsCo and AXM. We have worse numbers than they do.

The way the incident report describes the boy is really sad, and the mother's words in the newspaper. It's important that we change the ad. If you look at the ad, it makes the skating look like a lot of fun not dangerous enough. One way we could change the ad is to make it look dangerous like in reality.

SportsCo make a lot of money from skating but it's not the only place we make money from. We also make money from Team Sports, Fishing and Boating, Bicycling, Exercise Fitness, and Skiing. I think we should use more money to sell products for one of these things. It could help cover whatever it costs to change the ads and we will still probably sell skates but hopefully no kids will become paralyzed or injured like before.

Analysis and Problem Solving

Subscore: 2

- May state or imply a decision/conclusion/position
- Provides minimal analysis as support (e.g., briefly addresses only one idea from one document) or analysis is entirely inaccurate, illogical, unreliable, or unconnected to the decision/conclusion/position

Writing Effectiveness

Subscore: 2

- Does not develop convincing arguments; writing may be disorganized and confusing
- Does not provide elaboration on facts or ideas

Writing Mechanics

Subscore: 2

- Demonstrates minimal control of grammatical conventions with many errors that make the response difficult to read or provides insufficient evidence to judge
- Writes sentences that are repetitive or incomplete, and some are difficult to understand
- Uses simple vocabulary, and some vocabulary is used inaccurately or in a way that makes meaning unclear

CLA+ Scoring Rubric

Analysis and Problem Solving

Making a logical decision or conclusion (or taking a position) and supporting it by utilizing appropriate information (facts, ideas, computed values, or salient features) from the Document Library

1

May state or imply a decision/conclusion/position
Provides minimal analysis as support (e.g., briefly addresses only one idea from one document) or analysis is entirely inaccurate, illogical, unreliable, or unconnected to the decision/conclusion/position

2

States or implies a decision/conclusion/position
Provides analysis that addresses a few ideas as support, some of which is inaccurate, illogical, unreliable, or unconnected to the decision/conclusion/position

3

States or implies a decision/conclusion/position
Provides some valid support, but omits or misrepresents critical information, suggesting only superficial analysis and partial comprehension of the documents
May not account for contradictory information (if applicable)

4

States an explicit decision/conclusion/position
Provides valid support that addresses multiple pieces of relevant and credible information in a manner that demonstrates adequate analysis and comprehension of the documents; some information is omitted
May attempt to address contradiction/information or alternative decisions/conclusions/positions (if applicable)

5

States an explicit decision/conclusion/position
Provides strong support that addresses much of the relevant and credible information, in a manner that demonstrates very good analysis and comprehension of the documents
Refutes contradictory information or alternative decisions/conclusions/positions (if applicable)

6

States an explicit decision/conclusion/position
Provides comprehensive support, including nearly all of the relevant and credible information, in a manner that demonstrates outstanding analysis and comprehension of the documents
Thoroughly refutes contradictory evidence or alternative decisions/conclusions/positions (if applicable)

Writing Effectiveness

Constructing organized and logically cohesive arguments. Strengthening the writer's position by providing elaboration on facts or ideas (e.g., explaining how evidence bears on the problem, providing examples, and emphasizing especially convincing evidence)

Does not develop convincing arguments; writing may be disorganized and confusing
Does not provide elaboration on facts or ideas

Provides limited, invalid, or over-stated or very unclear information in a disorganized fashion or undermine own points
Any elaboration on facts or ideas tends to be vague, irrelevant, inaccurate, or unreliable (e.g., based entirely on writer's opinion); sources of information are often unclear

Provides limited or somewhat unclear arguments. Presents information that is not woven into arguments
Provides elaboration on facts or ideas a few times, some of which is valid; sources of information are sometimes unclear

Organizes response in a way that makes the writer's arguments apparent but not obvious
Provides valid elaboration on facts or ideas several times and cites sources of information

Organizes response in a logically cohesive way that makes it fairly clear how the writer's arguments

Provides valid and comprehensive elaboration on facts or ideas related to each argument and clearly cites sources of information

Writing Mechanics

Demonstrating facility with the conventions of standard written English (agreement, tense, capitalization, punctuation, and spelling) and control of the English language, including syntax (sentence structure) and diction (word choice and usage)

Demonstrates minimal control of grammatical conventions with many errors that make the response difficult to read or judge
Writes sentences that are repetitive or incomplete, and some are difficult to understand
Uses simple vocabulary, and some vocabulary is used inaccurately or in a way that makes meaning unclear

Demonstrates poor control of grammatical conventions with frequent minor errors and some severe errors
Consistently writes sentences with similar structure and length, and some may be difficult to understand
Uses simple vocabulary, and some vocabulary may be used inaccurately or in a way that makes meaning unclear

Demonstrates fair control of grammatical conventions with frequent minor errors
Writes sentences that read naturally but tend to have similar structure and length
Uses vocabulary that communicates ideas adequately but lacks variety

Demonstrates good control of grammatical conventions with few errors
Writes well-constructed sentences with some varied structure and length
Uses vocabulary that clearly communicates ideas but lacks variety

Demonstrates very good control of grammatical conventions
Consistently writes well-constructed sentences with varied structure and length
Uses varied and sometimes advanced vocabulary that effectively communicates ideas

Demonstrates outstanding control of grammatical conventions
Consistently writes well-constructed complex sentences with varied structure and length
Displays adept use of vocabulary that is precise, advanced, and varied

DOCUMENT : SCIENTIFIC & QUANTITATIVE REASONING

Fueling the Future

In a quest to solve the energy problems of the twenty-first century—that is, to find sustainable and renewable sources of energy that are less destructive to the environment yet economical enough to have mass appeal—scientists throughout the world are experimenting with innovative forms of fuel production. While oil is still the most common source of fuel, there is a finite amount of it, and new alternatives will become necessary to sustain the supply of energy that we are accustomed to.

Corn-based ethanol, the most common alternative to traditional fossil fuels (primarily coal, petroleum, and natural gas), is mixed into gasoline in small quantities, and it now accounts for about 10% of the fuel supply from sources within the United States. Because corn is grown on farmland, it is subject to price fluctuations based on supply and demand of the crop, as well as disruptions resulting from naturally occurring events, such as droughts and floods. At present, nearly 40% of the corn grown in the United States is used for fuel, and the demand for corn-based ethanol is rising. To meet this demand, wetlands, grasslands, and forests are all being converted into farmland with the sole intention of growing corn for more ethanol production. Corn grown for ethanol has become a more valuable commodity for farmers than crops grown for food, and this has negatively affected consumers worldwide, as shown by the increasing price of food over time.

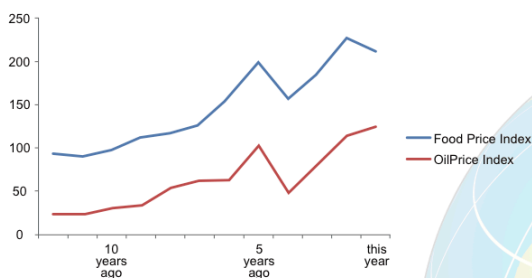


Figure 1: Food and oil price indices (based on information found at www.fao.org and www.indexmundi.com)

Another alternative that has gained attention in recent years is the harvesting of biofuel from algae. Biodiesel, a type of biofuel, is produced by extracting oil from algae, much like the process involved in creating vegetable oils from corn or soybeans. Ethanol can also be created by fermenting algae. Algae biofuel has some unique benefits that separate it from other fossil fuel alternatives. To begin with, while all fuels create

carbon dioxide when they are burned, algae have the ability to recapture and use that carbon dioxide during photosynthesis while they are growing. In this regard, the advantage is enormous. The process of growing algae actually absorbs more carbon dioxide than is released into the atmosphere when it is burned for fuel. Most manufacturing processes strive for “carbon neutrality”—or the balance between carbon emissions and depletion corresponding to a net carbon output of zero. Even better, algae-based biofuel can be described as “carbon negative.” Other forms of biofuel can make similar claims. For example, ethanol from corn also eliminates carbon dioxide in the atmosphere through photosynthesis. Unlike corn, however, algae grow in water, usually in man-made ponds built on land not used for crops. Additionally, algae do not require fresh water. Instead algae can be grown in salt water, and, in some cases, even sewage water and other waste material.

The most promising aspect of algae biofuel stems from its yield. When compared to other biofuel producers, algae’s fuel yield per harvested acre is over 500 times greater than that of corn.

The following chart compares commonly used biofuel crops on several important factors.

Table 1: Comparison of biofuel crops (based on information found at: algaefuel.org and clgas2org.wpengine.netdna-cdn.com)

Product	Oil Yield Gallons/Acre	Harmful Gas Emissions	Use of Water to Grow Crop	Fertilizer Needed to Grow Crop	Energy Used to Extract Fuel from Crop
Ethanol from Corn	18	high	high	high	high
Biodiesel from Soybeans	48	high	high	low-medium	medium-low
Biodiesel from Canola	127	medium	high	medium	medium-low
Biodiesel from Algae	10,000	negative	medium	low	high

QUESTIONS: SCIENTIFIC AND QUANTITATIVE REASONING

1. Which of the following negatively affects algae biofuel's ability to be a "carbon-negative" energy source?
- A. It takes 3000 liters of water to create one liter of biofuel from algae, which is highly inefficient and wasteful of resources.
 - B. The process of extracting biofuel from algae requires more energy than is generated by burning the biofuel itself.
 - C. The construction of facilities needed to extract algae biofuel would initially require the use of fossil fuels for energy.
 - D. Algae biofuel is about 25 years away from being commercially viable, by which point there will be more efficient alternative energy sources.
2. The graph shows that food and oil prices increase and decrease together. Which of the following is the most plausible explanation for this phenomenon?
- A. As the price of food increases due to supply and demand, the cost of oil also rises because less land is available for planting corn.
 - B. Food and oil suppliers dictate the prices of their goods. Therefore, the prices of food and oil rise as consumers can afford to pay more for commodities.
 - C. The prices of oil and food are simultaneously affected by global conditions, such as natural disasters, weather, famine, and political unrest.
 - D. Farmers plant more corn for ethanol when the price of oil increases. The price of food then rises because less food-yielding crops are being produced.
3. What additional information could be added to the table for evaluating the efficiency and viability of algae biofuel compared to other sources of biofuel?
- A. The average amount of money farmers earn per acre for each biofuel source.
 - B. The costs associated with the extraction of energy from each biofuel source.
 - C. The taxes collected by the government on the sale of each biofuel crop.
 - D. The level of financial support each type of biofuel has received from investors.
4. Which of the following could plausibly occur if algae become a highly efficient and cost-effective source of biofuel?
- A. The price of food would fall because more farmland could be used to produce food rather than corn harvested for ethanol.
 - B. The supply of fresh water would be reduced because of the demands of harvesting algae for biofuel.
 - C. The cost of fuel would rise as the world's markets become flooded with alternative sources of energy.
 - D. The amount of carbon in the air would increase because more fuel will be burned due to lower costs.

Answer Key:
1) B 2) D 3) B 4) A

DOCUMENT 1: CRITICAL READING & EVALUATION

Dear Nord County School Board,

We urge you to consider a ban on serving coffee in the Nord High School cafeteria. This is important for protecting and promoting good health practices in our teenagers. Caffeine is a harmful drug for growing brains and bodies. Many adults struggle to break their own addiction to coffee so allowing the teenagers at Nord High School to begin drinking coffee on a regular basis is a dangerous idea. Teenagers have less self-control and common sense about their own health than their adult counterparts.

There may be parents and researchers who claim that a daily cup or two of coffee for a teenager is not dangerous, but this is a misconception that is easily erased by simply looking at the facts. Teenagers need more sleep than most adults because their minds and bodies are still developing. Caffeine consumption disrupts their sleep cycles and leads to sleepiness during the school day. One study found that teenagers who fell asleep during class consumed 76% more caffeine than those who did not sleep during the school day. Additionally, caffeine consumption can lead to mood swings, impulsiveness, and loss of control. These are issues that many parents deal with. Serving coffee in the Nord High School cafeteria only worsens these problems and threatens the healthy functioning of our high school students.

Ban coffee from Nord High School and help Nord teenagers lead healthier lives.

Sincerely,

Garret Ricci

Garret Ricci
Parent of Nord High School students

PETITION TO KEEP COFFEE IN OUR SCHOOL CAFETERIA

To all Nord High School students:

Due to complaints from some parents, the Nord School Board is now considering a ban on coffee in our high school cafeteria. This would be an injustice to our school community! We have a right to make our own choices about our bodies and our consumption habits. Coffee is a healthy drink in moderation and is an important part of the school day for students who lead busy lives, balancing homework, friends, work, and extra-curricular activities. Just one cup of coffee during the day can help busy students stay alert and focused.

It's time that the Nord School Board treats high school students like the young adults that we are. They must give us the responsibility of making smart choices, and we will rise to the occasion. We must demand respect for our choices and our needs.

Oppose the ban on coffee in the Nord High School cafeteria by signing the petition below. Protect our rights!

Sincerely yours,

Lisa Browning

Nord High School Senior Class President

QUESTIONS: CRITICAL READING & EVALUATION

1. Which of the following statements, if true, would most seriously weaken Garret Ricci's claim?
 - A. Teenagers who are prone to mood swings and impulsiveness consume caffeine at the same rate as their peers.
 - B. Adults who consume a small amount of caffeine daily are able to multitask more efficiently.
 - C. Adults who consume caffeine regularly were not necessarily coffee drinkers as teenagers.
 - D. Eighty percent of caffeine consumed by teenagers is consumed in the form of soda and other caffeinated non-coffee beverages.

2. Which of the following is a significant flaw in the Garret Ricci's argument?
 - A. The author assumes that teenagers have less self-control than adults, without any evidence.
 - B. The author claims that sleeping during class is caused by caffeine consumption, while it may be that caffeine consumption is a result of sleepiness.
 - C. The author associates sleep and mood with health, without explaining the connection.
 - D. The author uses anecdotal evidence from parents and teenagers, rather than a substantial body of research.

3. On which point do Garret Ricci and Lisa Browning most clearly disagree?
 - A. the ability of teenagers to make reasonable judgments about their own health
 - B. the usefulness of coffee as a replacement for sleep
 - C. the effects of coffee on the human brain and body
 - D. the prevalence of coffee in a variety of cultural and commercial settings

4. It can be inferred that Lisa Browning would **most likely** agree with which of the following statements?
 - A. The School Board should not be allowed to make decisions about anything that affects the daily life of students.
 - B. The job of a class president is to protect the rights of students and represent their voices.
 - C. Parents who complain about coffee in the cafeteria have a negative view of teenagers.
 - D. Every high school student should enjoy the physical and mental benefits of coffee by drinking it daily.

5. Which of the following statements could be used as a counterargument to Garret Ricci's claim?
 - A. Coffee needs to be available in high school cafeterias for the teachers and staff members who rely on it.
 - B. Because of its bitter taste, most teenagers are unlikely to consume coffee, whether or not it is served in their high school cafeterias.
 - C. Teenagers will be exposed to coffee elsewhere, so it is important that they learn to consume it in school, with self-control and moderation.
 - D. It is the parents' job, not the school's, to determine whether their teenagers should consume caffeine.

DOCUMENT : CRITIQUE-AN-ARGUMENT

ARGUMENT

Law-enforcement agencies depend heavily on eyewitnesses to identify suspected criminals. Indeed, it is estimated that 77,000 people nationwide are put on trial each year because of eyewitness identification. Traditionally, eyewitnesses are asked to identify suspects in a police "lineup" where suspected criminals are presented along with known innocents, called fillers, in a simultaneous (all at once) lineup. However, nowadays the lineups typically involve photos, not actual people. New research conducted in a well-controlled laboratory setting suggests that presenting photographs in a sequential (one at a time) lineup significantly reduces eyewitnesses' identification of fillers from 18% in simultaneous lineups to 12% in sequential ones. It is clear that the sequential lineup is far superior to the simultaneous one, and it is imperative that law-enforcement agencies change the way in which eyewitnesses identify criminal suspects. This will greatly reduce the number of innocent people put on trial.

QUESTIONS: CRITIQUE-AN-ARGUMENT

1. Which of the following is the strongest argument against the speaker's position that law-enforcement agencies need to change eyewitness identifications from simultaneous to sequential lineups?

- A. Simultaneous lineups have traditionally been used and have always worked well, so it does not make sense to change things.
- B. Eyewitnesses using a sequential lineup may not be better at ruling out fillers because the rate of misidentification between the two groups is not that large.
- C. It is easier for eyewitnesses to rule out fillers in a simultaneous lineup because they are seeing everyone at the same time.
- D. People should have faith in the legal system because there are many steps in the judicial process to prevent an innocent person from going on trial.

2. The speaker states that the study was conducted in a well-controlled laboratory setting. The speaker probably intended this statement to

- A. establish that a laboratory study is better than a study that was conducted in the field because it is free of competing explanations for the difference between the two lineups.
- B. illustrate that a laboratory setting is one in which a placebo must be in place in order for researchers to draw an accurate conclusion about the two lineups.
- C. demonstrate that both real-world and scientific experiments can be conducted in laboratory settings because laboratory settings are neutral environments.
- D. reveal that the results of the study are not accurate because studies conducted in a laboratory setting are contrived and not a reflection of what happens in the real world.

3. Which of the following research results would best strengthen the case for law-enforcement agencies using sequential lineups instead of simultaneous ones?

- A. The same percentage of suspects was found guilty by juries regardless of whether a sequential or simultaneous lineup was used.
- B. Eyewitnesses presented with a sequential lineup made fewer overall selections than those presented with a simultaneous lineup.
- C. Eyewitnesses presented with a sequential lineup feel more confident about their choices than those presented with a simultaneous lineup.
- D. Fewer fillers were identified as criminals by eyewitnesses presented with a sequential lineup than those presented with a simultaneous one in real-life cases.

4. What assumption does the speaker make when stating that law-enforcement agencies can reduce the number of innocent people sent to prison if they use sequential lineups?
- A. Eyewitnesses could identify fillers as criminal suspects who then could be incorrectly put on trial and ultimately sent to prison.
 - B. If the simultaneous lineup is less accurate at identifying suspects, then more fillers are misidentified and incorrectly tried than if law-enforcement agencies only use sequential lineups.
 - C. If the sequential lineup is better at increasing the number of correctly identified suspects, then the fillers will no longer be needed, leading to fewer people being incorrectly put on trial.
 - D. People who act as fillers in multiple lineups could be incorrectly identified as suspects in one lineup but not in another.
5. Eyewitnesses from multiple cases were recruited to participate in a follow-up study where they were randomly assigned to one of two groups. Which one of the following research designs could be used to test the hypothesis that an officer's body language influences eyewitnesses' ability to correctly identify a suspect in a lineup?
- A. Have officers with knowledge of the cases present images in a sequential lineup to one group of eyewitnesses and in a simultaneous lineup to the other group.
 - B. Have officers with no knowledge of the cases present images in a sequential lineup to one group of eyewitnesses and in a simultaneous lineup to the other group.
 - C. Have officers with knowledge of the cases present images to one group of eyewitnesses and officers with no knowledge of the cases present images to the other group.
 - D. Two officers, one with and one without knowledge of the cases, present images to one group of eyewitnesses and another officer with knowledge of the cases presents images to the other group.

Answer Key:
1) B 2) A 3) D 4) B 5) C