SUBSCORE REPORTING COMPARISON BETWEEN THE LOG-LINEAR COGNITIVE DIAGNOSTIC MODEL AND THE HABERMAN MODELS

By

Peter J. Ramler

Submitted to the graduate degree program in Educational Psychology and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy

_____
John Poggio, PhD, Chairperson

_____
Jonathan Templin, PhD

_____
Meagan Patterson, PhD

_____
Suzanne Rice, PhD

_____
Lesa Hoffman, PhD

Date Defended: 28 November 2018

The dissertation committee for Peter Ramler certifies that this is the approved version of the following dissertation:

Subscore Reporting Comparison Between the Log-Linear Cognitive Diagnostic Model and The Haberman Models

_____

John Poggio, PhD, Chairperson

Date Approved: 5 December 2018

Abstract

Academic assessments are designed and used for a variety of purposes. One of the primary functions of any of these tests is to diagnose the respondent. In all forms of diagnosis, a single or series of assessments are conducted. Based on the outcome of these assessments, a decision can be made as to the condition of the person or object. For this study, the diagnosis is conducted on a person. This can be in a purely academic setting such as a K-12 school, a post-secondary situation such as a university or trade school, or certification and licensure testing such as medical or another professional field. The goal is to determine the level of a latent trait the respondent possesses through a test designed to identify and measure this trait (Rupp, Templin, & Henson, 2010). It is these subscores that can potentially add to the diagnostic ability of a specific test. If a subscore can provide additional information about the level of a latent trait, the diagnostic ability of the test could be increased. Low psychometric qualities such as having low subscore reliability or having high correlations between the subscores in question and the total score of the test are always of concern (Sinharay, Puhan, & Haberman, 2010). This study looks at the theoretical possibility to use Log-Linear Cognitive Diagnostic Models (LCDM) to examine candidate mastery levels of individual attributes (subscores) within a test while keeping reliability and validity of these subscores psychometrically suitable. This simulation study compared reliability and correlation estimates of five Classical Test Theory (CTT) based subscore methods with the LCDM method. Data sets simulated both CTT data and LCDM data. Both types of data were analyzed by the CTT methods and LCDM. One of the CTT methods produced higher reliability estimates and more accurate correlation estimates than the remaining four CTT methods. However, the LCDM produced the highest reliability estimates and most consistent correlation scores across both data sets.

Acknowledgements

I am grateful to each of my committee members for their knowledge, expertise, guidance, and support, as each has played a significant role in the completion of this work.

Dr. Poggio, I couldn't have done this without you. You believed in me from our first conversation. You have made me think when I didn't want to, work harder than I ever have, and in the end realize that this goal was attainable. I have always been a lifelong learner, as are you. What you gave me is confidence in myself to go out into the world with my knowledge and make a difference. Thank you.

Dr. Templin, I had no idea I would be relying on you as much as I did. Your knowledge on LCDM, R, Mplus, and the list goes on, is truly amazing. Your help allowed me to make connections to areas outside of my dissertation that I didn't even realize were there or possible. The combination of you and Dr. Poggio made this a wonderful experience. I wish you all the best at the University of Iowa, what a great position for you to build your knowledge base and share this with the world.

Dr. Patterson and Dr. Rice. You reminded me that there is a reason for statistics outside of statistics themselves. You allowed me to be myself in your classes, for this I am forever grateful. You also taught me not to take life too seriously and just because it is in print doesn't mean it's gospel. There is just as much junk out there as there is good stuff.

Dr. Hoffman, you saved me with your flexibility and willingness to join the committee at the last minute. Thank you and I wish you the best at your new position. They are fortunate to have you and I am proud to say you are at my alma mater.

Dr. Peyton, thank you for your advice going into my oral comprehensives. Also, thank you for the two years of being your GTA. You reminded me that teaching at any level is rewarding and fun. I had forgotten that. Thank you.

Dr. Skorupski showed me that there is life outside of KU and to enjoy it to its fullest. You also, inadvertently I think, demonstrated the multiple uses of both IRT and MCMC. All three of these lessons I take with me to the new chapter in my life.

Dr. Thomas for getting me through my master's program and introducing me to Dr. Poggio. For without this introduction none of this would have been possible.

Beverly Kelley. You went from "what?" To "where?" To "what will you do?" To "OK, let's get it finished." You supported everything in my life through this process, starting with the decision and ending with graduation. You cooked for my committee, listened to me complain about everything from a professor to R. You also celebrated everything with me from a professor to R. You learned to live your life on one floor of the house while I was on another in my office working. In the end words do not do this justice so I'm going to stop trying, please just understand these last few: Thank you and I love you.

My girls, Lindsay and Sydney, even though this was my goal, I wanted to make you proud of your dad. I hope this is a step in the right direction. Love you always.

Bray, Hunter, and Berit, my step-children. Thank you for putting up with me. You were actually there the entire time and I know I wasn't easy to be around a lot of it. Thank you for being patient and supportive. I love you all.

Last but certainly not least, my mom and dad. Throughout my entire life you have given me unquestioned support in my decisions even though I know you questioned some or most of

them, including this. I did it! Now on to the next decision, get ready. Thank you for choosing me. I love you both.

Table of Contents

List of Tables

Chapter One: Introduction

Diagnosis

Cognitive assessments are designed and used for a variety of purposes. Academic examples would be an English test to diagnose reading comprehension, a mathematics test to assess fluency with working with division, or a science test to show understanding of the scientific method. Non-academic tests such as certification and licensure tests measure competencies in medicine, investing, teaching, or nursing. There are also non-cognitive assessments that help to diagnosis non-cognitive latent traits such as depression or bi-polar disorder. One of the primary reasons for the assessment is to diagnose the respondent. Diagnosing something can take many forms, ranging from a medical examination to a mechanical inspection. In all forms of diagnosis, a single or series of assessments are conducted. Based on the outcome of these assessments, a decision can be made as to the condition of the person or object.

For this study, the diagnosis is conducted on a person. This simulation was for a certification and licensure test. The goal is to determine the level of a latent trait the respondent possesses through a test designed to identify and measure this trait. A latent trait is quality or characteristic possessed by the respondent that is not directly observable. The process involves specifying the diagnostic questions, selecting the diagnostic method, and applying and evaluating the data gathered from these methods. In the end, it is the purpose of the assessment that matters (Rupp et al., 2010).

There will be three simulated latent traits in this study. The simulation will produce reliability and correlation estimates to determine which of the six scoring methods yield

subscores giving the most additional information to the stakeholder concerning each of the three latent traits. This leads to the following research question:

> Do the subscore estimates provided by a LCDM (Henson, Templin, & Willse, 2009) provide an improvement (more reliable scores, lower correlations between scores, magnify subscore information in a beneficial manner) over Classical Test Theory (CTT) based reporting such as the raw subscore reporting, differential scoring, and Haberman methods #2, #3, and #4 (Sinharay & Haberman, 2008) used in this study?

Stakeholders

Numerous groups of stakeholders are involved in the testing process. Each group has its own set of criteria, outcomes, and standards for a given test. This potential list of stakeholders includes the agency sponsoring the test, institutions educating the test takers, the test developer, teachers, students, administrators, associations, licensing boards, universities, and the general public. At times, stakeholders want tests that give a total test score while at the same time providing diagnostic information in the form of subscores for student remediation, institutional evaluation, or program effectiveness (Feinberg & Wainer, 2014b; Haladyna, 2004; Puhan, Sinharay, Haberman, & Larkin, 2010). Stakeholders are interested in obtaining detailed diagnostic information about the levels of proficiency demonstrated by the test takers in the specific traits being examined (Rupp et al., 2010).

Subscores

The objective of an assessment is to provide evidence about the examinee's skill and knowledge in a particular area (Sinharay et al., 2010). Within any given test there may be subtests specific to lesser concepts of the overall test (Wainer et al., 2001). Besides the composite score, there are subscores. These are scores within a test that measure the lesser

constructs being assessed. A construct is a descriptive variable that is not directly detected. These subscores are used for complementary purposes such as correction, appraising, and the improvement of instruction (Sinharay & Haberman, 2008). At times the subscores have a low correlation, or orthogonality, to each other and to the total score. These types of subscores provide useful additional information. At other times, due to high correlation, or lack of orthogonality, subscores provide little additional information above the total score (Monaghan, 2006; Puhan et al., 2010; Wainer et al., 2001).

It is these subscores that can potentially add to the diagnostic ability of a specific test. If a subscore is able to provide additional information about the level of a latent trait, the diagnostic ability of the test could be increased. It could show in finer grain detail the level of which the respondent demonstrates the latent trait in question.

Subscores should be considered part of the overall test given to a candidate. The subtests consist of fewer items. Added together they form the test itself and are preferably integrated into the original test design. For example, if the overarching unidimensional test design contains 45 items, the subtests may contain totals of 21, 15, and 9 items. Psychometric problems arise when looking at subscores, which represent the performance of an individual on what is principally a shorter version of the test.

Standard 1.14 of the Standards for Educational and Psychological Testing states, "When interpretation of subscores, score differences, or profiles is suggested, the rationale and relevant evidence in support of such interpretation should be provided. Where composite scores are developed, the basis and rationale for arriving at the composites should be given." (AERA/APA/NCME, 2014)(p. 27). The value of both the full test (total score) and the subtests

(subscores) must be judged in the same manner and held to the same standard (Feinberg & Wainer, 2014b).

Assessments

In 2010 de la Torre quotes Wiggins (1998) and Stiggins (2002) with succinct reasoning for giving assessments. Wiggins purports assessment is "to educate and improve student performance and not merely to audit it." Stiggins is similar in proposing that assessment be used not only to ascertain the status of a student's learning but also to further the learning process (de la Torre, 2010). With the growing emphasis on accountability, the available resources for assessment are increasingly being spent on assessments primarily designed to audit the learner and not be a catalyst for instruction and learning (de la Torre, 2010).

The type of test and the specific stakeholder associated with the assessment dictate the interpretation of the respondents' test scores. Teachers, administrators, and the public are interested in growth and remediation possibilities. Assessments for this group of respondents should follow Wiggins (1998) and Stiggins (2002) thought that assessments should further learning, not just be an audit of what has been learned.

In this section, only educational testing has been mentioned. Other types of tests, such as certification and licensure tests, are designed for and used by a separate set of respondents and stakeholders. There is overlap in score interpretation, but the main difference is in the test taker. Licensure and certification testing typically have a cut score that the candidate must reach in order to receive the credential. This group cares about three things following a test: 1) Did I pass?, 2) If I failed, how badly did I do?, and 3) What do I need to study most in order to pass the next time? (Luecht, 2003).

If the candidate achieves a score that provides certification or licensure, this candidate is finished with the test. The candidates who fail to reach this cut point score are now interested in the various aspects of the test itself. Those who fail would like every possible advantage to pass the next test; this can include knowing the subscores of their previous test, for knowing how they scored on the internal aspects of a test may help prepare them for the next attempt (Sinharay & Haberman, 2008).

An example is a real estate licensure test. The test itself may be unidimensional in design, assessing a single construct. A specific topic may be real estate sales competency. Within that test, there could be subtests such as ethics, residential regulations, and financing options. If two respondents fail to reach the cut score with the same total score, it is possible that they could have completely different mastery levels of the tested material. Knowing their subscores could be beneficial to their study patterns leading up to a subsequent test attempt.

Summary

In many cases, the reported subscores are not psychometrically reliable enough to give stakeholders any additional information not already provided by the total score. Evaluating subscore data received from assessments that were fashioned to scale respondents on a unidimensional continuum almost always provides poor results from a diagnostic perspective (Rupp et al., 2010). This study looks at the possibility of using LCDM to estimate the probability of candidate mastery of individual attributes or subtests while keeping the reliability and validity of these subscores psychometrically fit.

When diagnosing respondents, it is important to clearly understand the level at which the attribute can reliably be measured. Cognitively diagnostic assessments look to provide a more pointed or finer grain of assessment of the latent variable. This in turn aids stakeholders of all levels in the understanding of respondent performance (Rupp et al., 2010).

The data to be used will be simulated from both a Rasch model and the LCDM. Tests A_1, A_2, A_3, and A_4 will use the Rasch Model to produce unidimensional and multidimensional data. The total test for all will be 30 items with subtests of 6, 10, and 14 items. All will have sample sizes of 300 and 1200 participants. Test B_1, B_2, B_3, and B_4 will use the LCDM to produce unidimensional and multidimensional data. All tests will be 30 items with subtests of 6, 10, and 14 items. There will be 100 replications for each test analysis.

Chapter 2, the Literature Review, will explore subscores. This will include what subscores are and some of the various methods currently used to calculate subscores. Particular attention will be given to the methods of Haberman (Sinharay & Haberman, 2008) both in the literature review and the research methods chapters. Raw subscore generation, differential scoring, and Haberman's methods two, three, and four will be used in the study and compared to the LCDM. The LCDM will also be discussed in detail in both the Literature Review and in Chapter 3, Research Design.

The Research Design section, Chapter 3, will describe the data being used and how it was compiled. Differential scoring and the Haberman methods will be examined in further detail as will the LCDM method. This will include the calculation of reliability and correlation between the individual subscores. This section will identify the various statistical programs used and the output received. Chapter 4 shows the results of all 16 data sets being analyzed by all six of the subscore methods. Chapter 5 discusses the results of Chapter 4 and brings back in the literature review along with the overall goal of the dissertation: which method produces the better subscore result and why?

Chapter Two: Literature Review

Subscores

This paper looks at the possible improvement of subscore reporting by using the LCDM method or reports compared to more traditional CTT methods such as the Haberman methods. It is looking for high reliability, above 0.85 (Ling, 2009), with low to moderate subscore correlations (Sinharay et al., 2010).

Subscores are a part of many academic and nonacademic assessments. The need for psychometrically sound subscore reporting is of growing value for diagnosis of the respondent. The Literature Review will look at current subscore reporting methods as well as using LCDM for an additional reporting option.

Within any given test, there may be subtests specific to the lesser constructs of the overall test purpose (Wainer et al., 2001). There may be one overriding purpose, for example, an achievement test in language arts or a veterinary assistant certification test, but there may also be additional reasons for looking at test scores. The goal of an assessment is to provide information about the examinee's skill and knowledge in a particular area (Sinharay et al., 2010).

Subscore validity.

As with all test construction, validity is the number one concern. Do the test scores show that the test measured what it was designed to measure? This is true for subscores as well. The use of subscores by a stakeholder group mean very little if score validity comes into question. Haladyna and Kramer (2004) put forth a list of subscore validity evidence interpretations. This list of ten items (See Appendix 1) is a sound reference at all stages of test development when subscore reporting and use is anticipated (Haladyna, 2004).

Problematic validity may also be in the forms of having low subscore reliability from the start or high correlation between the subscores in question and the total score of the test are potential negatives of reporting subscores (Sinharay et al., 2010). The dimensionality of a test often influences the usefulness and interpretation of the subscores. The internal structure of test dimensionality (unidimensional, one construct, versus multidimensional, two or more constructs) can provide additional evidence when looking at test reliability issues and deciding whether or not to report subscores (Ling, 2009). If the test is designed to be unidimensional, then providing useful subscores will be difficult (or arguably, uninformative at best or misguided at worst). If it is a multidimensional design, then subscore information may be more informative than the total score itself (Monaghan, 2006). There are four data sets created with a covariance matrix of 1 causing these to be unidimensional. The remaining data sets will have various covariance matrices to be intentionally multidimensional.

If using raw subscores, the calculation of subscores by definition is lower than the total score of the test. This results in lower reliability estimates for the subtests compared to the total test reliability estimate. This lower reliability is a primary drawback to reporting subscores. In order to raise the reliability to an acceptable level, the subscore must either be part of a multidimensional test, have a sufficient number of items, or post-hoc calculations must be done (Feinberg & Wainer, 2014b).

Subscore calculation methods.

A number of methods calculate initial subscores. Post-hoc augmentation methods are also available to recalculate these original subscores with the goal of producing more psychometrically sound results. Using Classical Test Theory (CTT), true subscores can be estimated in various ways: by using the observed subscore, using the observed total score, or by

a combination of the two. It is also possible to estimate the residual true subscore by regressing the true subscore on the true total score (S. J. Haberman, 2005).

The value of subscores to the various stakeholders is in the amount of additional information the subscore provide above and beyond that of a total test score. This is directly related to the extent the subscores are orthogonal to each other and, or the total test score. The smaller the correlation between the subscore and the remainder of the test, the greater the likelihood that the subscore is providing additional value (Feinberg & Wainer, 2014b; Monaghan, 2006). Reliability also plays a role in subscore value. It is recommended to not report a subscore with a reliability of less than 0.85 (Ling, 2009).

DETECT.

Ackerman and Shu (2009) use programs such as DETECT to determine subscore usefulness. DETECT uses an algorithm that searches through all possible item clusters to find the one that maximizes the DETECT statistic. The DETECT statistic measures the dimensionality of the given test. A DETECT score of < .1 shows unidimensionality, .1 to .5 shows weak dimensionality, .51 to 1.0 is moderate dimensionality, and > 1.01 relates to strong dimensionality (Sinharay, Puhan, & Haberman, 2011). The larger the DETECT score, the stronger the dimensionality, the better possibility of producing useful subscores.

Evidence-centered design.

Early communication between the test designer and the testing client helps to improve subscore reliability. The Evidence-Centered Design (ECD) approach is a good example of how communication early and often leads to increased validity (Feinberg & Wainer, 2014a; Huff, 2010). ECD is an example of a method of assessment development that is designed to provide worthwhile subscores regardless of test dimensionality. The assessment's design is the most essential factor in determining subscore value (Monaghan, 2006).

Item Discrimination

Item discrimination also plays a role in developing high reliability for subscores. Because the number of items in a subtest is lower than that of the total test, item discrimination becomes even more important. A test developer must not only keep in mind the length of the test but also the psychometric quality of questions being put forth. Both have influence over the final test product (Ebel, 1967).

Spearman-Brown Prophecy

It is also possible to use the Spearman-Brown Prophecy formula to ascertain the number of items needed to achieve the desired subscore reliability (Feinberg & Jurich, 2017),

$$k = \rho_{cc'}(1 - \rho_{ii'})/ \rho_{ii'}(1 - \rho_{cc'})$$

with k is the factor of increase (or decrease) needed for a test with reliability = $\rho_{ii'}$ to have a reliability = $\rho_{cc'}$ (Raykov & Marcoulides, 2011).

This is an a priori method of increasing the number of test items for each subtest. An issue this method underscores is that the number of items the formula suggests for measurement effectiveness is often unreasonable in a testing situation.

Post-Hoc Adjustments

When tests are not designed to specifically report subscores, post-hoc adjustments are necessary. Many reported subscores are often the result of retrofitting a unidimensional test to multidimensional subscores (Sinharay, Puhan, et al., 2011). The practice of post-hoc augmentation of subscores post-hoc has become popular and at times has been shown to add value and increase reliability (Sinharay et al., 2010). In 2001, Wainer et al. recommended that if the reliability of the subscores becomes too similar to the reliability of the total score, then the test is unidimensional, causing subscores to have little if any added value (Wainer et al., 2001).

A method to augment subscores is to regress each of the remaining subscores. Weights are given to each of the scores, so a candidate's subscore is the function of the results of the other areas of the test as well (Skorupski & Carvajal, 2010; Wainer et al., 2001). This falls under the general idea of using collateral information to boost subscore precision and reliability.

Empirical bayes estimation.

Another technique is referred to as the empirical Bayes estimation. The underlying driver of this approach is the assumption that by regressing an estimate toward an aggregate value, such as the mean of the collective body, the precision will be improved. The idea behind these approaches is to use information from the entire test to assist in estimating the subscale performance (Skorupski & Carvajal, 2010; Wainer et al., 2001).

Borrowing information.

In 2011, Sinharay, Haberman, and Wainer start their paper with four variations of how to increase the precision/meaningfulness of subscores by borrowing additional information from related total scores or subscores.

1. Using augmented subscores, which are a function of an examinee's subscore and that of the remaining subscores. (Wainer et al., 2001) (Haberman's 2nd method)
2. Calculating the Objective Performance Index (OPI), Yen (1987): a weighted average of the observed subscore and an estimate of the observed subscore obtained using a unidimensional Item Response Theory (IRT) model for the entire test.
3. Deriving a weighted average of a subscore and the total score. This was found by Sinharay (2010) to be similar to the augmented subscores generated in example #1. (Haberman's 4th method)
4. Using the estimated abilities or their resulting transformations from a Multivariate IRT (MIRT) model.

These methods demonstrate that the adjusted subscore reliability increased. The drawback to these methods was that the validity of the scores they were borrowing information from decreased. Also, the subscores in question became more correlated to the scores in which they were borrowing information (Sinharay, Haberman, & Wainer, 2011).

Proportional reduction in mean square error.

The proportional reduction in mean square error (PRMSE) is a common method of determining subscore value. Several methods utilize the computation of the corresponding mean-square error (MSE). The PRMSE is then calculated to determine if a subscore has contributed additional value.

The mean-square error demonstrates how close a regression line is to a set of data points. The error distance a point is from the regression line is squared to remove any negative values, and this also gives more weight to the larger differences. Finding the average of these squared differences gives the MSE. The smaller the MSE, and the smaller the distance between the points and the line, the better the regression line fits the set of points (Glen, 2018).

The lower the MSE, the better the subscore estimate. An increase in reliability reduces the MSE. As will be seen in the Haberman Methods section and in Chapter 3 the potential increase in reliability for each method is the driving force in subscore interpretive value.

Value added ratio and predicted value added ratio.

The formula for the Value Added Ratio (VAR) is $r_1/r_3^2r_4$. The reliability of the subscore is r_1. The disattenuated correlation of the subscore and the total score is r_3. The reliability of the total score is r_4 (Sinharay, Haberman, & Boughton, 2015). The subscore should only be used if VAR > 1.

In 2014, Feinberg and Wainer added their own version of how to find the Predicted Value Added Ratio (PVAR): $(PRMSE_s/PRMSE_x) = PVAR \approx 1.15 + .51*r_1 - .67*r_2$. In this adaptation,

r_1 is the subscore reliability and r_2 is the disattenuated correlation between the subscore and the remainder score. PVAR and VAR do not often lead to the same decision. PVAR is based primarily on simulation. Given the similarity in computation with the original VAR statistic, the use of PVAR is not clear (Feinberg & Wainer, 2014a; Sinharay et al., 2015).

Approximation of the true residual.

A_T: true subscore, E(A): subscore mean, B_T: true total score, E(B): total score mean

$$D_T = [A_T - E(A)] - \zeta[B_T - E(B)]$$

$$\zeta = Cov(A_T, B_T)/ \sigma^2(B_T)$$

The dependent variable D_T provides a measure of the information provided by the true subscore that is not provided by the true total score. If D_T is positive, the expected performance on the subscore is better than that of the total score. If D_T is negative, the expected performance of the subscore is weaker than that of the total score (S. J. Haberman, 2005).

Agreement and Correlation Methods

Babenko and Rogers (2014) cover three methods to calculate subscores. The first is PRMSE. This was outlined previously and will be covered in more detail within the Haberman Methods section.

The second is the agreement method. This method starts with subtests j and k expressed in the same metric, for example z-scores ($\mu=0$; $\sigma=1$). The difference between the two standard scores is calculated by $d_i = z_{ij} - z_{ik}$. In this model, z_{ij} is the standard score of individual i on subscore j and z_{ik} is the standard score of individual i on subscore k. If these are close in value, d_i is small, and the obtained subscore differences are no greater than chance.

In the third method, the correlation corrected for attenuation is demonstrated by ${}_c\rho_{jk}=\rho_{jk}/(\sqrt{\alpha_{jj}\alpha_{kk}})$. Here ρ_{jk} is the uncorrected correlation between the scores on subtests j and k, α_{jj} and α_{kk} are the internal consistency estimates for j and k. If ${}_c\rho_{jk}$ is less than .9, then the student's

performance on the two tests differ and reporting the subtest scores is appropriate (Babenko & Rogers, 2014).

Bi-Factor IRT

With the advent of computer-based testing, the evolution of mixed format testing was inevitable. Multiple choice, constructed response, and essay formats are a few of the more popular forms. With this shift have come two very important questions: 1. Do mixed format tests measure the same construct? 2. Is a unidimensional IRT scoring format appropriate for a multidimensional test? The answers may suggest a bi-factor approach. Bi-factor models assume a general factor, which influences all test items, and a number of specific factors, which influence mutually exclusive groups of items. There are three reasons given for using a bi-factor model in these situations:

1) It allows an examination of the distortions that may occur when unidimensional IRT models are fit to multidimensional data.

2) It enables researchers to empirically examine the utility of subscales.

3) It provides an alternative to non-hierarchical multidimensional representations of individual differences (Wang, Drasgow, & Liu, 2016).

The bi-factor model of estimation looks to be a promising method to accomplish just this. The subscores calculated show the uniqueness of the group factors and demonstrate the orthogonal nature of the subscores. Using MULTILOG (Thissen, 1991), the model simultaneously estimates three parameters for the multiple choice items and five parameters for the constructed response items. A parameter is a numerical, measurable characteristic of a population. A few examples of population parameters are the mean (μ), the variance (σ^2), and the standard deviation (σ).

Haberman Methods

Three of the subscore methods of comparison in this study will be referred to as the Haberman subscore methods. The practice of borrowing from other aspects of the test such as other subscores and the total test is not unique to Haberman. The various Haberman methods are prominent in subscore literature and take into account some of the methods referenced in previous paragraphs. Sinharay and Haberman have published numerous articles related to these methods and their effectiveness in improving the value of subscore reporting (Dai, Svetina, & Wang, 2017; S. Haberman, Sinharay, & Puhan, 2009; S. J. Haberman, 2008; Sha & McCoy, 2014; Sinharay & Haberman, 2008; Sinharay, Puhan, et al., 2011).

Estimates of the true subscore: S_T. All four methods utilize the computation of the corresponding mean-square error (MSE). Methods 2-4 use the proportional reduction in mean squared error (PRMSE) to determine if a subscore has contributed additional value (Glen, 2018). The lower the mean squared error the better the subscore estimate. An increase in reliability reduces the mean squared error.

Haberman symbol definitions.

E(S): Mean of the observed subscores

S: Observed subscore

S_t: Estimate of the true subscore

S_s: Estimate of the true subscore using method 2 $E(S) + \alpha[S - E(S)]$

α: Reliability of the subscores,

S_x: Estimate of the true subscore using method 3 $c[X - E(X)]$

E(X): Mean of the observed total scores

X: Observed total score

c: Constant used in method 3, this combines both subscores and total scores

S_{sx}: Estimate of the true subscore using method 4

β: Linear regression coefficient for the subscores

γ: Linear regression coefficient for the total scores

Cronbach's Alpha is used for all estimated reliability calculations (Sinharay & Haberman, 2008).

Method 1.

- $S_t = E(S)$ This is the simplest of the four models.
- The subscore estimate S_t is simply the mean of the observed subscores.
- Mean Squared Error, MSE = $E[S_t - E(S)]^2 = \sigma^2(S_t)$
- There is no PRMSE
- In the Haberman package in R, there is no output given for this method.
- This method is not used in this study.

Method 2.

- $S_s = E(S) + \alpha[S - E(S)]$ This model uses the subscores as the predictor of the true subscore estimate.
- Mean Squared Error, MSE = $E[S_t - S_s]^2 = \sigma^2(S_t)[1-\alpha(S_t,S)]$
- The PRMSE is reduced to the reliability of the observed subscores to the true subscores, $\alpha(S_t,S)$. This is the initial PRMSE value, and a reduction from this value is considered an improvement.

Method 3.

- $S_x = E(S) + c[X - E(X)]$ Model 3 is similar to Model 2. The value for c is defined further in Chapter 3.
- The observed total score(s) are used as the predictor of the true subscore estimate.
- $c = \rho(S_t,X)[\sigma(S_t)/\sigma(X)]$

- Mean Squared Error, MSE = $E[S_t - S_x]^2 = \sigma^2(S_t)[1-\alpha(S_t,X)]$
- The PRMSE is reduced to the reliability of the true subscores to the observed total scores, $\alpha(S_t,X)$.

Method 4.

- $S_{sx} = E(S) + \beta[S - E(S)] + \gamma[X - E(X)]$ Model 4 is the most complex of the four methods as it uses both subscores and total scores to predict the true subscore.
- $\tau = [\rho(X_t,X)\rho(S_t,X_t) - \rho(S,X)\rho(S_t,S)] / [1 - \alpha(S,X)]$
- $\gamma = [\sigma(S)/\sigma(X)]*[\rho(S_t,S)\tau]$
- $\beta = \rho(S_t,S)[\rho(S_t,S) - \rho(S,X)\tau]$
- Mean Squared Error, MSE = $E[S_t - S_{sx}]^2 = \sigma^2(S_t)[1 - \alpha(S_t,S) - \tau^2[1 - \alpha(S,X)]]$
- The reliability calculation for this model used in the PRMSE calculation is as follows: $1- [1 - \alpha(S, S_t)][1 - \alpha(X, S_t*S)]$.

  (Dai et al., 2017; S. Haberman et al., 2009; S. J. Haberman, 2008; Sha & McCoy, 2014; Sinharay & Haberman, 2008; Sinharay, Puhan, et al., 2011)

Sinharay and Haberman (2008) referenced three parts to this process. The subscore value is referred to by $PRMSE_s$. The total score value is represented by $PRMSE_x$. The observed scores for the subscore and the total score are demonstrated by $PRMSE_{sx}$ (S. J. Haberman & Sinharay, 2010; Sinharay & Haberman, 2008). Sinharay and Haberman (2008) stated that if $PRMSE_s$ was less than $PRMSE_x$, then the subscores do not add value and should not be reported. Also, $PRMSE_{sx}$ will always be at least as large as $PRMSE_s$ and $PRMSE_x$ and requires a substantial amount of computation. The only time $PRMSE_{sx}$ should be used is if it will provide a significantly larger value than the other two (Sinharay & Haberman, 2008). Other models for the PRMSE values include $PRMSE_s = \rho^2 (S_t,S)$ and $PRMSE_x = \rho^2 (S_t,X)$ (Feinberg & Jurich, 2017).

A Priori Differential Scoring

A direct definition of differential scoring is that it is an a priori and empirical weighting of the test items and the corresponding options. It is most effective when there are few variables in the composite and when these variables are not highly correlated. The differential scoring method used in this study will be found in Chapter 3, Methods (Stanley & Wang, 1968).

Differential scoring has been mentioned as one of the subscore methods being used in this study. It is not technically a subscore method, but instead it is an a priori decision of how to score the items, which in turn effects the results of the subscore methods. The subscore method used on the differential scoring is the raw scoring method. It is the simplest of the methods and gives a direct comparison between doing nothing to the test and subtest scores and making an a priori decision on how to score each section of the test.

Subscore Conclusion

Subscores are more likely to add value to stakeholders' reports when the reliabilities are high and the correlations to other subscores and the total score are low (Babenko & Rogers, 2014; Sinharay et al., 2010). An additional area to consider is the reliability combined with the validity of the subscore. High reliability does not always guarantee validity, but moderate reliability scores paired with higher score validity may be worth consideration (Ling, 2009).

The use of subscores can be enhanced if the subtests are specifically developed to measure at the multidimensional level for the given content or domain. These must be designed to possess weak to moderate correlation along with highly moderate to high reliability (Babenko & Rogers, 2014). Tests in general are most often designed to cover a wide spectrum of content; it is unlikely that a subscore consisting of a few items is able to precisely measure any single unique ability (S. Haberman et al., 2009). Simply increasing the number of items does not automatically increase the usefulness of the subscore (Puhan et al., 2010).

The Educational and Psychological Testing Standards states that when interpretation is based on a small subset of items the rationale and relevant evidence should be provided in support of these interpretations (AERA/APA/NCME, 2014; Haladyna, 2004). Subscores need validity just as much as total scores do. No matter what the content or how many questions are involved, subscore reliability should always be combined with construct validity evidence (Ling, 2009; Sinharay & Haberman, 2015; Skorupski & Carvajal, 2010).

Diagnostic Classification Models (DCMs)

The use of diagnostic classification models (DCMs) should be driven by the need for classifications concerning specific attributes of the respondents that relate to learning or behavior. One of the main objectives of diagnostic assessment is to provide the various stakeholders with purposeful and meaningful information (Rupp et al., 2010).

DCMs have been referred to by various names: cognitive psychometric models, cognitive diagnostic models, latent response models, restricted latent class models, multiple classification latent class models, structured located latent class models, and structured item response theory models (Rupp et al., 2010). Because of this, all variances referenced will be referred to as DCM to avoid confusion.

DCMs are a statistical tool in cognitive diagnosis that can take on an important role, as they can be employed in a variety of disciplines, including clinical psychology, educational assessment (Liu, 2012), and certification and licensure testing. The final criterion by which these will be judged is their usefulness to the stakeholders involved (Templin & Hoffman, 2013).

Formal definition of DCM.

A thorough definition of what DCMs are comes from Rupp and Templin (2008):

> Diagnostic classification models (DCM) are probabilistic, confirmatory multidimensional latent-variable models with a simple or complex loading structure. They are suitable for modelling observable categorical response variables and contain unobservable (i.e., latent) categorical predictor variables. The predictor variables are combined in compensatory and noncompensatory ways to generate latent classes. DCM enable multiple criterion-referenced interpretations and associated feedback for diagnostic purposes, which is typically provided at a relatively fine-grain size. This feedback can be, but does not have to be, based on a theory of response processing grounded in applied cognitive psychology. Some DCM are further able to handle complex sampling designs for items and respondents, as well as heterogeneity due to strategy use. (Rupp & Templin, 2008)(Rupp & Templin, 2008)(Rupp & Templin, 2008)(Rupp & Templin, 2008)(Rupp & Templin, 2008)(Rupp & Templin, 2008)(Rupp & Templin, 2008)(Rupp & Templin, 2008)(Rupp & Templin, 2008)(Rupp & Templin, 2008)(p. 228)

General attributes of DCMs.

A general way to address DCMs is that they are suitable whenever statistically driven classifications are needed for respondents. These classifications are based on the respondents answers to test items resulting in the mastery or non-mastery of multiple categorical latent traits or attributes (Madison & Bradshaw, 2015; Rupp & Templin, 2008; Templin & Bradshaw, 2013).

The attribute mastery or non-mastery status obtained represents ordered categorical states – an ordinal level of measurement provided by DCMs. They differ from many psychometric models due to the categorical rather than the continuous nature of the latent variables. It is this

categorical difference that allows for a mastery versus nonmastery distinction (Rupp et al., 2010; Templin & Bradshaw, 2013).

In contrast to traditional models such as Item Response Theory, DCMs can be used to understand a combination of skills, processes, and problem-solving strategies potentially involved in an assessment (de la Torre, 2009). As mentioned above, DCMs define the attributes an individual has or has not mastered based on test answers (Henson et al., 2009). These models are often used together with a Q-matrix, providing stakeholders with diagnostic information about the respondents (Chen, de la Torre, & Zhang, 2013). It is these various skills, processes, and attributes within a test that become the subtests and subsequently the subscores in which stakeholders are interested.

The DCMs are discrete latent variable models developed specifically for observing the presence or absence of multiple fine-grained skills for answering test questions. These models group respondents into unobserved or latent classes (de la Torre, 2009; Rupp et al., 2010; Rupp & Templin, 2008). Possibly the most significant difference in models falling within the DCM labeling is the direct estimate the models make in reference to a given respondent's probability of falling into a latent class (Rupp et al., 2010).

For additional details on the characteristics of individual DCMs, see Rupp and Templin (2008) pages 227-237 and a taxonomy of DCMs in Rupp, Templin, and Henson (2010) page 98.

Rupp et al. (2010), have a four-step framework for specifying the various diagnostic classification models:

1. The number of latent variables and their associated scale types

a. For K latent variables (attributes), there is a set of 2^K distinct attribute profiles in the universe of latent attribute classes (de la Torre & Chiu, 2016; Templin & Hoffman, 2013).
b. K = 4, there are 2^4 or 16 possible attribute classes
c. DCMs can be seen as confirmatory latent class models in that each of the 2^K possible mastery profiles can be represented by an individual latent class (Templin & Hoffman, 2013).

2. The number of observable response variables and their associated scale types
3. The mapping of latent variables onto the observable variables (i.e., the specification of the Q-matrix), as well as the selection of an appropriate model structure (i.e., additive or multiplicative combination of categorical latent variables)
4. The way the latent variables are mutually associated and hierarchically related (Rupp et al., 2010).

Compensatory and noncompensatory.

When looking at the general differences in the types of DCMs, one of the largest is between compensatory and noncompensatory models (Henson et al., 2009). Compensatory latent trait models allow for the respondent to "make up" for what is lacking in one attribute by having mastered a different attribute. A high probability of mastery in attribute one can compensate for a low probability of mastery in attribute two. For a respondent to receive a high probability of answering an item representing both attributes one and two, the respondent need only show mastery of either attribute, not both (Henson et al., 2009; Rupp et al., 2010).

In contrast, in noncompensatory latent trait models, a high value in one attribute cannot compensate for a low value in a separate attribute. Mastery in one does not "make up" for nonmastery in another (Henson et al., 2009; Rupp et al., 2010). Specific model computation will

be addressed in the LCDM section, but in general, when a model uses a sum, the latent variables are united in a compensatory method, and when a model uses a product, they are combined via a noncompensatory manner (Rupp et al., 2010).

Condensation rule.

A condensation rule for a DCM prescribes how the various attributes are combined and therefore describes how the attributes produce a latent response. There are a variety of condensation rules available for use, but the two most commonly used are conjunctive and disjunctive (Rupp et al., 2010).

With a conjunctive condensation rule, all 'A' latent trait variables activated by a certain test item, and thus involved in the response process, are multiplied by each other producing a numerical result of either 0 or 1:

Result = Attribute 1 * Attribute 2 *….Attribute A.

To perform well on an item under the conjunctive model, the respondent must know all required attributes. Missing even a single attribute can dramatically lower the probability of responding correctly (Henson et al., 2009; Rupp et al., 2010).

In the disjunctive condensation rule, all 'A' latent trait variables activated by a specific test item and involved in the response process are again multiplied together, but in a more complicated method:

Result = 1 – [(1-Attribute 1) * (1-Attribute 2) * …. (1-Attribute A)].

Mastering a subset, in some cases only one of the attributes, is satisfactory enough to have a high probability of a correct response (Henson et al., 2009; Rupp et al., 2010).

Why DCM instead of IRT?

Item response theory (IRT).

Classical Test Theory (CTT) and Item Response Theory (IRT) both provide a continuous measure of a respondent's ability and provide only a rank ordering of the examinees. Diagnostic information (i.e., mastery or nonmastery) is only obtained using additional analysis procedures. DCMs may provide an explanation of why a respondent does not perform well based on the specific skills that have been shown as nonmastered (Henson et al., 2009).

A desirable property of IRT models is the standard error attribute to the latent variable measured by a test is conditional on the latent variable's point estimate. This allows for tests to be constructed in order to provide more exacting measurement for areas of the scale determined to be of most importance. With this flexibility of precision comes a variability in reliability across the same scale. This variability causes an increase in the number of possible places of error. In DCMs, the latent variables are categorized, error is consolidated, and the reliability of the estimate increases.

IRT models work with continuous variables while DCMs are categorical; it is this difference in variable type that makes a direct reliability comparison difficult (Templin & Bradshaw, 2013).

Differences in various DCMs.

DCMs differ from psychometric approaches such as IRT in that instead of providing a latent trait estimate that falls along a continuous scale, DCMs provide a classification based on mastery or nonmastery of the skill. An IRT model may describe the performance of a respondent globally. Test takers with higher proficiencies are expected to have higher probabilities of answering all questions correctly. In turn, a DCM may describe the respondent's performance as

a function of the specific attributes tested, either individually or in combination (de la Torre, 2009; Templin & Bradshaw, 2013).

DCMs do afford the ability to isolate a respondent's location on a continuous scale as IRT models do. But, their classification-based measurement, as mentioned above, allows for an increase in reliability and could allow for an increase in measured dimensions from a test. The attribute size, or *grain size* as it is referred to, is the level of clarification in which a test designer wishes to construct and test the various attributes of a DCM based test. Tests analyzed with DCMs are able to analyze more dimensions with the same number of items. DCMs allow for the multidimensional assessment of a unidimensional test (Rupp et al., 2010; Templin & Bradshaw, 2013).

DCMs for subscores.

Traditional models for measuring proficiency, whether classical test theory or item response theory in nature, look at the respondent's latent trait proficiency as a unidimensional construct (de la Torre & Chiu, 2016). DCMs, by contrast, are well suited for providing diagnostic feedback due to their practical efficiency and increased reliability. Instead of hypothesizing a single proficiency continuum, DCMs see proficiency as a set of separable but interrelated knowledge within a domain. This allows for a finer-grain assessment of the respondent's performance (de la Torre & Chiu, 2016; Madison & Bradshaw, 2015).

The sought-after diagnosis is the mastery or nonmastery of the test material at the subtest level. This allows for classification of the respondent based on the DCM output and analyzes the level of latent trait demonstrated. It is this type of classification-based diagnosis that subscore reporting via DCM may provide (Rupp et al., 2010).

The primary unit of analysis for a DCM is the individual. These models can be tailored to provide diagnostic feedback for individuals and groups based on discrete attribute profiles. This

allows for the information to be used to tailor instruction and resources directly to those respondents who demonstrate nonmastery of specific attributes (de la Torre, 2009; Rupp et al., 2010). Recall that DCMs are suitable for supporting diagnoses because they provide statistically driven classifications of respondents according to one or more diagnostic criteria (Rupp et al., 2010; Templin & Bradshaw, 2013). In this instance, subscores are the diagnoses of the probability of mastery for a specific latent trait.

The Q-matrix.

A core element of the DCM design is the a priori specifications of which test items match with which latent trait attributes. This attribute to item alignment is known as a Q-matrix. The Q-matrix is traditionally items (rows) by attributes (columns) (Henson et al., 2009; Madison & Bradshaw, 2015; Rupp et al., 2010; Tatsuoka, 1983; Templin & Hoffman, 2013).

The Q-matrix design used for these data will have a deliberate set of items: 6 items for attribute 1, 10 items for attribute 2, and 14 items for attribute 3 for a total of 30 items. The matrix will contain 0s (no) and 1s (yes), indicating which attributes are measured by which specific item(s) (Liu, 2012; Madison & Bradshaw, 2015; Rupp et al., 2010). In addition to identifying the item to attribute relationships, the Q-matrix plays an important role in constraining the number of parameters that need to be estimated (Chen et al., 2013; Liu, 2012).

As the number of nonzero entries increases, the complexity of the Q-matrix increases as well. This complexity will vary depending upon the number of test items that represent multiple attributes, causing interactions between the items. A *complex* Q-matrix exists when items represent more than one attribute. A *simple* Q-matrix exists when one item represents a single attribute. (Madison & Bradshaw, 2015).

The assignment of items to attributes tends to be a subjective process with those experts involved. It is this subjectivity that has raised validity concerns about DCMs among researchers

(Chen et al., 2013; de la Torre & Chiu, 2016). The development of the Q-matrix is one of the most difficult, complex, and important aspects of the DCM process. Content experts, researchers, teachers, and psychometricians all play a role in the development and verification of the Q-matrix (Henson et al., 2009; Madison & Bradshaw, 2015).

The Q-matrix is the essential element of any DCM. It represents the operational theory that gives rise to the assessment itself and it represents the proposed hypothesis about the attribute structure being tested (Rupp et al., 2010). The verification of the Q-matrix provides content validity to the assessment process. If this matrix is not specified correctly, the inferences made from the DCM application are not valid (Gierl, 2009; Madison & Bradshaw, 2015). The complexity and importance of the Q-matrix and what it represents to the DCM process and results is a strong reminder that any changes, no matter how slight, will change the inferences made from the respondent scores (de la Torre & Chiu, 2016; Henson et al., 2009).

Log-linear Cognitive Diagnostic Model (LCDM)

The LCDM is a flexible diagnostic model that allows the relationships between the categorical variables to be demonstrated using a latent class design. It is this flexibility that allows for generalizations to diagnostic classification models (Henson et al., 2009; Templin & Hoffman, 2013). LCDM can be viewed as representing *model families* that consist of a range of compensatory and non-compensatory DCMs that are developed out of different parameter restrictions being placed on the LCDM (Rupp & Templin, 2008; Templin & Bradshaw, 2013). LCDMs model the attribute effects and interactions at the item level (Madison & Bradshaw, 2015; Rupp & Templin, 2008).

Common constraints.

Three constraints are common to the LCDM. The first set of constraints, shown directly after this paragraph, must be defined to ensure monotonicity. Monotonicity is the property stating that the probability of a correct response after mastering additional skills will be equal to or greater than the probability of a correct response prior to mastering the new skills.

$$P(X_{ij}=1|\alpha^{w}_{i}) \geq p(X_{ij}=1|\alpha_{i}) \text{ for all w}$$

$$\text{Where } \alpha^{w}_{ik} = \alpha_{ik}, \text{ where w} \neq k, \text{ 1 otherwise}$$

A second set of constraints is the identification of the Q-matrix. Identifying the Q-matrix is similar to a confirmatory analysis. Each item associated with each attribute will identify the attribute. Without the Q-matrix, attributes could change in their definition similar to what may occur in an exploratory factor analysis.

A third constraint is based on the Q-matrix attributes being defined as 0 or 1. This classifies the reference group as those who have not mastered any of the required attributes for an individual item. It also identifies the probability of those respondents who have not mastered any of the required attributes as the logit ($-\eta$) (The intercept of λ_{i0} replaces $-\eta_j$ as the identifier for the reference group.) (Henson et al., 2009).

Common similarities.

Most cognitive diagnostic models are parameterized to define the probability of a correct response. Items are either correct $X_{ij}=1$ or incorrect $X_{ij}=0$, and the LCDM is expressed in terms of the log-odds of a correct response for each of the items (Henson et al., 2009).

LCDM item parameters are comparable to the different effect levels found in an analysis of variance (ANOVA). The attributes are dummy coded with 0 or 1 having a logistic link for dichotomous data. In a similar vein, LCDM allows for interaction terms to be tested for a

difference from 0. This allows for models to be compensatory or non-compensatory as needed (Templin & Bradshaw, 2013; Templin & Hoffman, 2013).

Even though the LCDM item response function (IRF) is similar to the multidimensional IRT model specifications, the latent traits in LCDM are binary (mastery/nonmastery) rather than continuous. These traits are referenced as attributes (α). Mastery is indicated by $\alpha = 1$ and nonmastery by $\alpha = 0$ (Madison & Bradshaw, 2015).

General diagnostic model (GDM).

Log-Linear Cognitive Diagnostic Models can be understood as an extension of a binary general diagnostic model (GDM). The LCDM assumes the following for binary skills:

$$f(\lambda^{*}_{I}, h^{*}(q_i, \alpha)) = \lambda_{i0} + \Sigma_{d=1} (\lambda_{id} q_{id} \alpha_d) + \Sigma_{d<e} (\lambda_i^{de} q_{id} q_{ie} \alpha_d \alpha_e + \ldots.)$$

The dots indicate that higher order terms of three or more skills may be included in the item function. The parameters λ_i^{de} , λ_i^{def} …related to these interactions quantify the conjunctive effects of having two or more of the required skills that are not explained by the main effects λ_{id} of required ($q_{id} = 1$) skills (von Davier, 2014).

As a special case, GDMs general definition incorporates dichotomous latent variable models, including LCDM. The probability of a correct response is therefore defined as

$$P(X_{ij}=1|\boldsymbol{\alpha}_i) = \exp(\lambda_{i0} + \boldsymbol{\lambda}^{T}_{j}\mathbf{h}(\boldsymbol{\alpha}_i,\mathbf{q}_j))/1+ \exp(\lambda_{i0} + \boldsymbol{\lambda}^{T}_{j}\mathbf{h}(\boldsymbol{\alpha}_i,\mathbf{q}_j))$$

$\boldsymbol{\lambda}_j$ = vector of weights for the *j*th item

$\mathbf{h}(\boldsymbol{\alpha}_i,\mathbf{q}_j)$ = set of linear combinations of the $\boldsymbol{\alpha}_i$ and the Q-matrix for the *j*th item, q_j. The set of all weights included in the full LCDM with K latent dichotomous attributes.

λ_{i0} = defines the probability of a correct response for the reference group, those respondents who have not mastered any of the attributes.

$$\boldsymbol{\lambda}^{T}_{j}\mathbf{h}(\boldsymbol{\alpha}_i,\mathbf{q}_j)=\sum^{K}_{u=1} \lambda_{ju}(\alpha_u q_{ju}) + \sum^{K}_{u=1}\sum_{v>u} \lambda_{juv}(\alpha_u \alpha_v q_{ju} q_{jv}) + \ldots.$$

Below is an example of the probability of an item that requires the first two attributes of the Q-matrix:

$$P(X_{ij}=1|\boldsymbol{\alpha}_i) = \exp(\lambda_{j0} + \lambda_{j1}\alpha_1 + \lambda_{j2}\alpha_2 + \lambda_{j12}\alpha_1\alpha_2)/1+ \exp(\lambda_{j0} + \lambda_{j1}\alpha_1 + \lambda_{j2}\alpha_2 + \lambda_{j12}\alpha_1\alpha_2)$$

(Henson et al., 2009).

Items measuring only one skill will only have an intercept and a main effect for that skill. Items measuring two skills will have an intercept, two main effects, and a two-way interaction. Items measuring three skills will have an intercept, three main effects, three two-way interactions, and one three way interaction, and so on… (Templin & Hoffman, 2013).

Continuing from above, the intercept for item ($\lambda_{i,0}$) represents the log-odds of a respondent in the reference group, obtaining a correct answer for that item. The main effects ($\lambda_{i,1,(1)}$ and $\lambda_{i,1,(2)}$) will increase the log-odds of a correct response given the mastery of the respective attributes. The two-way interaction between the two attributes ($\lambda_{i,2,(1,2)}$) for a respondent permits the log-odds of a correct response to change given the mastery of both attributes (Templin & Bradshaw, 2013).

For an assessment with **A** attributes, the respondent has one of 2^A unique patterns of attribute mastery. The LCDM assumes item independence which is conditional on the respondent's attribute pattern. The Q-matrix and model provide a series of constraints on the general latent class model. This results in a fixed number of classes and a fixed item parameter structure. A general DCM, LCDM in this case, does not have extreme parameter constraints, but instead allows for attribute-specific main effects on the individual items (Madison & Bradshaw, 2015; Templin & Hoffman, 2013).

The LCDM does not require attributes to be isolated for identification of all of the unique classes. But, including items that do isolate attributes may increase the model accuracy by a

substantial amount. Constructing factorial simple items may be worth the effort to attain this increase in precision (Madison & Bradshaw, 2015).

LCDM subsumes DCMs.

Any DCM is able to be fit by constraining the LCDM parameters. This feature allows LCDM to be used for various DCM comparisons (Henson et al., 2009).

The use of LCDM instead of a specific DCM is due to the flexibility of the LCDM in general. The LCDM subsumes most commonly used DCMs (i.e., Deterministic Inputs Noisy And Gate, DINA; Noisy Inputs Deterministic And Gate, NIDA; Reduced Reparameterized Unified Model, RUM; Deterministic Inputs Noisy OR Gate, DINO; Noisy Inputs Deterministic Or Gate, NIDO; and the Compensatory Reparameterized Unified Model, C RUM). A LCDM can take the form of each of the previously mentioned DCMs by using specific parameter restrictions. The LCDM will also allow for parameterizations that are not possible with other DCMs. This provides precise and flexible information about the structure of the items within each specific test (Henson et al., 2009; Templin & Bradshaw, 2013; Templin & Hoffman, 2013).

These common DCMs can be viewed as special LCDM cases in which certain a priori parameters are constrained to match the specific DCM (Madison & Bradshaw, 2015). The flexibility of the LCDM does not mean that the constraints must be done a priori. These options can be tested empirically to guide model specification that may lead to the best possible representation of relationships between items and attributes. The flexibility of model constraint parameter options also provides for model comparison opportunities (Bradshaw & Templin, 2014; Henson et al., 2009; Madison & Bradshaw, 2015).

Earlier in the literature review the difference between compensatory and noncompensatory DCMs was discussed. With the LCDM there is no need for this distinction.

$P(X_{ij}=1|\boldsymbol{\alpha}_i) = \exp(\lambda_{j0} + \lambda_{j1}\alpha_1 + \lambda_{j2}\alpha_2 + \lambda_{j12}\alpha_1\alpha_2)/1+ \exp(\lambda_{j0} + \lambda_{j1}\alpha_1 + \lambda_{j2}\alpha_2 + \lambda_{j12}\alpha_1\alpha_2)$ is the probability of an item that requires the first two attributes of the Q-matrix.

For the model to be compensatory, there are only interactions to consider, no main effects. λ_{j1} and λ_{j2} would be equal to zero. For the model to be noncompensatory, the interaction effect, λ_{j12}, offsets one of the main effects involved. The model itself, set up through the Q-matrix and model constraints, removes the need for specifying compensatory or noncompensatory models (Rupp et al., 2010).

Reliability

Reliability is one of the most important characteristics of an assessment. Score reliability is the degree to which scores in a particular sample are precise (Kline, 2016). Reliability for DCMs references the stability of examinee classifications of re-examination (Madison & Bradshaw, 2015; Templin & Bradshaw, 2013). It is the general notion of consistency of the scores across instances of the testing procedure (AERA/APA/NCME, 2014).

Measurement of reliability.

The subscores of a test are the components that make up the composite test. The reliability of the composite test is greater than the reliability of the components (Lord & Novick, 1968). Subscore calculations using Classical Test Theory methods use traditional forms of calculating reliability. Cronbach's Alpha is used by the raw method, the differential scoring method, and three Haberman methods in this study and are denoted by $\rho^2(x,y)$, where x and y are subscores or total scores (Sinharay & Haberman, 2008).

$\alpha = p/p\text{-}1 = \sum Cov(X_i,X_j) / Var(X)$

What alpha is and what alpha isn't:

1. Alpha in an index of internal consistency, the degree to which components are interrelated. The higher the covariance the higher the alpha and vice versa.
2. Alpha is considered a lower bound of the composite reliability only when error scores are uncorrelated.
3. Alpha is the mean of all possible split-half coefficients.
4. Alpha is not an index of unidimentionality. (Raykov & Marcoulides, 2011)

Composite reliability.

A facet of reliability to keep in mind when looking at the three Haberman methods is how a transformed reliability score affects the composite reliability score. He (2009) discusses at length and with various Classical Test Theory methods the possible equations that could be used to calculate a composite reliability given multiple test components or subtests. Below is an example from Feldt and Brennan (1989):

$R_{STRAT.a} = 1 - (\sum_i \sigma^2_i(1 - r_i)) / \sigma^2_c$ where:

$R_{STRAT.a}$: the reliability of the composite scores

r_i : the reliability of stratum *i*

σ^2_i : the variance of stratum *i*

σ^2_c : the variance of the composite scores (He, 2009)

This will not be looked at in this study but it is worth keeping in mind for further possible study and while looking at the results of the Haberman methods. It brings up the question, does the augmentation of subscore reliability affect the reliability of the composite score?

LCDM reliability.

Traditional forms of calculating reliability are not available for DCM attribute reliability. The latent traits are categorical and therefore the variance of the latent trait and the error variance are not independent. Direct estimates of these variances are not available, causing difficulty in

constructing traditional reliability estimates. With this being said, Lord's (1980) premise that *"true score ξ and ability Θ are the same thing expressed on different scales of measurement"* – (p. 46) is adopted where DCM reliability is concerned.

Using the concept of reliability observed in the first paragraph of this section, score reliability is the degree to which scores in a particular sample are precise (Kline, 2016). Reliability for DCMs references the stability of examinee re-examination classification (Madison & Bradshaw, 2015; Templin & Bradshaw, 2013).

Reliability is the general notion of consistency of the scores across instances of the testing procedure (AERA/APA/NCME, 2014). The goal is to capture the consistency of a respondent's categorized estimate from a DCM over hypothetically repeated observations. The calculation is enabled by simulated repeated draws from the respondent's posterior distribution. Tetrachoric correlation coefficients will be used as the measure of reliability for the LCDM (Templin & Bradshaw, 2013). Tetrachoric correlation will be used for subscore correlation, and polychoric correlation, an extension of tetrachoric correlation, will be used for reliability (Uebersax, 2015).

For a given respondent, any two hypothetical test administrations are independent and therefore the correlation between α_{ea1} and α_{ea2} is zero. But, in a sample across respondents, the correlation is non-zero. It is this non-zero correlation that represents an estimate of the DCM attribute reliability. Below are the three general steps for calculating DCM reliability:

1. For each attribute α and examinee *e*, calculate the respondent's attribute mastery ($\hat{\rho}_{ea}$) probability.
2. Create the replication contingency table size (size 2x2 for binary attributes)

3. Calculate the attribute reliability using the tetrachoric correlation of $\alpha_{.a1}$ and $\alpha_{.a2}$ (Templin & Bradshaw, 2013)

Templin & Bradshaw (2013) have shown that for all models, the DCM analogs produced higher reliability for the latent attributes measured by the test and extracted by each model**.** Templin & Bradshaw (2013) also point out that the comparison between CTT, IRT and DCM reliabilities uses separate metrics: the Pearson and the polychoric correlation coefficients. This difference in metrics causes the comparisons between the various reliabilities to not necessarily be directly equivalent. (Templin & Bradshaw, 2013). Even so, reliability is a measure of consistency (AERA/APA/NCME, 2014). For IRT and CTT assessments, it is the consistency of scores. For LCDM, reliability is measuring the consistency of attribute classification.

Validity

Subscore validity.

Validity is a unitary concept. It is the degree to which all accumulated evidence supports the intended interpretation of the test scores for their proposed use. It references the degree to which evidence and theory support the interpretations of the test scores for the proposed uses of the tests. Validating a test can be seen as a process of constructing and evaluating arguments in defense of or opposed to the intended interpretations or the test scores and their significance to their proposed use (AERA/APA/NCME, 2014).

As validity relates to subscores, Standard 1.14 and 1.15 of the Standards for Educational and Psychological Testing address this specifically: Standard 1.14: “When interpretation of subscores, score differences or profiles is suggested, the rationale and relevant evidence in support of such interpretations should be provided” and Standard 1.15: “When interpretation of performance on specific items, or small subsets of items is suggested, the rationale and relevant

evidence in support of such interpretation should be provided. When interpretation of individual item responses is likely but is not recommended by the developer, the user should be warned against making such interpretations" (AERA/APA/NCME, 2014) (p. 27). Validity evidence is necessary for all levels and sizes of tests and subtests.

Subscore validity comparisons.

The validity of the subscore reporting methods will be compared using the reliability of the subscores produced by each method. Correlations between subscores will also be examined. The idea behind reporting subscores is to provide additional, useful information to aid in furthering the goal of the stakeholder in question. This goal may be preparation, remediation, evaluation, or any combination of these or additional goals set by the individual stakeholder (Sinharay & Haberman, 2008; Sinharay et al., 2010).

Classical test theory (CTT) subscore validity issues.

The validity of traditional subscore reporting, such as the Haberman methods used in this study, come into question with subscore reliabilities under 0.85 (Ling, 2009) and an increase in subscore-to-subscore and subscore-to-total-score correlation (Skorupski & Carvajal, 2010). The correlation coefficient, between -1 and 1, measures the degree of linear association between two quantitative variables. The closer the absolute value of the coefficient is to 1, the stronger the relationship between variables (Coladarci & Cobb, 2014). The smaller the correlation between the subscore and the remainder of the test, the greater the likelihood that the subscore is providing additional value (Feinberg & Wainer, 2014b; Sinharay et al., 2010).

Moving the opposite direction, the closer the reliability of the subscore becomes to the reliability of the total score, the more unidimensional the test becomes and the less added value the subscores have over the total score (Wainer et al., 2001). High subscore correlation

(Pearson's r) with other subscores and/or the total score also reduce the value of subscore reporting (Sinharay et al., 2010).

LCDM subscore validity.

DCMs differ from many psychometric models due to the categorical rather than continuous nature of the latent variables. It is this categorical nature that allows for a mastery versus nonmastery distinction (Rupp et al., 2010; Templin & Bradshaw, 2013). The DCMs define the mastery level of an attribute for an individual based on the answers he or she gives to test questions (Henson et al., 2009). Instead of theorizing a single proficiency continuum, DCMs see ability as distinguishable but interrelated knowledge within a domain. This allows for a finer-grain assessment of the respondent's performance (de la Torre & Chiu, 2016; Madison & Bradshaw, 2015). These models are often used together with a Q-matrix providing stakeholders diagnostic information about the respondents (Chen et al., 2013).

The subscore output for LCDM is continuous dichotomous: mastery versus nonmastery with 0.5 set as the mastery level. The total score is continuous. The correlation model will be a 2x2 cross-classification between mastery and non-mastery, 1 and 0 (Corder & Foreman, 2014; Uebersax, 2015).

The development of the Q-matrix plays a substantial role in the validity of the DCM process. Content experts, researchers, teachers, and psychometricians all play a role in the development and verification of the Q-matrix (Henson et al., 2009; Madison & Bradshaw, 2015). The verification of the Q-matrix provides content validity to the assessment process, and if not specified correctly, the inferences made from the DCM application are not valid (Gierl, 2009; Madison & Bradshaw, 2015).

DCMs serve to support the validity of the interpretations with empirical evidence. The weight the various stakeholders place on each piece of evidence is dependent upon the belief

system of that stakeholder and the goals the stakeholder has for the test and subsequent subscores. By showing the mastery/nonmastery of the latent traits based on the items of the test pertaining to each attribute, DCMs are able to provide stakeholders with the individualization they need (Rupp et al., 2010).

Simulation

The purpose of using simulated data is to predict possible changes in the Haberman and LCDM methods given the various changes in data analyzed. Decisions were made to alter the number of respondents, the number of subscore items, and test dimensionality. It was also decided to have 100 replications (Maria, 1997).

These decisions were made to best answer the research question referring to the Log-linear Cognitive Diagnostic Model providing improved subscore information over the Classical Test Theory based subscore reporting methods, specifically those attributed to Haberman.

The simulation varies the sample sizes, subtest sizes, and dimensionality. This will allow for examination of the reliability and correlations to determine which, if any, of the methods amplify the available information for the stakeholders in a psychometrically valid fashion.

The data was produced as strictly dichotomous data from the Rasch Model and from an LCDM. The options were 0 for incorrect and 1 for correct. Both models produced a probability for each item for each participant. The 0 or 1 was based on this probability of answering the item correctly. Dichotomous data was selected due to the scoring of most tests of this format, either a 1 for correct or 0 for incorrect. Partial credit is rarely given for incorrect answers.

Initial data from a past certification test was to be used for the study. This data came from an adaptive test format. The CTT subscore models are not available for use with adaptive testing. With CTT models the number of subtest items must remain static. Therefore, the simulation produced a test with a total of 30 items: Subtest 1 having 6 items, subtest 2 having 10 items, and

subtest 3 having 14 items. Using the original data would have meant not using the CTT models but changing to IRT based subscore methods. This study was designed to compare LCDM with CTT, a simulation at this point in the study was the best option.

The use of three attributes or traits is similar to the original data and to that of the licensure test mentioned as an example. The original data contained six attributes and the example above contained four attributes. If a respondent were to fail this simulated test he/she would most likely want to know which areas (attributes) he/she needed to work on the most to have the best chance of passing a future attempt. This is similar to the original data containing six attributes and the example above of four attributes. Having the best possible information, high reliability with low correlation, on which attributes to study is what these stakeholders are looking for. Using LCDM or a CTT method may assist these stakeholders in this decision-making process.

Chapter Three: Methods

Data Summary

Four areas will be examined in the methods section:

1. Model. Unidimensional and multidimensional data will be produced from a covariance matrix using four covariances. The Rasch model and LCDM will be used to generate four data sets each.
    - The covariances for each model are found in Appendix 3, Table 12 and Table 13
2. Dimensionality: unidimensional and multidimensional.
3. Sample sizes of 300 and 1200.
4. Subtest size. All tests will have a total test size of 30. The subtests will be 6, 10, and 14 items.

There will be 100 replications of each sample size, reliability, dimensionality, and model. See Appendix 3, Table 14 and Table 15 for a summary of the test specifications.

Models.

Data for test A_1, A_2, A_3, and A_4 will use the Rasch model. The Rasch model is a 1-parameter logistic (1-PL) model used to calculate the probability of and individual obtaining a positive (correct) dichotomous answer having a given theta (ability level) on a question with a given difficulty level (b parameter):

$$P(1|\Theta) = \exp(\Theta-b) / (1 + \exp(\Theta-b))$$

$P(1|\Theta)$: the probability of a correct answer with the given theta, latent trait ability level

Θ: theta, the latent trait ability level

b: the difficulty level of the item (Andrich, 1988)

The distributions for the Rasch models are as follows. The theta distributions will be random multivariate normal with a mean of zero and a covariances of 1, .9, .6, and .3. The distribution is multivariate due to the three attributes within the test itself. This provides for three possible theta levels when producing the data.

The b parameter will have a distribution of random normal with a mean of zero and a standard deviation of one. This allows for one difficulty level within the test itself. As seen below, it is the fluctuation in theta level that causes the fluctuation in reliability. The b parameter does not influence the reliability of Rasch Model data. Below is the reason for using a multivariate theta distribution and random normal beta distribution for the Rasch model parameters. It is the theta value that affects the reliability of the Rasch model, not the beta value.

Rasch model reliability estimates.

ρ^2: reliability

σ^2_T: true score variance

σ^2_E: error variance

SE: Standard Error

N: sample size

$I(\Theta)$: Item Information for a given theta

$I_j(\Theta)$: Information for item j with given theta

Θ : theta, ability parameter

b : item difficulty parameter

$\rho^2 = \sigma^2_T / (\sigma^2_T + \sigma^2_E)$

$\sigma^2_E = \Sigma SE^2 / N$

$SE = 1 / \sqrt{(I(\Theta)}$

$I(\Theta) = \sum I_j(\Theta)$

$I_j(\Theta) = 1 / [\exp(\Theta\text{-}b_j)]*[1 + \exp(-(\Theta\text{-}b_j))]^2$

- If σ^2_E increases, reliability decreases
- If σ^2_E decreases, reliability increases
- If SE increases, σ^2_E increases
- As $I_j(\Theta)$ decreases, SE increases (Skorupski, 2017)

The covariance matrices were constructed starting with a correlation matrix. The correlations were chosen based on the type of dimensionality of the test. By definition unidimensional (one dimension) tests do not have subtest correlations because they are unidimensional. In academic as well as licensure tests, unidimensional tests may have subsets of questions that come together as a subtest even under an unidimensional umbrella.

Multidimensional tests correlations were chosen based on the idea that the various attributes that make up the dimensionality of the test should not be highly correlated. To convert from the correlation to variance/covariance matrix, each cell was multiplied by the standard deviation of each attribute. For consistency, the standard deviation of each attribute was set at 1. This made the correlation matrix for each data set, A_1 through A_4, the variance/covariance matrix. The correlations, which in turn became the covariances, were chosen to demonstrate a variety of total test and subtest reliabilities and correlations. A unidimensional test has a covariance of one; both A_1 and B_1 are the unidimensional tests.

The covariance matrices produce a variety of subscore reliability estimates that can be seen in Table 14 and produce a variety of correlations between subscores that will be seen in Chapter 4 and Appendices 5 and 9.

A second full set of data will be produced using LCDM Parameters will be estimated to produce the separate unidimensional and multidimensional data sets:

$$P(X_{ij}=1|\boldsymbol{\alpha}_i) = \exp(\lambda_{i0} + \boldsymbol{\lambda}^T_j\mathbf{h}(\boldsymbol{\alpha}_i,\mathbf{q}_j))/1+ \exp(\lambda_{i0} + \boldsymbol{\lambda}^T_j\mathbf{h}(\boldsymbol{\alpha}_i,\mathbf{q}_j))$$

LCDM parameter estimate distributions.

λ_{i0} Intercept, a random uniform distribution between -2 and 0

$\boldsymbol{\lambda}^T_j$ main effects, a random uniform distribution between 0 and 4

$\mathbf{h}(\boldsymbol{\alpha}_i,\mathbf{q}_j)$: $\boldsymbol{\alpha}_i$,attribute class vector, $\mathbf{q}_j$, Q-matrix

h: random multivariate normal vector coinciding with the Q-matrix item to attribute assignments

Dimensions.

Unidimensional and multidimensional.

The dimensionality of a test often influences the usefulness and interpretation of the subscores. When testing a specific construct, such as reading ability, it can be broken into multiple attributes and each of these mapped onto its own scale; this would be a multidimensional testing procedure. Or, the construct can be mapped onto a single scale. This latter option is a unidimensional testing of the construct. The internal structure of test dimensionality (unidimensional, one construct, versus multidimensionality, two or more constructs) can provide additional evidence when looking at test reliability issues and deciding whether or not to report subscores (Ling, 2009).

Unidimensional tests are designed to test one attribute. All of the questions on the test relate to a single test score describing the level of knowledge a person has in this one area (Andrich, 1988). Unidimensional tests are designed to produce one total test score and one total

test reliability. Subscores and subscore reliabilities are questionable at best in unidimensional tests and add little if any value to the total score.

Multidimensional tests have specific subtests within the overall test itself that are designed to assess lesser attributes within an overall attribute or construct. If it is a multidimensional design, then subscore information may be more informative than the total score itself (Monaghan, 2006).

Sample sizes.

The sample sizes of 300 and 1200 were chosen to examine the effect sample size has on the various subscore methods. The CTT based Haberman methods have proven to work at both small and large sample sizes. Does the sample size affect the reliability of the LCDM and does the model have difficulty converging at the lower sample sizes?

Subtest sizes.

The subtest sizes of 6, 10, and 14 were chosen to look at how the number of subtest items effected the subscore output estimates. Starting with a reliability estimate for a subtest size of 6, how does the increase to 10 and 14 items affect the LCDM reliability estimates? The reliability estimates of the Haberman methods should increase as the number of items increases. This can be demonstrated with the Spearman-Brown Prophecy:

$$k = \rho_{cc'}(1 - \rho_{ii'})/ \rho_{ii'}(1 - \rho_{cc'})$$

The variable *k* is the factor of increase (or decrease) needed for a test with reliability = $\rho_{ii'}$ to have a reliability = $\rho_{cc'}$ (Raykov & Marcoulides, 2011).

Q-matrix.

The Q-matrix developed for each test will be a simple Q-matrix. It will be the same Q-matrix for each test. Individual items will represent only one of the three attributes. There will be no interactions. See Appendix 2.

Subscore Methods

In this study, there are six approaches of looking at subscore reporting: 1. Raw subscore, 2. differential scoring, 3-5. Three Haberman methods (methods 2, 3, and 4), and 6. LCDM.

Initial reliability scores for each attribute will be calculated and reported using Cronbach's Alpha:

$$\alpha = p/p\text{-}1 = \sum Cov(X_i,X_j) / Var(X)$$ (Raykov & Marcoulides, 2011)

Correlation calculations will use the Pearson product-moment coefficient of correlation (Pearson's r),

$$r = \sum(X\text{-}E(X))*(Y\text{-}E(Y)) / S_X S_Y$$

where E(X) and E(Y) are the mean values of X and Y, and S_X and S_Y are the standard deviations of X and Y (Coladarci & Cobb, 2014)

Standard deviations for each method will use the following:

$$SD = \sqrt{(\sum(X\text{-}Xbar)^2/(n\text{-}1))}$$

where X is the score, Xbar is the mean of the scores and n is the number of scores (Coladarci & Cobb, 2014)

This is used as a measure of the variance of the reliability and correlation scores. A smaller standard deviation is desired among scores. A larger variance and therefore larger standard deviation demonstrates a wider spread among scores and is not as consistent. A smaller variance and smaller standard deviation shows a closer connection between scores and more consistency in the results (Coladarci & Cobb, 2014).

Raw subscores.

Raw subscores are the most basic subscore reporting method. It may be the easiest to produce and the easiest to explain to stakeholders. Reliability will be calculated using Cronbach's Alpha, and Pearson's r will be used for correlation calculations.

Differential scoring.

The calculation methods for this method are the same as those found in the raw subscores method. Reliability will use Cronbach's Alpha and Pearson's r for calculating correlations. The difference is in the calculating of the actual subscores and total score. The calculation for scoring using this type of differential scoring is as follows:

> ...let the positive weight equal q, the proportion of examinees failing the item, and let the negative weight equal -p, where p is the proportion passing the item. Thus a difficult item passed by only .05 of the examinees would be scored .95 if passed and -.05 if failed. The mean score for each item over all examinees is qp + (-pq) = 0 and thus the mean test score for all examinees is also zero,... Although this weighting scheme is not being recommended, it is logically more defensible than simply assigning weights according to difficulty (Stanley & Wang, 1968).

Haberman methods.

Haberman symbol definitions.

- α: Reliability of the subscores, Cronbach's Alpha
- E(S): Mean of the observed subscores for an individual attribute
- S: Observed subscore for an individual attribute
- E(X): Mean of the observed total scores
- X: Observed total score
- S_t: Estimate of the true subscore for an individual attribute
- S_s: Estimate of the true subscore using method 2
- S_x: Estimate of the true subscore using method 3
- c: constant used in method 3, this combines both subscores and total scores
- S_{sx}: Estimate of the true subscore using method 4

- β: Linear regression coefficient for the subscores, method 4
- γ: Linear regression coefficient for the total scores, method 4

Cronbach's Alpha is used for all estimated reliability calculations (Sinharay & Haberman, 2008). Pearson's r will be used for all correlation calculations.

Method 1.

- $S_t = E(S)$
- This is the least complex of the four models. The mean of the subscore is used as the true subscore. With this method, each of the respondents would have the same subscore.
- Mean Squared Error, $MSE = E[S_t - E(S)]^2 = \sigma^2(S_t)$
- There is no proportional reduction in Mean Squared Error (MSE)
- In the Haberman package in R there is no output given for this method.
- This method will not be used in the study.

Method 2.

- $S_s = E(S) + \rho^2(S_t,S)[S - E(S)]$
- This model uses the subscores as the predictor of the true subscore estimate.
- $MSE = E[S_t - S_s]^2 = \sigma^2(S_t)[1- \rho^2(S_t,S)]$
- The MSE calculation from this method is the starting value for PRMSE comparisons. A reduction from this value is considered an improvement in subscore estimation.
- The PRMSE is reduced to the reliability of the observed subscores to the true subscores, $\rho^2(S_t,S)$.

Method 3.

- $S_x = E(s) + c[x - E(x)]$

- Model 3 is similar to Model 2. The observed total score(s) are used as the predictor of the true subscore estimate.
- $c = \rho(S_t,X)[\sigma(S_t)/\sigma(X)]$
- $MSE = E[S_t - S_x]^2 = \sigma^2(S_t)[1- \rho^2(S_t,X)]$
- The PRMSE is reduced to the reliability of the true subscores to the observed total scores, $\rho^2(S_t,X)$.

Method 4.

- $S_{sx} = E(S) + \beta[S - E(S)] + \gamma[X - E(X)]$
- Model 4 is the most complex of the four methods as it uses both subscores and total scores to predict the true subscore.
- $\tau = [\rho(X_t,X)\rho(S_t,X_t) - \rho(S,X)\rho(S_t,S)] / [1 - \rho^2(S,X)]$
- $\gamma = [\sigma(S)/\sigma(X)]*[\rho(S_t,S)\tau]$
- $\beta = \rho(S_t,S)[\rho(S_t,S) - \rho(S,X)\tau]$
- $MSE = E[S_t - S_{sx}]^2 = \sigma^2(S_t)[1 - \rho^2(S_t,S) - \tau^2[1 - \rho^2(S,X)]]$
- The reliability calculation for this model used in the PRMSE calculation is as follows: $1-[1 - \rho^2(S, S_t)][1 - \rho^2(X, S_t*S)]$.

(Dai et al., 2017; S. Haberman et al., 2009; S. J. Haberman, 2008; Sha & McCoy, 2014; Sinharay & Haberman, 2008; Sinharay, Puhan, et al., 2011)

Log-linear Cognitive Diagnostic Model (LCDM)

The LCDM is a classification model with dichotomous independent variables. Instead of a subscore as in previous methods, it produces a percent probability of mastery for each attribute.

Input.

The input for the LCDM analysis consists of two matrices. The first is the Q-matrix. This was discussed in Chapter Two. Because each item is only assigned to a single attribute, the Q-

matrix will be a simple, not complex, Q-matrix. There will be no interactions to consider. The second is a matrix of respondents by items. In this data sample, each respondent's row will contain 0s (answered incorrectly) or 1s (answered correctly) for each of the 30 items. The Q-matrix is found in Appendix 2; the respondent to item matrices will be generated by LCDM.

R and Mplus.

The LCDM package uses the R scripts from Dr. Andre A. Rupp and Dr. Oliver Wilhelm; these are available from

http://www.education.umd.edu/EDMS/fac/Rupp/R%20Files%20for%20Mplus%20Input%20File%20Generation.zip

Additionally, some scripts have been modified or created by Dr. Jonathan Templin. This is the LCDM routine used for this research.

(1) Formats the data in accordance with the R functions expectations

(2) Writes the Mplus script

(3) Runs Mplus using the MplusAutomation package

(4) Collects Mplus output

(5) Converts Mplus output to LCDM parameters

(6) Estimates attribute reliability (Templin and Bradshaw, 2013)

(7) Creates EAP and MAP estimates for attribute profiles and marginal attributes

a) MAP (Maximum a posteriori): One method of assigning latent class attribute probability profiles. Can be difficult to interpret due to no direct probability estimates for each individual attribute for each person.

b) EAP (Expected a posteriori): A second method of assigning latent class attribute probabilities. This is more suitable due to the estimated expected value for each individual attribute for each person (Rupp et al., 2010).

There are three specific pieces that are needed for the attribute probability analysis:

1. X_r : the respondent's item responses
2. π_{ic} : the manner in which the other respondents with the same attribute vector respond to each item
3. υ_c : the proportion of respondents having the same attribute profile in the population

These three pieces are put together to produce the probability that a respondent has mastered a specific attribute.

Attribute probability symbol definitions.

- $P(X_{i1}=1|\alpha_1)$: The probability of getting a correct answer for item 1 that is matched with attribute 1
- α_{rc} : the posterior probability that r belongs to c for the attribute vector $\boldsymbol{\alpha_c}$
- $\prod$: the product over all I items
- X_{ir} : the observed response of respondent r to item i
- υ_c : probability of any respondent belonging to latent class c
- π_{ic} : the response probability of a respondent in latent class c for item i
- μ_c : the kernel: estimates the probability density function
- $\lambda_{i,0}$: the intercept, a respondent who obtains zero correct for vector $\boldsymbol{\alpha_c}$
- $\boldsymbol{\alpha_c}$: vector for a attribute class c
- $\boldsymbol{\lambda_i}$: vector for item i
- $\mathbf{q_i}$: the set of Q-matrix entries for item i

LCDM steps.

- $P(X_{i1}=1|\alpha_1) = \exp(\lambda_0 + \lambda_{11}\alpha_1)/1+ \exp(\lambda_0 + \lambda_{11}\alpha_1)$
    - λ_0 : the log odds of a respondent who obtains zero correct for all of the attributes
    - λ_{11} : the log odds of item 1 attribute 1

- α_1 : the log odds of attribute 1
- $\lambda_{11}\alpha_1$: the log odds of person i correctly answering Item 1 Attribute 1
- λ : the weights that represent the changes in the predicted values

- $\mu_c = \sum\gamma_{1(a)}\alpha_{ca} + \ldots$. (all main effects and interactions for that α_c)
- $\upsilon_c = \exp(\mu_c)/\sum \exp(\mu_c)$ for c=1, 1 to c
  - Using μ_c the probability of any respondent belonging to a specific latent class is calculated.
- $\pi_{ic} = \exp(\lambda_{i,0} + \boldsymbol{\lambda_i^T h(\alpha_c + q_i)})/[1 + \exp(\lambda_{i,0} + \boldsymbol{\lambda_i^T h(\alpha_c + q_i)})]$
  - Calculate to see the individual probability of a respondent answering item i correctly while in latent class c.
- $\alpha_{rc} = \upsilon_c \prod \pi_{ic}^{Xir}(1- \pi_{ic})^{1-Xir} / \Sigma\, \upsilon_c \prod \pi_{ic}^{Xir}(1- \pi_{ic})^{1-Xir}$
  - This final model puts everything together to calculate the posterior probability that the respondent belongs to class c given the attribute vector for c (Henson et al., 2009; Templin & Bradshaw, 2013)

LCDM correlation.

The Mplus syntax will produce the probability of attribute mastery and the reliability for each attribute. See #6 in the "This routine" section of the R script listed previously.

Tetrachoric correlation.

Correlation for the LCDM will be the tetrachoric correlation. This measures binary data that are fundamentally continuous: it gives what the correlation would be if it were measured on a continuous scale.

A two by two contingency table is developed. See Table 1. After the probability of mastery for an individual is determined, two draws are made from the posterior distribution of the individual and compared to the original probability. There are four possibilities for each pair

of drawings: 1-1, 1-0, 0-1, or 0-0 (1 is correct, 0 is incorrect). Once this has been completed for all the respondents for that particular attribute, the cosine formula below is used to calculate the correlation, which becomes the correlation of the LCDM for the attributes.

Table 1Tetrachoric Correlation 2x2 Contingency Table

	1	0
1	Σ(p*p)/N **A**	Σ(p*(1-p))/N **B**
0	[(Σ(1-p))*p]/N **C**	[Σ(1-p)(1-p)]/N **D**

p: The probability of attribute mastery

$r = \cos(180(1 + \sqrt{[(BC)/(AD)]}))$

("More correlation coefficients," 2016; Templin & Bradshaw, 2013)

LCDM Reliability

Polychoric correlation.

Reliability for the LCDM will be the polychoric correlation. This measures binary data that are fundamentally continuous: it gives what the correlation would be if it were measured on a continuous scale. r is always positive and always between 0 and 1.

An 8 by 8 contingency table is developed. After the probability of membership in one of the eight potential classes for an individual is determined, two draws are made from the posterior distribution for each class for each individual and compared to the original probability which becomes the reliability of the LCDM for the attribute.

Reporting Methods

Research question.

Do the subscore estimates provided by a LCDM (Henson et al., 2009) provide an improvement (more reliable scores, lower correlations between scores, magnify subscore information in a beneficial manner) over Classical Test Theory (CTT) based reporting such as the raw subscore reporting, differential scoring, and Haberman methods #2, #3, and #4 (Sinharay & Haberman, 2008) used in this study?

The improvement will be measured using reliability estimates, correlation estimates, and standard deviations. Does the LCDM method magnify the subscore information in a more beneficial manner than the raw scores or the augmented Haberman scores? This is in large part a question of method validity.

Validity.

As mentioned earlier in the paper, the reliability of the subscores should be .85 or above, (Ling, 2009), and the correlation between one subscore and the others as well as between the subscore and the total score are of prime importance. The question of difference in validity presents itself with the methods of producing subscore estimates. The most accurate subscore estimates, those that provide the most accurate information, are those that produce the highest possible reliability with the lowest possible correlation to other subscores and/or the total score (Babenko & Rogers, 2014; Feinberg & Wainer, 2014b; Haladyna, 2004; Sinharay et al., 2010; Skorupski & Carvajal, 2010).

Raw scores.

The raw score provides raw subscores, which are the number correct out of the total possible for each subscore. Looking singularly at the raw score, the reliability of the individual

subscores as well as the correlations between subscores and total scores are possible. This is the least evasive and costs the least amount in time and effort to explore.

Differential scoring.

The differential score methods are similar to those of the raw subscores. The exception is a priori adjustments made to the respondents scores. This specific adjustment can be found in Chapter 2**.** Differential scoring is an attempt to adjust for the difficulty of the item based on the proportion of respondents answering the item correctly. The reliability and correlation calculation methods are the same as the raw subscore method.

Haberman methods.

Haberman methods provide transformed subscores (attributes). The Haberman methods provide three possible altered subscores above and beyond the raw score for each subscore category. There is a procedure involved, PRMSE, to help the stakeholder choose the subscore of most value. The Haberman package in R provides output that identifies the subscore of most value using the PRMSE as the guide.

The Haberman methods increase the reliability of the subscore by borrowing from subscores (Method 2), total scores (Method 3), or both simultaneously (Method 4). This association most often in fact does increase the reliability of the subscore, but, at the cost of a potential increase in correlation between the estimated subscores and the attribute and/or total score used in the process. Correlations between these data will be examined to see if the procedure may actually decrease the potential information gained from the original raw score process.

LCDM.

LCDM provides a probability of percent mastery by the respondent of the attribute. The differential method makes an attempt to take into account the difficulty of the individual items in

a linear manner. The LCDM process does this more comprehensively by looking at all the item scores of all the respondents. LCDM considers the difficulty of the items answered correctly. This provides a more informative picture of what the respondent has successfully mastered.

Final comparisons.

The method that produces the highest reliabilities coupled with the lowest correlations should be the method that is the most valid and the method that provides stakeholders with the most information about the particular subscores/attributes in question. As has been mentioned in the literature review and in the reporting methods section, these are the characteristics of subscores that provide the most accurate information (Feinberg & Wainer, 2014b; Haladyna, 2004; Sinharay et al., 2010; Skorupski & Carvajal, 2010).

Data Collection

Data that will be collected and analyzed is presented in the following paragraphs. Given the 16 simulated tests outlined at the beginning of this chapter, the following data will be produced, captured and analyzed: total reliability and subscore reliability for each test for each method and subtest correlations for each test for each method.

The data comparisons have 100 replications for each sample size, dimension, model, and initial reliability. The mean of each 100 replications is calculated to find the reported reliability, correlation, and standard deviations.

Raw scores.

- Reliability will be calculated using Cronbach's Alpha
    - Total test reliability and standard deviation
    - Individual reliability for each of the three attributes and standard deviations
- Correlations between each of the attributes
    - Pearson's r will be used for all correlation calculations

Differential scoring.

- Reliability will be calculated using Cronbach's Alpha.
    - Total test reliability and standard deviation
    - Individual reliability for each of the three attributes and standard deviations
- Correlations between each of the attributes
    - Pearson's r will be used for all correlation calculations

Haberman methods (#2, #3, and #4).

- Reliability will be calculated using Cronbach's Alpha.
    - Total test reliability and standard deviations (New attribute reliabilities used to calculate a new composite total reliability)
    - Individual reliability and standard deviations for each of the three attributes
- Correlations between each of the attributes for each method
    - Pearson's r will be used for all correlation calculations.

LCDM.

- Reliability estimates for the attributes will use polychoric correlation
    - Individual reliability and standard deviations for each of the three attributes
- Correlations between each of the attributes
    - Tetrachoric correlation between each attribute with standard deviation
- Nonconvergent data sets within each set of 100 repetitions will be removed prior to final calculations of reliability and correlation.
    - Convergence demonstrates the algorithm has found a parameter setting in which the next possible iteration falls within a certain predetermined error space. This error space is determined by default by the algorithm.

  - An appropriate error space leads to good model fit. Nonconvergence often indicates poor model fit.

- Model Fit
  - Rasch model: The b parameters of the first data set of 1200Cov1 data will be correlated to the item difficulties produced by Mplus using the first set of 1200Cov1 data.
  - LCDM: The intercepts and main effects of the first set of 1200Cov1 will be used to produce a second set of data. The correlation between these two sets will be compared.
  - The correlation between data sets in both models should be 1.

Methods summary.

- Simulate data sets using the Rasch Model
  - 2 unidimensional
  - 6 multidimensional
  - Total test items of 30
    - Subtest lengths of 6, 10, and 14
  - Sample sizes of 300 and 1200
- Simulate 8 data sets using the log-linear cognitive diagnostic method
  - 2 total unidimensional
  - 6 multidimensional
  - Total test items of 30
    - Subtest lengths of 6, 10, and 14
  - Sample sizes of 300 and 1200

- Subscore methods
    - Raw subscores
    - Differential scoring
    - Haberman methods 2, 3, and 4
    - Log-linear Cognitive Diagnostic Method
- Data gathered
    - Reliability estimates for all subtests and total tests for all 16 data sets
    - Correlations between subtests for all 16 data sets
    - Reliability and correlation comparisons between all data sets
    - Model fit statistics
        - Rasch model data set: 1200 participants, covariance of 1
        - LCDM data set: 1200 participants, covariance of 1

To produce a subscore that provides additional information above and beyond the total test score, the subscore reliability should be at or above 0.85 (Ling, 2009). Which of the methods will produce a psychometrically acceptable subscore reliability? The second aspect is the correlation between the subscores. The higher the correlation between the two subscores, the closer the test becomes to being unidimensional and the less value the subscores have at providing additional information (Feinberg & Wainer, 2014b).

Chapter Four: Results

Results

The purpose of the paper is to psychometrically investigate if LCDM subscore analyzation provides an improvement in subscore reporting quality compared to the CTT subscore methods. Specifically, this is reporting raw subscores, using a specific a priori differential method, and three of the four Haberman methods of post hoc augmentation reporting.

The simulation produced reliability scores for each method and data set as well as correlation scores for each method and data set. Ling recommends not reporting a subscore with a reliability of less than 0.85 (Ling, 2009). It is also recommended that high correlation between subscores reduces the potential information that can be gathered from the individual subscore (Sinharay et al., 2010; Skorupski & Carvajal, 2010)

Appendices 4 through 23 give general results for these 16 data sets produced by two methods (the Rasch model and LCDM) and analyzed for reliability and subtest correlation by six subscore analysis methods: LCDM, raw analysis, analysis following an a priori linear differential adjustment, and Haberman methods 2, 3, and 4.

Data Generation

Rasch data simulation model.

The Rasch model is a 1-PL logistic model used to calculate the probability of an individual obtaining a positive (correct) dichotomous answer having a given theta (ability level) on a question with a given difficulty level (b parameter) (see Chapter 3):

$$P(1|\Theta) = \exp(\Theta\text{-}b) / (1 + \exp(\Theta\text{-}b))$$

The distributions for the Rasch models are as follows: the theta distributions were random multivariate normal with a mean of zero, a SD of 1, and covariances of 1, .9, .6, and .3.

Multivariate normal was used due to the three attributes and the three possible theta levels involved with these attributes. This allows for using the Rasch model to simulate multidimensional data. The b parameter had a distribution of random normal with a mean of zero and a standard deviation of one. There is only one difficulty parameter with a Rasch model so only one distribution will be used.

LCDM data simulation model.

$$P(X_{ij}=1|\boldsymbol{\alpha}_i) = \exp(\lambda_{i0} + \boldsymbol{\lambda}^T_j\mathbf{h}(\boldsymbol{\alpha}_i,\mathbf{q}_j))/1+ \exp(\lambda_{i0} + \boldsymbol{\lambda}^T_j\mathbf{h}(\boldsymbol{\alpha}_i,\mathbf{q}_j))$$

λ_{i0} Intercept, a random uniform distribution between -2 and 0

$\boldsymbol{\lambda}^T_j$ main effects, a random uniform distribution between 0 and 4

$\mathbf{h}(\boldsymbol{\alpha}_i,\mathbf{q}_j)$: $\boldsymbol{\alpha}_i$,attribute class vector, $\mathbf{q}_j$, Q-matrix

h: random multivariate normal vector coinciding with the Q-matrix item to attribute assignments

Analysis Methods

Raw.

Cronbach's Alpha was used for reliability. Pearson's r was used for correlation (see Chapter 3).

Differential.

Cronbach's Alpha was used for reliability. Pearson's r was used for correlation (see Chapter 3).

Haberman 2, 3 and 4 (see Chapter 2).

Method 2.

- $S_s = E(S) + \rho^2(S_t,S)[S - E(S)]$

Method 3.

- $S_x = E(s) + c[x - E(x)]$

Method 4.

- $S_{sx} = E(S) + \beta[S - E(S)] + \gamma[X - E(X)]$

(Dai et al., 2017; S. Haberman et al., 2009; S. J. Haberman, 2008; Sha & McCoy, 2014; Sinharay & Haberman, 2008; Sinharay, Puhan, et al., 2011)

LCDM analysis method.

LCDM steps (see Chapter 2).

- $P(X_{i1}=1|\alpha_1) = \exp(\lambda_0 + \lambda_{11}\alpha_1)/1+ \exp(\lambda_0 + \lambda_{11}\alpha_1)$
- $\mu_c = \sum\gamma_{1(a)}\alpha_{ca} + \ldots$ (all main effects and interactions for that α_c)
- $\upsilon_c = \exp(\mu_c)/\sum \exp(\mu_c)$ for c=1, 1 to c
- $\pi_{ic} = \exp(\lambda_{i,0} + \boldsymbol{\lambda_i^T h(\alpha_c + q_i)})/[1 + \exp(\lambda_{i,0} + \boldsymbol{\lambda_i^T h(\alpha_c + q_i)})]$
- $\alpha_{rc} = \upsilon_c \prod \pi_{ic}^{Xir}(1- \pi_{ic})^{1-Xir} / \Sigma\, \upsilon_c \prod \pi_{ic}^{Xir}(1- \pi_{ic})^{1-Xir}$

(Henson et al., 2009; Templin & Bradshaw, 2013)

Reliability methods.

Classical test theory.

Initial reliability scores for each attribute were calculated and reported using Cronbach's Alpha (see Chapter 3).

$$\alpha = p/p\text{-}1 = \sum Cov(X_i,X_j) / Var(X) \text{ (Raykov \& Marcoulides, 2011)}$$

LCDM.

Polycoric correlation.

Reliability for the LCDM was the polychoric correlation. This measures binary data that are fundamentally continuous: it gives what the correlation would be if it were measured on a continuous scale. (see Chapter 2).

Correlation methods.

Classical test theory.

Correlation calculations used the Pearson product-moment coefficient of correlation (Pearson's r) (see Chapter 3).

$$r = \sum(X-E(X))*(Y-E(Y)) / S_X S_Y$$

(Coladarci & Cobb, 2014)

LCDM.

Tetrachoric correlation.

Correlation for the LCDM was the tetrachoric correlation. This measures binary data that are fundamentally continuous: it gives what the correlation would be if it were measured on a continuous scale. (see Chapter 3).

Model fit.

Rasch model: The b parameters of the first data set of 1200Cov1 data were correlated to the item difficulties produced by Mplus using the first set of 1200Cov1 data. This resulted in a correlation of .997

LCDM: The intercepts and main effects of the first set of 1200Cov1 were used to produce a second set of data. The correlation between the two sets of data was 1.

Standard deviation.

This is used as a measure of the variance of the reliability and correlation scores. A smaller standard deviation is desired among scores. A smaller variance and smaller standard deviation shows a closer connection between scores and more consistency in the results (see Chapter 3).

$$SD = \sqrt{(\sum(X-Xbar)^2/(n-1))}$$ (Coladarci & Cobb, 2014)

Convergence rates.

Table 2 shows the convergence rates for the LCDM Analysis (see Chapter 3).

Table 2 Convergence Rates for the LCDM Analysis

Convergence Rates LCDM Analysis		
Data Set	LCDM Data	Rasch Data
300Cov1	59%	34%
300Cov.3	100%	90%
300Cov.6	100%	82%
300Cov.9	98%	37%
1200Cov1	57%	44%
1200Cov.3	100%	100%
1200Cov.6	100%	100%
1200Cov.9	100%	87%

Data

The data for this simulation are presented by reliability estimates followed by correlation estimates. Standard deviations are presented at the end of the chapter. Unless additional data is needed, the data examples shown will be 300Cov.6 and 1200Cov.6 (Sample size of 300 and 1200 with a covariance of 0.6). For the full set of data outcomes, see Appendices 4 through 23.

Reliability estimates.

Appendix 3, Tables 14 and 15 show a summary of the raw score method of calculating total test as well as subscore reliability estimates for the 16 different data sets. As the covariance between attributes increased, the reliability increased. Also, as the sample size increased across attributes, the individual attribute reliability estimate increased:

$$\alpha = p/p\text{-}1 = \sum Cov(X_i,X_j) / Var(X)$$

Table 3 shows the total score reliability estimates using the Rasch data and the LCDM data.

Table 3 Rasch Data and LCDM Data Total Score Reliability Estimates

Total Score Rasch Data				Total Score LCDM Data			
	Raw RelTot	Dif RelTot	Hab RTot		Raw RelTot	Dif RelTot	Hab RTot
300Cov.6	0.792	0.795	0.792	300Cov.6	0.774	0.783	0.774
SD	0.020	0.019	0.020	SD	0.042	0.040	0.042
300Cov.9	0.833	0.835	0.833	300Cov.9	0.824	0.831	0.824
SD	0.016	0.016	0.016	SD	0.035	0.033	0.035
1200Cov.6	0.793	0.796	0.793	1200Cov.6	0.768	0.778	0.768
SD	0.010	0.010	0.010	SD	0.035	0.033	0.035
1200Cov.9	0.833	0.835	0.833	1200Cov.9	0.820	0.828	0.820
SD	0.007	0.008	0.007	SD	0.030	0.029	0.030

Tables 4 and 5 show the reliability estimates for the six subscore reporting methods using the Rasch data and the LCDM data. In both tables, the reliability estimates using the LCDM method are higher than those of the five CTT methods. The LCDM method has higher reliability estimates when analyzing the LCDM data compared to analyzing the Rasch model data.

Table 4 Rasch Data Reliability Estimates, Six Subscore Methods

Rasch Data									
	LCDM			Raw			Differential		
	RelA1	RelA2	RelA3	RelA1	RelA2	RelA3	RelA1	RelA2	RelA3
300Cov.6	0.853	0.908	0.945	0.513	0.637	0.719	0.515	0.640	0.720
SD	0.036	0.022	0.017	0.048	0.035	0.030	0.047	0.034	0.029
1200Cov.6	0.835	0.894	0.931	0.524	0.643	0.715	0.527	0.645	0.718
SD	0.019	0.015	0.010	0.025	0.022	0.015	0.025	0.022	0.014

	Haberman 2			Haberman 3			Haberman 4		
	RM2A1	RM2A2	RM2A3	RM3A1	RM3A2	RM3A3	RM4A1	RM4A2	RM4A3
300Cov.6	0.513	0.637	0.719	0.499	0.575	0.666	0.618	0.693	0.751
SD	0.048	0.035	0.030	0.066	0.047	0.038	0.038	0.029	0.026
1200Cov.6	0.524	0.643	0.715	0.502	0.578	0.663	0.623	0.696	0.748
SD	0.025	0.022	0.015	0.032	0.024	0.018	0.019	0.018	0.012

Table 5 Reliability of LCDM Data, Six Subscore Methods

LCDM Data									
	LCDM			Raw			Differential		
	RelA1	RelA2	RelA3	RelA1	RelA2	RelA3	RelA1	RelA2	RelA3
300Cov.6	0.930	0.979	0.991	0.538	0.662	0.719	0.548	0.672	0.729
SD	0.049	0.020	0.010	0.138	0.090	0.078	0.135	0.088	0.073
1200Cov.6	0.915	0.972	0.991	0.526	0.645	0.723	0.536	0.657	0.733
SD	0.060	0.024	0.009	0.129	0.093	0.063	0.126	0.090	0.059

	Haberman 2			Haberman 3			Haberman 4		
	RM2A1	RM2A2	RM2A3	RM3A1	RM3A2	RM3A3	RM4A1	RM4A2	RM4A3
300Cov.6	0.538	0.662	0.719	0.388	0.476	0.589	0.600	0.689	0.737
SD	0.138	0.090	0.078	0.074	0.063	0.069	0.100	0.075	0.069
1200Cov.6	0.526	0.645	0.723	0.362	0.460	0.588	0.578	0.672	0.738
SD	0.129	0.093	0.063	0.050	0.057	0.051	0.103	0.080	0.056

The raw method, the a priori differential method, and the Haberman 2 (H2) method show lower reliability estimates that do not compare favorably with the remaining three methods. The LCDM, the Haberman 3 (H3), and the Haberman 4 (H4) methods all show increased reliability estimates to increase subscore information and value (Sinharay et al., 2010). See Appendices 4, 6, 8, and 10 for complete data set and method reliability estimates.

Correlation estimates.

Table 6 shows Rasch data correlations from all methods. Table 7 demonstrates data correlation estimates from all six methods using the LCDM data sets. See Appendices 5, 7, 9, and 11 for complete data set and method correlation estimates.

The correlation estimates were identical to three decimal places for the raw, differential, and H2 methods. The correlation estimates for H3 are 1 for all data sets (see Appendices 5 and 9). The LCDM and H4 correlations are similar and both are higher than the remaining CTT methods, excluding H3.

Table 6 Correlations for Six Subscore Methods, Rasch Data

Rasch Data									
	LCDM			Raw			Differential		
	CorA1A2	CorA1A3	CorA2A3	CorA1A2	CorA1A3	CorA2A3	CorA1A2	CorA1A3	CorA2A3
300Cov.6	0.775	0.764	0.761	0.344	0.366	0.407	0.344	0.366	0.407
SD	0.099	0.116	0.087	0.045	0.057	0.048	0.045	0.057	0.048
1200Cov.6	0.783	0.776	0.743	0.350	0.370	0.407	0.350	0.370	0.407
SD	0.061	0.054	0.053	0.025	0.025	0.024	0.025	0.025	0.024
	Haberman 2			Haberman 3			Haberman 4		
	CM2A1A2	CM2A1A3	CM2A2A3	CM3A1A2	CM3A1A3	CM3A2A3	CM4A1A2	CM4A1A3	CM4A2A3
300Cov.6	0.344	0.366	0.407	1.000	1.000	1.000	0.765	0.776	0.759
SD	0.045	0.057	0.048	0.000	0.000	0.000	0.059	0.069	0.062
1200Cov.6	0.350	0.370	0.407	1.000	1.000	1.000	0.764	0.777	0.759
SD	0.025	0.025	0.024	0.000	0.000	0.000	0.031	0.032	0.029

Table 7 Correlations for Six Subscore Methods, LCDM Data

LCDM Data									
	LCDM			Raw			Differential		
	CorA1A2	CorA1A3	CorA2A3	CorA1A2	CorA1A3	CorA2A3	CorA1A2	CorA1A3	CorA2A3
300Cov.6	0.630	0.645	0.618	0.261	0.274	0.291	0.261	0.274	0.291
SD	0.105	0.095	0.079	0.060	0.068	0.055	0.060	0.068	0.055
1200Cov.6	0.632	0.629	0.611	0.245	0.260	0.283	0.245	0.260	0.283
SD	0.054	0.049	0.042	0.041	0.044	0.035	0.041	0.044	0.035
	Haberman 2			Haberman 3			Haberman 4		
	CM2A1A2	CM2A1A3	CM2A2A3	CM3A1A2	CM3A1A3	CM3A2A3	CM4A1A2	CM4A1A3	CM4A2A3
300Cov.6	0.261	0.274	0.291	1.000	1.000	1.000	0.577	0.602	0.563
SD	0.060	0.068	0.055	0.000	0.000	0.000	0.098	0.104	0.085
1200Cov.6	0.350	0.370	0.407	1.000	1.000	1.000	0.764	0.777	0.759
SD	0.025	0.025	0.024	0.000	0.000	0.000	0.031	0.032	0.029

Standard deviations.

Standard deviations (SD) varied among the six subscore methods. This was particularly apparent when comparing reliability and correlation estimates between the Rasch data and the LCDM data. Table 8 is a comparison of the SD for the raw method using both the Rasch data and LCDM data:

$$SD = \sqrt{(\sum(X\text{-}Xbar)^2/(n\text{-}1))}$$ (Coladarci & Cobb, 2014)

Using the 1200Cov.6, RelA2 output, the reliability is 0.645 and a SD of 0.093 for the LCDM data. A confidence interval of +/- 2 SD extends the possible reliability to between 0.459 and 0.831. (0.372)

For the Rasch data, the reliability is 0.643 with a SD of 0.022. Given +/- 2 SD, the confidence interval is 0.599 to 0.687. (a width of 0.088 SD) The reliability estimate for the Rasch data is within a smaller possible interval.

The SD are an important aspect of the evaluation process when the decision to add additional value to a subscore is being evaluated. The SD should always be examined for variance prior to making a subscore method decision.

Table 8 LCDM and Haberman 4 Reliability SD Comparison

	LCDM Raw				Rasch Raw			
	RelA1	RelA2	RelA3	RelTot	RelA1	RelA2	RelA3	RelTot
300Cov.6	0.538	0.662	0.719	0.774	0.513	0.637	0.719	0.792
SD	0.138	0.090	0.078	0.042	0.048	0.035	0.030	0.020
1200Cov.6	0.526	0.645	0.723	0.768	0.524	0.643	0.715	0.793
SD	0.129	0.093	0.063	0.035	0.025	0.022	0.015	0.010

The complete output from the six methods and 16 data sets can be found in Appendices 4 through 23. These data are presented in a variety of tables allowing the reader to compare methods using various criteria.

Chapter 5: Discussion

Discussion

This analysis was designed to compare the level of subscores produced by six different reporting methods: raw scores, a priori differential, post hoc Haberman methods 2, 3, and 4, and the post hoc LCDM method. It was specifically designed to compare the potential improvement of the LCDM method over the five CTT methods. Comparing the six methods, reliability and correlation estimates were the primary means of determining which method produced subscores of increased benefit to the stakeholder. The analysis design was determined by the following research question:

> Do the subscore estimates provided by a LCDM (Henson et al., 2009) provide an improvement (more reliable scores, lower correlations between scores, magnify subscore information in a beneficial manner) over Classical Test Theory (CTT) based reporting such as the raw subscore reporting, differential scoring, and Haberman methods #2, #3, and #4 (Sinharay & Haberman, 2008) used in this study?

Results of this study show the Haberman #4 method providing the most improvement over raw subscore reporting with the highest increase in reliability and with the most consistent correlation results using either Rasch data or LCDM data. Overall LCDM provides the most improvement in reporting subscores of either Rasch or LCDM data. This improvement is higher than the Haberman #4 reliability scores and correlations closer to the covariance matrix used to create the simulated data than any of the CTT methods. The study answers the research question positively in favor of the LCDM.

CTT methods.

There are five CTT methods used to analyze the simulated data. The first method is referred to as the raw score method. Cronbach's Alpha is used to calculate the reliability of the subscores and the total score (see Chapter 3). For all CTT models, reliability estimates increase as the number of items within the attribute increase:

$$\alpha = p/p\text{-}1 = \sum Cov(X_i,X_j) / Var(X) \text{ (Raykov \& Marcoulides, 2011).}$$

Pearson's r is used to calculate the correlations between the subscores (see Chapter 3):

$$r = \sum(X\text{-}E(X))*(Y\text{-}E(Y)) / S_X S_Y \text{ (Coladarci \& Cobb, 2014)}$$

The standard deviations decrease with an increase in sample size (see Chapter 3):

$$SD = \sqrt{(\sum(X\text{-}Xbar)^2/(n\text{-}1))} \text{ (Coladarci \& Cobb, 2014)}$$

The raw data method calculations produce the best combination of higher reliability and lower subscore correlation estimates using the LCDM data (see Appendices 4, 5, 8 and 9).

The differential method is not a post hoc method of computing subscores. As described in Chapter 2, this method adjusts scores linearly a priori. The methods of estimating reliability and correlation do not change from those of the raw method. Therefore, due to the linear score adjustment, the results do not differ. Refer to Appendices 4, 5, 8, and 9 for specific results. The additional pre-calculations, a linear change in scores, did not affect the reliability and correlation estimates. This differential method does not provide additional information about the specific subscores.

The Haberman methods.

The Haberman methods 2, 3, and 4 are designed to use additional data, post hoc, to improve subscore information. The models and variable definitions for each can be found in Chapter 2 and Chapter 3. Method 2 did not provide additional information above the raw or

differential methods. Method 2 uses the reliability of the subscore. Including the subscore reliability to the H2 method did not add additional information already found in the raw score or differential scoring methods. See Appendices 4, 5, 8, and 9 to compare reliability and correlation estimates.

Looking at the model for H 3, $S_x = E(S) + c[X - E(X)]$, the full model is defined in Chapter 3. The variable *c* is used as additional data to increase the information provided by this method. C contains the reliability and standard deviation of the total score. By indirectly increasing the number of items, the reliability of the subscore increases. This is done through the reliability of the total score and the standard deviation of the total score. This increase in reliability for each subscore does provide an improvement in the subscores generated by this method.

However, the correlation between all the subscore combinations, in both the Rasch and LCDM data simulation methods and all the covariance matrices used, is 1. This singular correlation value cancels the multidimensionality of the covariance matrices used to formulate the data. The H 3 method demonstrates all the simulated tests are unidimensional. Contrary to the increased subscore reliabilities, this individual correlation value provides no additional useful information about the specific subscores.

The H 4 method is the most complex of the CTT methods. H 4 uses both the subscores and the total score to attempt to add value to the original subscore output. Each of the variables in this model has been defined in Chapter 2 and Chapter 3.

$$S_{sx} = E(S) + \beta[S - E(S)] + \gamma[X - E(X)]$$

The combination of subscore and total score data provide H 4 with increased reliability and subscore correlations. H 4 provides the most improvement of subscore reporting for the five

CTT methods. For a complete comparison of reliability and correlation estimates for the five CTT methods, see Appendices 18 through 22.

The raw model produces reliability and correlation estimates free from additional a priori and post hoc calculations. This differential method and the H 2 method did not provide improvement over the raw score method. The H 3 method is an improvement in reliability, but the correlation estimate of 1 for all of the data sets removes any potential use on multidimensional data. The H 4 method provides subscore improvement. This post hoc method has more reliable subscores, and between subscore correlations are closer to the covariance matrices that produced both the multidimensional and unidimensional data.

LCDM.

The LCDM method of analyzing subscores looks at the various aspects of the test simultaneously. This method takes into account the individual log odds of successfully answering an item, the main effects for the items, the probability of an individual answering an item correctly while belonging to one of the attribute classes, and the probability that the individual belongs to a specific attribute class. LCDM does this using four separate models simultaneously. These models can be found in Chapter 2. The analysis output from this specific LCDM analysis is broken into six areas: 1) LCDM item parameters, 2) The attribute reliability, 3) EAP marginal estimates, 4) EAP profile estimates, 5) MAP marginal estimates, and 6) MAP profile estimates. For further explanation, see Appendix 24.

In all data sets, at all covariance levels, and in all sample sizes, the LCDM method outperformed each of the CTT methods in reliability estimates (see Appendices 5 and 9). One of the criteria for improving subscore information is demonstrating superior reliability scores. The LCDM method produced consistency in reporting suggested in both the Standards For

Educational and Psychological Testing and in the 2013 paper by Templin and Bradshaw (AERA/APA/NCME, 2014) (Templin & Bradshaw, 2013).

The LCDM correlation scores do not provide as consistent information as the reliability scores. The correlations are comparable to those found in the Haberman 4 method. Both sets of correlations are similar to the covariance used to simulate the Rasch model data and LCDM data.

This lack of consistency makes for interesting comparisons and potential decisions on which method to use. Note that in Tables 9 and 10, the reliabilities of the LCDM and H4 are higher than the raw model. But, if the stakeholder was using a multidimensional test, in this case with a covariance of .6, the raw method produces lower correlation estimates, which is a criteria for providing additional information via subscores (Feinberg & Wainer, 2014a). In this situation the question arises if high reliability guarantees validity or would moderate reliability scores paired with lower subscore correlation be worth consideration (Ling, 2009). For a complete comparison of reliability and correlation estimates for the LCDM method, see Appendix 23.

Table 9 Correlations for Rasch Data, Covariance .6, Attributes 1&3

	LCDM	Raw	Dif	H2	H3	H4
Sample	CorA1A3	CorA1A3	CorA1A3	CorA1A3	CorA1A3	CorA1A3
1200Cov.6	0.776	0.370	0.370	0.370	1.000	0.777
SD	0.054	0.025	0.025	0.025	0.000	0.032

Table 10 Reliability for Rasch Data, Covariance .6, Attributes 1 and 3

Attribute 1

	LCDM	Raw	Dif	H2	H3	H4
Sample	RelA3	RelA3	RelA3	RM2A3	RM3A3	RM4A3
300Cov.6	0.853	0.513	0.515	0.513	0.499	0.618
SD	0.036	0.048	0.047	0.048	0.066	0.038
1200Cov.6	0.835	0.524	0.527	0.524	0.502	0.623
SD	0.019	0.025	0.025	0.025	0.032	0.019

Attribute 3

Sample	LCDM RelA3	Raw RelA3	Dif RelA3	H2 RM2A3	H3 RM3A3	H4 RM4A3
300Cov.6	0.945	0.719	0.720	0.719	0.666	0.751
SD	0.017	0.030	0.029	0.030	0.038	0.026
1200Cov.6	0.931	0.715	0.718	0.715	0.663	0.748
SD	0.010	0.015	0.014	0.015	0.018	0.012

The research question asks if the subscore estimates provided by a LCDM deliver an improvement over CTT based subscore reporting. One of the criteria for this improvement is an increase in subscore reliability. The LCDM method does provide this. The question in the area of reliability is whether or not polychoric correlations can be rightfully compared to the Cronbach Alpha reliability estimates used by the CTT methods. While both are designed to show consistency of test replication, it is up to the individual psychometrician to decide if the separate methods can be legitimately compared.

This simulation shows that the LCDM method of subscore analysis provides similar correlations but higher reliabilities than any of the CTT methods. Of the six subscore analysis methods used in this simulation, Haberman method 3 provided the least useful information concerning the subscores. This was due to the repeated correlation of 1 with a SD of 0 for each of the data sets. Even though the reliabilities for this method were above those of the raw, differential, and H 2 methods, the consistent correlation of 1 suggests that all data sets were unidimensional when in fact they had varying covariance structures: 0.3, 0.6, 0.9, and 1 (see Appendices 4, 5, 8, and 9).

The CTT method that provided the most information was the H 4 method. The reliabilities were all greater than any of the other CTT methods, although not all were above Ling's (2009) suggestion of 0.85. The correlations for H 4 were larger than three of the CTT methods, except for Haberman 3, and were related to the covariance used to simulate the data.

The SD was associated to the type of data set simulated. LCDM showed lower SD with LCDM data than Rasch Model data, and Haberman showed lower SD with Rasch Model data than LCDM data.

The method that did provide reliability scores above 0.85, 0.890 and above, was the LCDM method. The correlations were similar to H 4, and these were tied closely to the covariance matrices used to produce the data. The SD for either method was not affected by the type of data being analyzed.

This simulation answers one aspect of the research question: the LCDM method produces higher subscore reliabilities than the five CTT methods. What it does not conclusively answer is whether or not the correlation estimates provide improved information over four of the five CTT methods.

Implications.

Looking closely at the three Haberman methods and the LCDM it is noticed that the control of variance is of primary concern. The CTT methods all use Cronbach's Alpha to calculate reliability. The denominator in this formula is the variance of X, the test or subtest. By decreasing the variance, the reliability increases. Increasing the number of items is a popular method to increase reliability (this is demonstrated in the third paragraph of this section). Haberman increases items first with subscores (H2), then with total scores (H3), and finally with both (H4). The more items used, the more items fall in the tails of the distribution, the less variance there is. The less variance there is the higher the reliability.

LCDM takes a different route to controlling variance. The LCDM method uses three steps (see chapters' 2 and 3) to convert continuous outcomes to a dichotomous outcome with a Bernoulli distribution. This distribution has no tails, consolidating the error, and increasing

reliability. LCDM produces subscore estimates with the least amount of variance and the highest amount of reliability.

Shorter tests are needed to obtain acceptable reliability estimates. A popular method to increase reliability is to add more items until the desired reliability is met. Looking at the Rasch data, 300cov.3, and using the Spearman-Brown Prophecy formula, the H 4 method would need approximately 21 items in subtest 1, 42 items in subtest 2, and 71 items in subtest 3 to ascertain the same reliability estimates as the original LCDM output.

LCDM analysis allows for the grain size of the attribute to vary with each administration of a test by adjusting the Q-matrix. If the results of this test show respondents struggling in attribute 2, a second test could be produced dividing attribute two into three attributes of smaller grain size. Looking deeper into the possible areas that need attention for each respondent. This ability to adjust attribute grain size is

At the beginning of the paper diagnosis was mentioned as a reason for giving a test. By producing scores that are more reliable than its CTT counterparts the LCDM method instills an increased level of confidence in the potential final diagnosis.

Combining the four previous paragraphs allows for a stakeholder to expand the breadth and depth of potential diagnostic as well as evaluative testing available to be given to the respondent. With high reliability estimates using a small number of items more can be tested, evaluated, and diagnosed in less time than currently afforded.

Limitations and positives to CTT.

A significant drawback to all Haberman methods is that the number of items within an attribute cannot change. This number must be set at the beginning of the testing process and therefore Haberman cannot be used in adaptive testing situations. H 1 was not reported in this simulation because the reported subscore for the participants is the mean of the subscores. H 2,

even though it was tied to subscore reliability, did not add additional information. H 3 produced correlations of 1 for all data sets. The reliabilities were stronger but this lack of variance with the correlation output made all data sets seem unidimensional, and no additional data could be obtained from the subscores. The H 4 method gave the most information of all the CTT methods. The reliability scores were the highest and the correlations were tied closely to the covariance matrix used in the data simulation.

All five CTT methods could be used in any testing situation in which the item number per attribute does not vary. In this simulation Attribute 1 always had 6 items, Attribute 2 always had 10 items, and Attribute 3 always had 14 items. Smaller scaled licensure tests, within school academic tests, and noncognitive tests could potentially benefit from the Haberman methods.

Limitations and positives to LCDM.

Some possible limitations of using LCDM are access to and familiarity with the software needed to complete the LCDM analysis. The convergence rates for the LCDM method are shown in Table 11. The convergence rates for the LCDM data are higher. The LCDM data was analyzed by an LCDM method. The convergence rates for the Rasch model generated data are lower.

The resulting use of fewer data sets is an area that should be examined. How does having a convergence rate of 98%, LCDM data 300Cov.9, affect the results compared to a convergence rate of 37% for Rasch data 300Cov.9?

Table 11 Convergence Rates for the LCDM Analysis

Convergence Rates LCDM Analysis		
Sample	**LCDM Data**	**Rasch Data**
300Cov1	59%	34%
300Cov.3	100%	90%
300Cov.6	100%	82%
300Cov.9	98%	37%
1200Cov1	57%	44%

1200Cov.3	100%	100%
1200Cov.6	100%	100%
1200Cov.9	100%	87%

To produce a subscore that provides additional information, the subscore reliability should be at or above 0.85 (Ling, 2009). The second aspect of producing additional information is the correlation between subscores. The higher the correlation between the subscores, the more unidimensional the test becomes, and the less value the subscores have at providing additional information (Feinberg & Wainer, 2014b).

LCDM does produce output that answers the original psychometric issue of having low reliability estimates. As is seen in the data, Attribute 1, with only 6 items, LCDM produces Rasch data reliabilities at or above 0.774 with SD at or below 0.045, and LCDM data reliabilities at or above 0.890 with SD at or below 0.081 (see Appendix 12).

Areas of Further Study

Replication of this simulation.

This study looks at introductory aspects of subscore reporting. One half of the data is produced with a 1-PL Rasch Model; the Q-matrix is a simple matrix; the covariance matrices are set equally; and the number of attributes is limited to three therefore limiting the number of possible attribute classes to eight. This study is a starting point for comparison between CTT subscore reporting methods and the LCDM Method. This would also be a starting point for expanding into more complex areas such as 2-PL IRT models and complex Q-Matrices.

Alternate reliability and correlation methods.

The LCDM method of subscore reporting is not able to use traditional reliability estimates from Cronbach's Alpha nor the correlation estimates of Pearson's r. In this study, tetrachoric correlations were used for reliability and polychoric correlations for subscore correlations. Alternatives to these include the Biserial and Point Biserial correlations for dichotomous items. The Spearman Rank-Order correlation coefficient is designed to measure the relationship between two ordinal variables, such as mastery and nonmastery (Corder & Foreman, 2014; Ferguson, 1959).

Templin and Bradshaw (2013) use the same reliability system for both a 1-PL IRT model and a DCM analysis (Templin & Bradshaw, 2013). An additional method, the Attribute Hierarchy Method (AHM), is used for classifying participant's responses into a set of structural attribute patterns (Gierl, 2009). The alternative methods used by LCDM to calculate reliability and correlation estimates need to be continually examined for validity in cross method comparisons.

LCDM compared to IRT subscore reporting methods.

Touched on briefly in Chapter 2, IRT provides a continuous measure of an examinee's ability. The additional information provided by IRT scoring is the rank order of the respondents. The standard error attribute of IRT is desirable in constructing tests to specific areas of the scale. This flexibility comes with a variability in the reliability across the scale.

LCDM provides additional analysis by placing examinees in diagnostic classes. In addition to this diagnostic analysis, the LCDM is able to provide an explanation as to why an examinee was placed within a specific class. The LCDM latent variables are categorized and therefore possible error is consolidated and reliability estimates increase (Henson et al., 2009). LCDM reliability has been shown to be uniformly higher than IRT models. This increase in reliability makes LCDM analysis an attractive alternative to IRT, but the reliability estimates are calculated from different models making reliability difficult to directly compare.

In mixed format testing, an LCDM alternative to the bi-factor IRT model.

The advances within computer-based testing have allowed for formats in addition to multiple choice within testing scenarios. Constructed response and essay answers are two popular formats. A bi-factor IRT model assumes an overall general construct with various individualized constructs influencing exclusive groups of items. The method shows the uniqueness of the group factors and the orthogonality of the subscores (Wang et al., 2016).

This type of testing directly influences the development and validity of the Q-matrix. How does a complex Q-matrix include all the possibilities without becoming overly burdensome or in worst case, invalid? The development of the Q-matrix is a subjective endeavor and is often looked at as a validity issue within the LCDM analysis process. Adding constructed response and essay answers to the "item to attribute" assignments may bring the Q-matrix more scrutiny.

Q-matrix validity.

The assignment of items to attributes for an academic based test can lead to a simple Q-matrix similar to the matrix used in this study. As mentioned previously, the complexity of the subject matter and of the test design, could lead to misspecifications within the Q-matrix.

A testing area not touched upon to this point is the noncognitive genre. The items and the attributes are at times not as straightforward to assign as in cognitive testing. It is this complexity of assignment that may call the validity of these particular Q-matrices into question.

The validity of the Q-matrix, cognitive or noncognitive based, affects the correct estimation of model parameters and ultimately the correct classification of the participant (de la Torre & Chiu, 2016). As with all validity concerns, Q-matrix validity is on a continuum, and continued verification of Q-matrix practices is essential.

References

AERA/APA/NCME. (2014). *Standards for educational and psychological testing*. Wahington D.C.: American Educational Research Association.

Andrich, D. (1988). *Rasch models for measurement* (Vol. 68). Newbury Park, CA: Sage.

Babenko, O., & Rogers, T. W. (2014). Comparison and properties of correlational and agreement methods for determining whether or not to report subtest scores. *International Journal of Learning, Teaching and Educational Research, 4*(1).

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. doi:10.1007/s11336-013-9350-4

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. doi:10.1111/j.1745-3984.2012.00185.x

Coladarci, T., & Cobb, C. (2014). *Fundamentals of Statistical Reasoning in Education* (4th ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Corder, G. W., & Foreman, D. I. (2014). *Nonparametric statistics a step-by-step approach* (2nd Edition ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Dai, S., Svetina, D., & Wang, X. (2017). Reporting subscores using R: A sofware review. *Journal of educational and behavioral statistics, 42*(5), 617-638.

de la Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied psychological measurement, 33*(3), 163-183. doi:10.1177/0146621608320523

de la Torre, J. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement, 47*(2), 227-249. doi:10.1111/j.1745-3984.2010.00110.x

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*(2), 253-273. doi:10.1007/s11336-015-9467-8

Ebel, R. L. (1967). The relation of item discrimination to test reliability1. *Journal of Educational Measurement, 4*(3), 125-128. doi:10.1111/j.1745-3984.1967.tb00579.x

Feinberg, R. A., & Jurich, D. (2017). Guidelines for interpreting and reporting subscores. *Educational measurement, issues and practice, 36*(1), 5-13.

Feinberg, R. A., & Wainer, H. (2014a). A simple equation to predict a subscore's value. *Educational Measurement: Issues and Practice, 33*(3), 55-56. doi:10.1111/emip.12035

Feinberg, R. A., & Wainer, H. (2014b). When can we improve subscores by making them shorter?: The case against subscores with overlapping items. *Educational measurement, issues and practice, 33*(3), 47-54. doi:10.1111/emip.12037

Ferguson, G. A. (1959). *Statistical analysis in psychology and education*. New York, New York: McGraw Hill.

Gierl, M. J. (2009). Reliability and attribute-based scoring in cognitive diagnostic assessment. *Journal of Educational Measurement, 46*(3), 293-313. doi:10.1111/j.1745-3984.2009.00082.x

Glen, S. (2018). Statistics How To. Retrieved from https://owl.english.purdue.edu/owl/resource/560/10/

Haberman, S., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology, 62*(1), 79-95. doi:10.1348/000711007X248875

Haberman, S. J. (2005). When can subscores have value? *ETS Research Report Series, 2005*(1), i-15. doi:10.1002/j.2333-8504.2005.tb01985.x

Haberman, S. J. (2008). When can subscores have value? *Journal of educational and behavioral statistics, 33*(2), 204-229. doi:10.3102/1076998607302636

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika, 75*(2), 209-227. doi:10.1007/s11336-010-9158-4

Haladyna, T. M. (2004). The validity of subscores for a credentialing test. *Evaluation & the health professions, 27*(4), 349-368. doi:10.1177/0163278704270010

He, Q. (2009). *Estimating the reliability of composite scores*. Retrieved from http://dera.ioe.ac.uk/id/eprint/1060

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

Huff, K. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education, 23*(4), 310-324. doi:10.1080/08957347.2010.510956

Kline, R. B. (2016). *Principles abd Practice of Structural Equation Modeling* (4th ed.). New York, New York: Guilford Publications.

Ling, G. (2009). *Why the major field (business) test does not report subscores of individual test-takers—reliability and construct validity evidence.* Paper presented at the Annual meeting of the National Council on Measurement in Education, San Diego, CA.

Liu, J. (2012). Data-driven learning of Q-matrix. *Applied psychological measurement, 36*(7), 548-564. doi:10.1177/0146621612456591

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. United States of America: Addison-Wesley Publishing Co.

Luecht, R. M. (2003). Applications of multidimensional diagnostic scoring for certification and licensure tests.

Madison, M. J., & Bradshaw, L. (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. doi:10.1177/0013164414539162

Maria, A. (1997). *Introduction to modeling and simulation*. Paper presented at the Proceedings of the 29th conference on Winter simulation, Atlanta, Georgia, USA.

Monaghan, W. (2006). The facts about subscores. *Educational Testing Services*.

More correlation coefficients. (2016). Retrieved from https://www.andrews.edu/~calkins/math/edrm611/edrm13.htm#TETRA

Puhan, G., Sinharay, S., Haberman, S., & Larkin, K. (2010). The utility of augmented subscores in a licensure exam: an evaluation of methods using empirical data. *Applied Measurement in Education, 23*(3), 266-285. doi:10.1080/08957347.2010.486287

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to Psychometric Theory*. New York: Routledge.

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

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement : journal of the International Measurement Confederation, 6*(4), 219-262. doi:10.1080/15366360802490866

Sha, S., & McCoy, T. (2014). A comparison of two augmented subscore methods and the role of score distribution.

Sinharay, S., Haberman, S., & Boughton, K. (2015). Too simple to be useful: A comment on Feinberg and Wainer (2014). *Educational Measurement: Issues and Practice, 34*(3), 6-8.

Sinharay, S., & Haberman, S. J. (2008). Reporting subscores: A survey. *Research Memorandum*(08-18).

Sinharay, S., & Haberman, S. J. (2015). Comments on “A Note on Subscores” by Samuel A. Livingston. *Educational Measurement: Issues and Practice, 34*(2), 6-7. doi:10.1111/emip.12071

Sinharay, S., Haberman, S. J., & Wainer, H. (2011). Do adjusted subscores lack validity? Don’t blame the messenger. *Educational and Psychological Measurement, 71*(5), 789-797. doi:10.1177/0013164410391782

Sinharay, S., Puhan, G., & Haberman, S. J. (2010). Reporting diagnostic scores in educational testing: Temptations, pitfalls, and some solutions. *Multivariate behavioral research, 45*(3), 553-573. doi:10.1080/00273171.2010.483382

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice, 30*(3), 29-40. doi:10.1111/j.1745-3992.2011.00208.x

Skorupski, W. P. (2017). *EPSY 922 Item Response Theory.* Paper presented at the The University of Kansas, Lawrence, KS.

Skorupski, W. P., & Carvajal, J. (2010). A comparison of approaches for improving the reliability of objective level scores. *Educational and Psychological Measurement, 70*(3), 357-375. doi:10.1177/0013164409355694

Stanley, J. C., & Wang, M. D. (1968). *Differential weighting: A survey of methods and empirical studies*. New York, NY: College Entrance Examination Board.

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. doi:10.1111/j.1745-3984.1983.tb00212.x

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

Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational measurement, issues and practice, 32*(2), 37-50. doi:10.1111/emip.12010

Uebersax, J. (2015). The tetrachoric and polychoric correlation coefficients. Statistical methods for rater agreement. Retrieved from http://www.john-uebersax.com/stat/tetra.htm

von Davier, M. (2014). The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM). Research report. ETS RR-14-40. *ETS Research Report Series*.

Wainer, H., Vevea, J., Camacho, F., Reeve III, B., Rosa, K., Nelson, L., . . . Thissen, D. (2001). Augmented scores - "Borrowing Strength" to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test Scoring* (pp. 343-387). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Wang, W., Drasgow, F., & Liu, L. (2016). Classification accuracy of mixed format tests: A bi-factor approach. *Frontiers in Psychology, 7*. doi:10.3389/fpsyg.2016.00270

Appendices

Appendix 1 Subscore Sources of Validity

Sources of Validity Evidence for Subscore Interpretation			
Source of Evidence	***Type of Validity Evidence***	***Description of This Evidence for a Unidimensional Approach***	***Description of This Evidence for a Multidimensional Approach***
Construct definition	Logical	A single definition is sufficient. Sub domains are irrelevant.	A single definition is insufficient. Subdomains are considered real and important.
Practice, job, or task analysis	Procedural	A practice analysis makes no differentiation among subdomains in terms of critically.	A practice analysis reveals differences among subdomains.
Test specifications	Procedural	This action emphasizes the importance of content distribution and representation in the test.	Test specifications call for test design to be undertaken with specific content area quotas.
Item development	Procedural	All items are fungible.	Each item has a unique content identification and elicits a cognitive demand.
Test design	Procedural	All items in the pool are exchangeable. IT matters little which items are chosen.	Each item is chosen to satisfy the test specifications in items of content and cognitive operation.
Subscore differences	Empirical	All possible subscores should be equally difficult.	All possible subscores might be equally difficult but can be unequal.
Intercorrelations among subscores	Empirical	Correlations, when corrected for attenuation, should be approaching or near unity to support unidimensionality.	After corrections for attenuation, correlations among subscores should be less than unity.
Factor analysis	Empirical	A confirmatory factor analysis should show one and only one factor and no evidence for hypothesized subscores.	A confirmatory factor analysis should generate some evidence for subscores.
Item analysis	Empirical	Item discriminations using total score or appropriate subscore should provide identical results to support unidimensionality.	If the item analysis results differ as a function of the criterion score, then subscore validity is supported.
Subscore reports	Empirical	Subscores should be different enough for each candidate for a subscore report to be informative.	If the item analysis of differences among subscores is statistically large and potentially meaningful to candidates, then this evidence supports multidimensionality.

Appendix 2 Q-Matrix

Question/Attribute	Attribute 1	Attribute 2	Attribute 3
Q1	1		
Q2	1		
Q3	1		
Q4	1		
Q5	1		
Q6	1		
Q7		1	
Q8		1	
Q9		1	
Q10		1	
Q11		1	
Q12		1	
Q13		1	
Q14		1	
Q15		1	
Q16		1	
Q17			1
Q18			1
Q19			1
Q20			1
Q21			1
Q22			1
Q23			1
Q24			1
Q25			1
Q26			1
Q26			1
Q27			1
Q28			1
Q29			1
Q30			1

Appendix 3 Data Specification Tables

Table 12 Rasch Model Covariance Structure

Test	Covariance A1	Covariance A2	Covariance A3
Simulated Data Sets: Variance/Covariance			
A1 Unidimensional	1	1	1
A2 Multidimensional	.3	.3	.3
A3 Multidimensional	.6	.6	.6
A4 Multidimensional	.9	.9	.9

Table 13 LCDM Covariance Structure

Test	Covariance A1	Covariance A2	Covariance A3
Simulated Data Sets: LCDM			
B1 Unidimensional	1	1	1
B2 Multidimensional	.3	.3	.3
B3 Multidimensional	.6	.6	.6
B4 Multidimensional	.9	.9	.9

Table 14 Simulated data specifications Variance/Covariance Matrix

Test	Replications	Sample Size	Total Test	Subtests	Total Test Reliability	Subtest Reliability
Simulated Data Sets: Variance/Covariance						
A1, Cov 1 Unidimensional	100	300	30	6, 10, 14	≈0.844	≈0.510, 0.640, 0.719
	100	1200	30	6, 10, 14	≈0.844	≈0.515, 0.642, 0.719
A2, Cov 0.3 Multidimensional	100	300, 1200	30	6, 10, 14	≈0.728	≈0.516, 0.640, 0.713
	100	1200	30	6, 10, 14	≈0.732	≈0.518, 0.642, 0.718
A3, Cov 0.6 Multidimensional	100	300	30	6, 10, 14	≈0.792	≈0.513, 0.637, 0.719
	100	1200	30	6, 10, 14	≈0.793	≈0.524, 0.643, 0.715
A4, Cov 0.9 Multidimensional	100	300	30	6, 10, 14	≈0.833	≈0.515, 0.644, 0.717
	100	1200	30	6, 10, 14	≈0.833	≈0.514, 0.644, 0.716

Table 15 Simulated data specifications LCDM

Test	Replications	Sample Size	Total Test	Subtests	Total Test Reliability	Subtest Reliability
Simulated Data Sets: Log-Linear Cognitive Diagnostic Model						
B1, Cov 1 Unidimensional	100	300	30	6, 10, 14	≈0.857	≈0.537, 0.663, 0.719
	100	1200	30	6, 10, 14	≈0.853	≈0.525, 0.645, 0.723
B2, Cov 0.3 Multidimensional	100	300	30	6, 10, 14	≈0.719	≈0.539, 0.662, 0.719
	100	1200	30	6, 10, 14	≈0.713	≈0.524, 0.645, 0.722
B3, Cov 0.6 Multidimensional	100	300	30	6, 10, 14	≈0.774	≈0.538, 0.662, 0.719
	100	1200	30	6, 10, 14	≈0.768	≈0.526, 0.645, 0.723
B4, Cov 0.9 Multidimensional	100	300	30	6, 10, 14	≈0.824	≈0.539, 0.662, 0.720
	100	1200	30	6, 10, 14	≈0.820	≈0.525., 0.645, 0.723

Appendix 4 Rasch Data Reliability by Method

Rasch Data Reliability by Method

	LCDM			Rasch Raw				Differential			
	RelA1	**RelA2**	**RelA3**	**RelA1**	**RelA2**	**RelA3**	**RelTot**	**RelA1**	**RelA2**	**RelA3**	**RelTot**
300Cov1	0.943	0.962	0.969	0.510	0.640	0.719	0.844	0.513	0.643	0.721	0.846
SD	0.023	0.012	0.010	0.046	0.035	0.023	0.013	0.045	0.035	0.024	0.013
300Cov.3	0.810	0.889	0.929	0.516	0.640	0.713	0.728	0.517	0.642	0.715	0.734
SD	0.045	0.025	0.019	0.042	0.032	0.027	0.023	0.045	0.032	0.026	0.022
300Cov.6	0.853	0.908	0.945	0.513	0.637	0.719	0.792	0.515	0.640	0.720	0.795
SD	0.036	0.022	0.017	0.048	0.035	0.030	0.020	0.047	0.034	0.029	0.019
300Cov.9	0.926	0.947	0.960	0.515	0.644	0.717	0.833	0.516	0.646	0.718	0.835
SD	0.028	0.017	0.012	0.044	0.037	0.027	0.016	0.045	0.037	0.027	0.016
1200Cov1	0.943	0.956	0.966	0.515	0.642	0.717	0.844	0.519	0.644	0.719	0.845
SD	0.012	0.007	0.006	0.027	0.021	0.016	0.008	0.025	0.022	0.016	0.008
1200Cov.3	0.774	0.870	0.920	0.518	0.642	0.718	0.732	0.520	0.645	0.721	0.738
SD	0.027	0.018	0.013	0.028	0.019	0.017	0.012	0.027	0.019	0.017	0.012
1200Cov.6	0.835	0.894	0.931	0.524	0.643	0.715	0.793	0.527	0.645	0.718	0.796
SD	0.019	0.015	0.010	0.025	0.022	0.015	0.010	0.025	0.022	0.014	0.010
1200Cov.9	0.910	0.937	0.953	0.514	0.644	0.716	0.833	0.517	0.646	0.719	0.835
SD	0.017	0.010	0.008	0.027	0.020	0.016	0.007	0.026	0.020	0.016	0.008

	Haberman 2			Haberman 3			Haberman 4			Haberman
	RM2A1	**RM2A2**	**RM2A3**	**RM3A1**	**RM3A2**	**RM3A3**	**RM4A1**	**RM4A2**	**RM4A3**	**RTot**
300Cov1	0.510	0.640	0.719	0.862	0.848	0.848	0.869	0.852	0.849	0.844
SD	0.046	0.035	0.023	0.061	0.030	0.018	0.070	0.032	0.019	0.013
300Cov.3	0.516	0.640	0.713	0.263	0.384	0.527	0.543	0.654	0.721	0.728
SD	0.042	0.032	0.027	0.061	0.054	0.045	0.039	0.030	0.026	0.023
300Cov.6	0.513	0.637	0.719	0.499	0.575	0.666	0.618	0.693	0.751	0.792
SD	0.048	0.035	0.030	0.066	0.047	0.038	0.038	0.029	0.026	0.020
300Cov.9	0.515	0.644	0.717	0.767	0.778	0.800	0.780	0.792	0.809	0.833
SD	0.044	0.037	0.027	0.059	0.036	0.025	0.050	0.027	0.021	0.016
1200Cov1	0.515	0.642	0.717	0.852	0.845	0.846	0.853	0.846	0.846	0.844
SD	0.027	0.021	0.016	0.028	0.015	0.011	0.030	0.015	0.011	0.008
1200Cov.3	0.518	0.642	0.718	0.261	0.381	0.532	0.543	0.655	0.726	0.732
SD	0.028	0.019	0.017	0.030	0.032	0.027	0.025	0.018	0.016	0.012
1200Cov.6	0.524	0.643	0.715	0.502	0.578	0.663	0.623	0.696	0.748	0.793
SD	0.025	0.022	0.015	0.032	0.024	0.018	0.019	0.018	0.012	0.010
1200Cov.9	0.514	0.644	0.716	0.756	0.778	0.801	0.767	0.789	0.808	0.833
SD	0.027	0.020	0.016	0.030	0.018	0.014	0.024	0.014	0.011	0.007

Appendix 5 Rasch Data Correlation by Method

Rasch Data Correlation by Method

	LCDM			Raw			Differential		
	CorA1A2	CorA1A3	CorA2A3	CorA1A2	CorA1A3	CorA2A3	CorA1A2	CorA1A3	CorA2A3
300Cov1	0.981	0.980	0.986	0.578	0.615	0.682	0.578	0.615	0.682
SD	0.017	0.021	0.016	0.043	0.036	0.032	0.043	0.036	0.032
300Cov.3	0.418	0.439	0.409	0.171	0.183	0.204	0.171	0.183	0.204
SD	0.165	0.162	0.173	0.058	0.056	0.054	0.058	0.056	0.054
300Cov.6	0.775	0.764	0.761	0.344	0.366	0.407	0.344	0.366	0.407
SD	0.099	0.116	0.087	0.045	0.057	0.048	0.045	0.057	0.048
300Cov.9	0.968	0.957	0.959	0.523	0.553	0.611	0.523	0.553	0.611
SD	0.028	0.035	0.027	0.041	0.041	0.039	0.041	0.041	0.039
1200Cov1	0.993	0.993	0.991	0.578	0.613	0.680	0.578	0.613	0.680
SD	0.005	0.006	0.006	0.023	0.021	0.018	0.023	0.021	0.018
1200Cov.3	0.431	0.445	0.398	0.172	0.186	0.203	0.172	0.186	0.203
SD	0.085	0.088	0.071	0.027	0.029	0.028	0.027	0.029	0.028
1200Cov.6	0.783	0.776	0.743	0.350	0.370	0.407	0.350	0.370	0.407
SD	0.061	0.054	0.053	0.025	0.025	0.024	0.025	0.025	0.024
1200Cov.9	0.968	0.966	0.960	0.518	0.548	0.613	0.518	0.548	0.613
SD	0.023	0.016	0.019	0.022	0.023	0.020	0.022	0.023	0.020

	Haberman 2			Haberman 3			Haberman 4		
	CM2A1A2	CM2A1A3	CM2A2A3	CM3A1A2	CM3A1A3	CM3A2A3	CM4A1A2	CM4A1A3	CM4A2A3
300Cov1	0.578	0.615	0.682	1.000	1.000	1.000	0.994	0.995	0.997
SD	0.043	0.036	0.032	0.000	0.000	0.000	0.007	0.006	0.004
300Cov.3	0.171	0.183	0.204	1.000	1.000	1.000	0.399	0.423	0.410
SD	0.058	0.056	0.054	0.000	0.000	0.000	0.095	0.101	0.091
300Cov.6	0.344	0.366	0.407	1.000	1.000	1.000	0.765	0.776	0.759
SD	0.045	0.057	0.048	0.000	0.000	0.000	0.059	0.069	0.062
300Cov.9	0.523	0.553	0.611	1.000	1.000	1.000	0.978	0.981	0.976
SD	0.041	0.041	0.039	0.000	0.000	0.000	0.018	0.016	0.019
1200Cov1	0.578	0.613	0.680	1.000	1.000	1.000	0.999	0.999	0.999
SD	0.023	0.021	0.018	0.000	0.000	0.000	0.001	0.002	0.001
1200Cov.3	0.172	0.186	0.203	1.000	1.000	1.000	0.399	0.425	0.407
SD	0.027	0.029	0.028	0.000	0.000	0.000	0.045	0.048	0.045
1200Cov.6	0.350	0.370	0.407	1.000	1.000	1.000	0.764	0.777	0.759
SD	0.025	0.025	0.024	0.000	0.000	0.000	0.031	0.032	0.029
1200Cov.9	0.518	0.548	0.613	1.000	1.000	1.000	0.981	0.983	0.980
SD	0.022	0.023	0.020	0.000	0.000	0.000	0.008	0.009	0.009

Appendix 6 Rasch Data Reliability by Attribute

Rasch Data Reliability by Attribute

	LCDM RelA1	Raw RelA1	Dif RelA1	H2 RM2A1	H3 RM3A1	H4 RM4A1
300Cov1	0.943	0.510	0.513	0.510	0.862	0.869
SD	0.023	0.046	0.045	0.046	0.061	0.070
300Cov.3	0.810	0.516	0.517	0.516	0.263	0.543
SD	0.045	0.042	0.045	0.042	0.061	0.039
300Cov.6	0.853	0.513	0.515	0.513	0.499	0.618
SD	0.036	0.048	0.047	0.048	0.066	0.038
300Cov.9	0.926	0.515	0.516	0.515	0.767	0.780
SD	0.028	0.044	0.045	0.044	0.059	0.050
1200Cov1	0.943	0.515	0.519	0.515	0.852	0.853
SD	0.012	0.027	0.025	0.027	0.028	0.030
1200Cov.3	0.774	0.518	0.520	0.518	0.261	0.543
SD	0.027	0.028	0.027	0.028	0.030	0.025
1200Cov.6	0.835	0.524	0.527	0.524	0.502	0.623
SD	0.019	0.025	0.025	0.025	0.032	0.019
1200Cov.9	0.910	0.514	0.517	0.514	0.756	0.767
SD	0.017	0.027	0.026	0.027	0.030	0.024

	LCDM RelA2	Raw RelA2	Dif RelA2	H2 RM2A2	H3 RM3A2	H4 RM4A2
300Cov1	0.962	0.640	0.643	0.640	0.848	0.852
SD	0.012	0.035	0.035	0.035	0.030	0.032
300Cov.3	0.889	0.640	0.642	0.640	0.384	0.654
SD	0.025	0.032	0.032	0.032	0.054	0.030
300Cov.6	0.908	0.637	0.640	0.637	0.575	0.693
SD	0.022	0.035	0.034	0.035	0.047	0.029
300Cov.9	0.947	0.644	0.646	0.644	0.778	0.792
SD	0.017	0.037	0.037	0.037	0.036	0.027
1200Cov1	0.956	0.642	0.644	0.642	0.845	0.846
SD	0.007	0.021	0.022	0.021	0.015	0.015
1200Cov.3	0.870	0.642	0.645	0.642	0.381	0.655
SD	0.018	0.019	0.019	0.019	0.032	0.018
1200Cov.6	0.894	0.643	0.645	0.643	0.578	0.696
SD	0.015	0.022	0.022	0.022	0.024	0.018
1200Cov.9	0.937	0.644	0.646	0.644	0.778	0.789
SD	0.010	0.020	0.020	0.020	0.018	0.014

	LCDM RelA3	Raw RelA3	Dif RelA3	H2 RM2A3	H3 RM3A3	H4 RM4A3
300Cov1	0.969	0.719	0.721	0.719	0.848	0.849
SD	0.010	0.023	0.024	0.023	0.018	0.019
300Cov.3	0.929	0.713	0.715	0.713	0.527	0.721
SD	0.019	0.027	0.026	0.027	0.045	0.026
300Cov.6	0.945	0.719	0.720	0.719	0.666	0.751
SD	0.017	0.030	0.029	0.030	0.038	0.026
300Cov.9	0.960	0.717	0.718	0.717	0.800	0.809
SD	0.012	0.027	0.027	0.027	0.025	0.021
1200Cov1	0.966	0.717	0.719	0.717	0.846	0.846
SD	0.006	0.016	0.016	0.016	0.011	0.011
1200Cov.3	0.920	0.718	0.721	0.718	0.532	0.726
SD	0.013	0.017	0.017	0.017	0.027	0.016
1200Cov.6	0.931	0.715	0.718	0.715	0.663	0.748
SD	0.010	0.015	0.014	0.015	0.018	0.012
1200Cov.9	0.953	0.716	0.719	0.716	0.801	0.808
SD	0.008	0.016	0.016	0.016	0.014	0.011

Total	Raw RelTot	Dif RelTot	Hab RTot
300Cov1	0.844	0.846	0.844
SD	0.013	0.013	0.013
300Cov.3	0.728	0.734	0.728
SD	0.023	0.022	0.023
300Cov.6	0.792	0.795	0.792
SD	0.020	0.019	0.020
300Cov.9	0.833	0.835	0.833
SD	0.016	0.016	0.016
1200Cov1	0.844	0.845	0.844
SD	0.008	0.008	0.008
1200Cov.3	0.732	0.738	0.732
SD	0.012	0.012	0.012
1200Cov.6	0.793	0.796	0.793
SD	0.010	0.010	0.010
1200Cov.9	0.833	0.835	0.833
SD	0.007	0.008	0.007

Appendix 7 Rasch Data Correlation by Attribute

Rasch Data Correlation by Attribute

	LCDM CorA1A2	Raw CorA1A2	Dif CorA1A2	H2 CM2A12	H3 CM3A12	H4 CM4A12
300Cov1	0.981	0.578	0.578	0.578	1.000	0.994
SD	0.017	0.043	0.043	0.043	0.000	0.007
300Cov.3	0.418	0.171	0.171	0.171	1.000	0.399
SD	0.165	0.058	0.058	0.058	0.000	0.095
300Cov.6	0.775	0.344	0.344	0.344	1.000	0.765
SD	0.099	0.045	0.045	0.045	0.000	0.059
300Cov.9	0.968	0.523	0.523	0.523	1.000	0.978
SD	0.028	0.041	0.041	0.041	0.000	0.018
1200Cov1	0.993	0.578	0.578	0.578	1.000	0.999
SD	0.005	0.023	0.023	0.023	0.000	0.001
1200Cov.3	0.431	0.172	0.172	0.172	1.000	0.399
SD	0.085	0.027	0.027	0.027	0.000	0.045
1200Cov.6	0.783	0.350	0.350	0.350	1.000	0.764
SD	0.061	0.025	0.025	0.025	0.000	0.031
1200Cov.9	0.968	0.518	0.518	0.518	1.000	0.981
SD	0.023	0.022	0.022	0.022	0.000	0.008

	LCDM CorA1A3	Raw CorA1A3	Dif CorA1A3	H2 CM2A13	H3 CM3A13	H4 CM4A13
300Cov1	0.980	0.615	0.615	0.615	1.000	0.995
SD	0.021	0.036	0.036	0.036	0.000	0.006
300Cov.3	0.439	0.183	0.183	0.183	1.000	0.423
SD	0.162	0.056	0.056	0.056	0.000	0.101
300Cov.6	0.764	0.366	0.366	0.366	1.000	0.776
SD	0.116	0.057	0.057	0.057	0.000	0.069
300Cov.9	0.957	0.553	0.553	0.553	1.000	0.981
SD	0.035	0.041	0.041	0.041	0.000	0.016
1200Cov1	0.993	0.613	0.613	0.613	1.000	0.999
SD	0.006	0.021	0.021	0.021	0.000	0.002
1200Cov.3	0.445	0.186	0.186	0.186	1.000	0.425
SD	0.088	0.029	0.029	0.029	0.000	0.048
1200Cov.6	0.776	0.370	0.370	0.370	1.000	0.777
SD	0.054	0.025	0.025	0.025	0.000	0.032
1200Cov.9	0.966	0.548	0.548	0.548	1.000	0.983
SD	0.016	0.023	0.023	0.023	0.000	0.009

	LCDM CorA2A3	Raw CorA2A3	Dif CorA2A3	H2 CM2A23	H3 CM3A23	H4 CM4A23
300Cov1	0.986	0.682	0.682	0.682	1.000	0.997
SD	0.016	0.032	0.032	0.032	0.000	0.004
300Cov.3	0.409	0.204	0.204	0.204	1.000	0.410
SD	0.173	0.054	0.054	0.054	0.000	0.091
300Cov.6	0.761	0.407	0.407	0.407	1.000	0.759
SD	0.087	0.048	0.048	0.048	0.000	0.062
300Cov.9	0.959	0.611	0.611	0.611	1.000	0.976
SD	0.027	0.039	0.039	0.039	0.000	0.019
1200Cov1	0.991	0.680	0.680	0.680	1.000	0.999
SD	0.006	0.018	0.018	0.018	0.000	0.001
1200Cov.3	0.398	0.203	0.203	0.203	1.000	0.407
SD	0.071	0.028	0.028	0.028	0.000	0.045
1200Cov.6	0.743	0.407	0.407	0.407	1.000	0.759
SD	0.053	0.024	0.024	0.024	0.000	0.029
1200Cov.9	0.960	0.613	0.613	0.613	1.000	0.980
SD	0.019	0.020	0.020	0.020	0.000	0.009

Appendix 8 LCDM Data Reliability be Method

LCDM Data Reliability by Method

	LCDM			LCDM Raw				Differential			
	RelA1	RelA2	RelA3	RelA1	RelA2	RelA3	RelTot	RelA1	RelA2	RelA3	RelTot
300Cov1	0.999	1.000	1.000	0.537	0.663	0.719	0.857	0.547	0.672	0.728	0.863
SD	0.004	0.001	0.000	0.139	0.088	0.079	0.030	0.136	0.086	0.075	0.028
300Cov.3	0.911	0.973	0.988	0.539	0.662	0.719	0.719	0.549	0.672	0.728	0.731
SD	0.062	0.025	0.014	0.139	0.089	0.078	0.049	0.135	0.087	0.074	0.046
300Cov.6	0.930	0.979	0.991	0.538	0.662	0.719	0.774	0.548	0.672	0.729	0.783
SD	0.049	0.020	0.010	0.138	0.090	0.078	0.042	0.135	0.088	0.073	0.040
300Cov.9	0.968	0.990	0.995	0.539	0.662	0.720	0.824	0.549	0.671	0.730	0.831
SD	0.024	0.009	0.006	0.137	0.090	0.078	0.035	0.134	0.088	0.074	0.033
1200Cov1	0.999	1.000	1.000	0.525	0.645	0.723	0.853	0.535	0.657	0.733	0.859
SD	0.003	0.000	0.000	0.130	0.093	0.063	0.027	0.127	0.090	0.060	0.025
1200Cov.3	0.890	0.964	0.989	0.524	0.645	0.722	0.713	0.535	0.657	0.733	0.726
SD	0.081	0.030	0.012	0.129	0.093	0.063	0.039	0.126	0.090	0.059	0.037
1200Cov.6	0.915	0.972	0.991	0.526	0.645	0.723	0.768	0.536	0.657	0.733	0.778
SD	0.060	0.024	0.009	0.129	0.093	0.063	0.035	0.126	0.090	0.059	0.033
1200Cov.9	0.964	0.986	0.995	0.525	0.645	0.723	0.820	0.536	0.656	0.733	0.828
SD	0.025	0.011	0.005	0.130	0.094	0.063	0.030	0.127	0.090	0.059	0.029

	Haberman 2			Haberman 3			Haberman 4			Haberman
	RM2A1	RM2A2	RM2A3	RM3A1	RM3A2	RM3A3	RM4A1	RM4A2	RM4A3	RTot
300Cov1	0.537	0.663	0.719	0.926	0.880	0.867	0.950	0.886	0.871	0.857
SD	0.139	0.088	0.079	0.134	0.037	0.033	0.193	0.041	0.034	0.030
300Cov.3	0.539	0.662	0.719	0.219	0.335	0.484	0.557	0.669	0.723	0.719
SD	0.139	0.089	0.078	0.067	0.070	0.085	0.122	0.083	0.076	0.049
300Cov.6	0.538	0.662	0.719	0.388	0.476	0.589	0.600	0.689	0.737	0.774
SD	0.138	0.090	0.078	0.074	0.063	0.069	0.100	0.075	0.069	0.042
300Cov.9	0.539	0.662	0.720	0.646	0.681	0.729	0.715	0.747	0.776	0.824
SD	0.137	0.090	0.078	0.071	0.054	0.052	0.060	0.054	0.053	0.035
1200Cov1	0.525	0.645	0.723	0.895	0.871	0.863	0.905	0.876	0.867	0.853
SD	0.130	0.093	0.063	0.061	0.030	0.026	0.074	0.033	0.026	0.027
1200Cov.3	0.524	0.645	0.722	0.196	0.320	0.489	0.537	0.652	0.726	0.713
SD	0.129	0.093	0.063	0.051	0.066	0.062	0.124	0.090	0.061	0.039
1200Cov.6	0.526	0.645	0.723	0.362	0.460	0.588	0.578	0.672	0.738	0.768
SD	0.129	0.093	0.063	0.050	0.057	0.051	0.103	0.080	0.056	0.035
1200Cov.9	0.525	0.645	0.723	0.629	0.671	0.730	0.692	0.734	0.776	0.820
SD	0.130	0.094	0.063	0.048	0.037	0.034	0.059	0.052	0.041	0.030

Appendix 9 LCDM Data Correlation by Method

LCDM Data Correlation by Method

	LCDM			Raw			Differential		
	CorA1A2	CorA1A3	CorA2A3	CorA1A2	CorA1A3	CorA2A3	CorA1A2	CorA1A3	CorA2A3
300Cov1	1.000	1.000	1.000	0.625	0.642	0.708	0.625	0.642	0.708
SD	0.000	0.000	0.000	0.085	0.081	0.065	0.085	0.081	0.065
300Cov.3	0.327	0.345	0.308	0.128	0.135	0.135	0.128	0.135	0.135
SD	0.134	0.128	0.111	0.056	0.060	0.060	0.056	0.060	0.060
300Cov.6	0.630	0.645	0.618	0.261	0.274	0.291	0.261	0.274	0.291
SD	0.105	0.095	0.079	0.060	0.068	0.055	0.060	0.068	0.055
300Cov.9	0.918	0.917	0.909	0.448	0.458	0.505	0.448	0.458	0.505
SD	0.036	0.039	0.036	0.068	0.071	0.062	0.068	0.071	0.062
1200Cov1	1.000	1.000	1.000	0.600	0.634	0.698	0.600	0.634	0.698
SD	0.000	0.000	0.000	0.083	0.077	0.056	0.083	0.077	0.056
1200Cov.3	0.316	0.312	0.305	0.113	0.123	0.134	0.113	0.123	0.134
SD	0.059	0.059	0.054	0.031	0.032	0.031	0.031	0.032	0.031
1200Cov.6	0.632	0.629	0.611	0.245	0.260	0.283	0.245	0.260	0.283
SD	0.054	0.049	0.042	0.041	0.044	0.035	0.041	0.044	0.035
1200Cov.9	0.922	0.921	0.911	0.430	0.453	0.498	0.430	0.453	0.498
SD	0.022	0.022	0.016	0.059	0.060	0.043	0.059	0.060	0.043

	Haberman 2			Haberman 3			Haberman 4		
	CM2A1A2	CM2A1A3	CM2A2A3	CM3A1A2	CM3A1A3	CM3A2A3	CM4A1A2	CM4A1A3	CM4A2A3
300Cov1	0.625	0.642	0.708	1.000	1.000	1.000	0.985	0.987	0.993
SD	0.085	0.081	0.065	0.000	0.000	0.000	0.016	0.018	0.008
300Cov.3	0.128	0.135	0.135	1.000	1.000	1.000	0.289	0.313	0.276
SD	0.056	0.060	0.060	0.000	0.000	0.000	0.110	0.126	0.112
300Cov.6	0.261	0.274	0.291	1.000	1.000	1.000	0.577	0.602	0.563
SD	0.060	0.068	0.055	0.000	0.000	0.000	0.098	0.104	0.085
300Cov.9	0.448	0.458	0.505	1.000	1.000	1.000	0.884	0.889	0.875
SD	0.068	0.071	0.062	0.000	0.000	0.000	0.054	0.056	0.052
1200Cov1	0.578	0.613	0.680	1.000	1.000	1.000	0.999	0.999	0.999
SD	0.023	0.021	0.018	0.000	0.000	0.000	0.001	0.002	0.001
1200Cov.3	0.172	0.186	0.203	1.000	1.000	1.000	0.399	0.425	0.407
SD	0.027	0.029	0.028	0.000	0.000	0.000	0.045	0.048	0.045
1200Cov.6	0.350	0.370	0.407	1.000	1.000	1.000	0.764	0.777	0.759
SD	0.025	0.025	0.024	0.000	0.000	0.000	0.031	0.032	0.029
1200Cov.9	0.518	0.548	0.613	1.000	1.000	1.000	0.981	0.983	0.980
SD	0.022	0.023	0.020	0.000	0.000	0.000	0.008	0.009	0.009

Appendix 10 LCDM Data Reliability by Attribute

LCDM Data Reliability by Attribute

	LCDM RelA1	Raw RelA1	Dif RelA1	H2 RM2A1	H3 RM3A1	H4 RM4A1
300Cov1	0.999	0.537	0.547	0.537	0.926	0.950
SD	0.004	0.139	0.136	0.139	0.134	0.193
300Cov.3	0.911	0.539	0.549	0.539	0.219	0.557
SD	0.062	0.139	0.135	0.139	0.067	0.122
300Cov.6	0.930	0.538	0.548	0.538	0.388	0.600
SD	0.049	0.138	0.135	0.138	0.074	0.100
300Cov.9	0.968	0.539	0.549	0.539	0.646	0.715
SD	0.024	0.137	0.134	0.137	0.071	0.060
1200Cov1	0.999	0.525	0.535	0.525	0.895	0.905
SD	0.003	0.130	0.127	0.130	0.061	0.074
1200Cov.3	0.890	0.524	0.535	0.524	0.196	0.537
SD	0.081	0.129	0.126	0.129	0.051	0.124
1200Cov.6	0.915	0.526	0.536	0.526	0.362	0.578
SD	0.060	0.129	0.126	0.129	0.050	0.103
1200Cov.9	0.964	0.525	0.536	0.525	0.629	0.692
SD	0.025	0.130	0.127	0.130	0.048	0.059

	LCDM RelA2	Raw RelA2	Dif RelA2	H2 RM2A2	H3 RM3A2	H4 RM4A2
300Cov1	1.000	0.663	0.672	0.663	0.880	0.886
SD	0.001	0.088	0.086	0.088	0.037	0.041
300Cov.3	0.973	0.662	0.672	0.662	0.335	0.669
SD	0.025	0.089	0.087	0.089	0.070	0.083
300Cov.6	0.979	0.662	0.672	0.662	0.476	0.689
SD	0.020	0.090	0.088	0.090	0.063	0.075
300Cov.9	0.990	0.662	0.671	0.662	0.681	0.747
SD	0.009	0.090	0.088	0.090	0.054	0.054
1200Cov1	1.000	0.645	0.657	0.645	0.871	0.876
SD	0.000	0.093	0.090	0.093	0.030	0.033
1200Cov.3	0.964	0.645	0.657	0.645	0.320	0.652
SD	0.030	0.093	0.090	0.093	0.066	0.090
1200Cov.6	0.972	0.645	0.657	0.645	0.460	0.672
SD	0.024	0.093	0.090	0.093	0.057	0.080
1200Cov.9	0.986	0.645	0.656	0.645	0.671	0.734
SD	0.011	0.094	0.090	0.094	0.037	0.052

	LCDM RelA3	Raw RelA3	Dif RelA3	H2 RM2A3	H3 RM3A3	H4 RM4A3
300Cov1	1.000	0.719	0.728	0.719	0.867	0.871
SD	0.000	0.079	0.075	0.079	0.033	0.034
300Cov.3	0.988	0.719	0.728	0.719	0.484	0.723
SD	0.014	0.078	0.074	0.078	0.085	0.076
300Cov.6	0.991	0.719	0.729	0.719	0.589	0.737
SD	0.010	0.078	0.073	0.078	0.069	0.069
300Cov.9	0.995	0.720	0.730	0.720	0.729	0.776
SD	0.006	0.078	0.074	0.078	0.052	0.053
1200Cov1	1.000	0.723	0.733	0.723	0.863	0.867
SD	0.000	0.063	0.060	0.063	0.026	0.026
1200Cov.3	0.989	0.722	0.733	0.722	0.489	0.726
SD	0.012	0.063	0.059	0.063	0.062	0.061
1200Cov.6	0.991	0.723	0.733	0.723	0.588	0.738
SD	0.009	0.063	0.059	0.063	0.051	0.056
1200Cov.9	0.995	0.723	0.733	0.723	0.730	0.776
SD	0.005	0.063	0.059	0.063	0.034	0.041

Total	Raw RelTot	Dif RelTot	Hab RTot
300Cov1	0.857	0.863	0.857
SD	0.030	0.028	0.030
300Cov.3	0.719	0.731	0.719
SD	0.049	0.046	0.049
300Cov.6	0.774	0.783	0.774
SD	0.042	0.040	0.042
300Cov.9	0.824	0.831	0.824
SD	0.035	0.033	0.035
1200Cov1	0.853	0.859	0.853
SD	0.027	0.025	0.027
1200Cov.3	0.713	0.726	0.713
SD	0.039	0.037	0.039
1200Cov.6	0.768	0.778	0.768
SD	0.035	0.033	0.035
1200Cov.9	0.820	0.828	0.820
SD	0.030	0.029	0.030

Appendix 11 LCDM Data Correlation by Attribute

LCDM Data Correlation by Attribute

	LCDM CorA1A2	Raw CorA1A2	Dif CorA1A2	H2 CM2A1A2	H3 CM3A1A2	H4 CM4A1A2
300Cov1	1.000	0.625	0.625	0.625	1.000	0.985
SD	0.000	0.085	0.085	0.085	0.000	0.016
300Cov.3	0.327	0.128	0.128	0.128	1.000	0.289
SD	0.134	0.056	0.056	0.056	0.000	0.110
300Cov.6	0.630	0.261	0.261	0.261	1.000	0.577
SD	0.105	0.060	0.060	0.060	0.000	0.098
300Cov.9	0.918	0.448	0.448	0.448	1.000	0.884
SD	0.036	0.068	0.068	0.068	0.000	0.054
1200Cov1	1.000	0.600	0.600	0.578	1.000	0.999
SD	0.000	0.083	0.083	0.023	0.000	0.001
1200Cov.3	0.316	0.113	0.113	0.172	1.000	0.399
SD	0.059	0.031	0.031	0.027	0.000	0.045
1200Cov.6	0.632	0.245	0.245	0.350	1.000	0.764
SD	0.054	0.041	0.041	0.025	0.000	0.031
1200Cov.9	0.922	0.430	0.430	0.518	1.000	0.981
SD	0.022	0.059	0.059	0.022	0.000	0.008

	LCDM CorA1A3	Raw CorA1A3	Dif CorA1A3	H2 CM2A1A3	H3 CM3A1A3	H4 CM4A1A3
300Cov1	1.000	0.642	0.642	0.642	1.000	0.987
SD	0.000	0.081	0.081	0.081	0.000	0.018
300Cov.3	0.345	0.135	0.135	0.135	1.000	0.313
SD	0.128	0.060	0.060	0.060	0.000	0.126
300Cov.6	0.645	0.274	0.274	0.274	1.000	0.602
SD	0.095	0.068	0.068	0.068	0.000	0.104
300Cov.9	0.917	0.458	0.458	0.458	1.000	0.889
SD	0.039	0.071	0.071	0.071	0.000	0.056
1200Cov1	1.000	0.634	0.634	0.613	1.000	0.999
SD	0.000	0.077	0.077	0.021	0.000	0.002
1200Cov.3	0.312	0.123	0.123	0.186	1.000	0.425
SD	0.059	0.032	0.032	0.029	0.000	0.048
1200Cov.6	0.629	0.260	0.260	0.370	1.000	0.777
SD	0.049	0.044	0.044	0.025	0.000	0.032
1200Cov.9	0.921	0.453	0.453	0.548	1.000	0.983
SD	0.022	0.060	0.060	0.023	0.000	0.009

	LCDM CorA2A3	Raw CorA2A3	Dif CorA2A3	H2 CM2A2A3	H3 CM3A2A3	H4 CM4A2A3
300Cov1	1.000	0.708	0.708	0.708	1.000	0.993
SD	0.000	0.065	0.065	0.065	0.000	0.008
300Cov.3	0.308	0.135	0.135	0.135	1.000	0.276
SD	0.111	0.060	0.060	0.060	0.000	0.112
300Cov.6	0.618	0.291	0.291	0.291	1.000	0.563
SD	0.079	0.055	0.055	0.055	0.000	0.085
300Cov.9	0.909	0.505	0.505	0.505	1.000	0.875
SD	0.036	0.062	0.062	0.062	0.000	0.052
1200Cov1	1.000	0.698	0.698	0.680	1.000	0.999
SD	0.000	0.056	0.056	0.018	0.000	0.001
1200Cov.3	0.305	0.134	0.134	0.203	1.000	0.407
SD	0.054	0.031	0.031	0.028	0.000	0.045
1200Cov.6	0.611	0.283	0.283	0.407	1.000	0.759
SD	0.042	0.035	0.035	0.024	0.000	0.029
1200Cov.9	0.911	0.498	0.498	0.613	1.000	0.980
SD	0.016	0.043	0.043	0.020	0.000	0.009

Appendix 12 Reliability Attribute 1

Reliability Attribute 1

	Rasch Data						LCDM Data					
	LCDM RelA1	**Raw RelA1**	**Dif RelA1**	**H2 RM2A1**	**H3 RM3A1**	**H4 RM4A1**	**LCDM RelA1**	**Raw RelA1**	**Dif RelA1**	**H2 RM2A1**	**H3 RM3A1**	**H4 RM4A1**
300Cov1	0.943	0.510	0.513	0.510	0.862	0.869	0.999	0.537	0.547	0.537	0.926	0.950
SD	0.023	0.046	0.045	0.046	0.061	0.070	0.004	0.139	0.136	0.139	0.134	0.193
300Cov.3	0.810	0.516	0.517	0.516	0.263	0.543	0.911	0.539	0.549	0.539	0.219	0.557
SD	0.045	0.042	0.045	0.042	0.061	0.039	0.062	0.139	0.135	0.139	0.067	0.122
300Cov.6	0.853	0.513	0.515	0.513	0.499	0.618	0.930	0.538	0.548	0.538	0.388	0.600
SD	0.036	0.048	0.047	0.048	0.066	0.038	0.049	0.138	0.135	0.138	0.074	0.100
300Cov.9	0.926	0.515	0.516	0.515	0.767	0.780	0.968	0.539	0.549	0.539	0.646	0.715
SD	0.028	0.044	0.045	0.044	0.059	0.050	0.024	0.137	0.134	0.137	0.071	0.060
1200Cov1	0.943	0.515	0.519	0.515	0.852	0.853	0.999	0.525	0.535	0.525	0.895	0.905
SD	0.012	0.027	0.025	0.027	0.028	0.030	0.003	0.130	0.127	0.130	0.061	0.074
1200Cov.3	0.774	0.518	0.520	0.518	0.261	0.543	0.890	0.524	0.535	0.524	0.196	0.537
SD	0.027	0.028	0.027	0.028	0.030	0.025	0.081	0.129	0.126	0.129	0.051	0.124
1200Cov.6	0.835	0.524	0.527	0.524	0.502	0.623	0.915	0.526	0.536	0.526	0.362	0.578
SD	0.019	0.025	0.025	0.025	0.032	0.019	0.060	0.129	0.126	0.129	0.050	0.103
1200Cov.9	0.910	0.514	0.517	0.514	0.756	0.767	0.964	0.525	0.536	0.525	0.629	0.692
SD	0.017	0.027	0.026	0.027	0.030	0.024	0.025	0.130	0.127	0.130	0.048	0.059

Appendix 13 Reliability Attribute 2

Reliability Attribute 2

	Rasch Data						LCDM Data						
	LCDM RelA2	Raw RelA2	Dif RelA2	H2 RM2A2	H3 RM3A2	H4 RM4A2		LCDM RelA2	Raw RelA2	Dif RelA2	H2 RM2A2	H3 RM3A2	H4 RM4A2
300Cov1	0.962	0.640	0.643	0.640	0.848	0.852	**300Cov1**	1.000	0.663	0.672	0.663	0.880	0.886
SD	0.012	0.035	0.035	0.035	0.030	0.032	**SD**	0.001	0.088	0.086	0.088	0.037	0.041
300Cov.3	0.889	0.640	0.642	0.640	0.384	0.654	**300Cov.3**	0.973	0.662	0.672	0.662	0.335	0.669
SD	0.025	0.032	0.032	0.032	0.054	0.030	**SD**	0.025	0.089	0.087	0.089	0.070	0.083
300Cov.6	0.908	0.637	0.640	0.637	0.575	0.693	**300Cov.6**	0.979	0.662	0.672	0.662	0.476	0.689
SD	0.022	0.035	0.034	0.035	0.047	0.029	**SD**	0.020	0.090	0.088	0.090	0.063	0.075
300Cov.9	0.947	0.644	0.646	0.644	0.778	0.792	**300Cov.9**	0.990	0.662	0.671	0.662	0.681	0.747
SD	0.017	0.037	0.037	0.037	0.036	0.027	**SD**	0.009	0.090	0.088	0.090	0.054	0.054
1200Cov1	0.956	0.642	0.644	0.642	0.845	0.846	**1200Cov1**	1.000	0.645	0.657	0.645	0.871	0.876
SD	0.007	0.021	0.022	0.021	0.015	0.015	**SD**	0.000	0.093	0.090	0.093	0.030	0.033
1200Cov.3	0.870	0.642	0.645	0.642	0.381	0.655	**1200Cov.3**	0.964	0.645	0.657	0.645	0.320	0.652
SD	0.018	0.019	0.019	0.019	0.032	0.018	**SD**	0.030	0.093	0.090	0.093	0.066	0.090
1200Cov.6	0.894	0.643	0.645	0.643	0.578	0.696	**1200Cov.6**	0.972	0.645	0.657	0.645	0.460	0.672
SD	0.015	0.022	0.022	0.022	0.024	0.018	**SD**	0.024	0.093	0.090	0.093	0.057	0.080
1200Cov.9	0.937	0.644	0.646	0.644	0.778	0.789	**1200Cov.9**	0.986	0.645	0.656	0.645	0.671	0.734
SD	0.010	0.020	0.020	0.020	0.018	0.014	**SD**	0.011	0.094	0.090	0.094	0.037	0.052

Appendix 14 Reliability Attribute 3

Reliability Attribute 3

	Rasch Data							LCDM Data					
	LCDM RelA3	**Raw RelA3**	**Dif RelA3**	**H2 RM2A3**	**H3 RM3A3**	**H4 RM4A3**		**LCDM RelA3**	**Raw RelA3**	**Dif RelA3**	**H2 RM2A3**	**H3 RM3A3**	**H4 RM4A3**
300Cov1	0.969	0.719	0.721	0.719	0.848	0.849	**300Cov1**	1.000	0.719	0.728	0.719	0.867	0.871
SD	0.010	0.023	0.024	0.023	0.018	0.019	**SD**	0.000	0.079	0.075	0.079	0.033	0.034
300Cov.3	0.929	0.713	0.715	0.713	0.527	0.721	**300Cov.3**	0.988	0.719	0.728	0.719	0.484	0.723
SD	0.019	0.027	0.026	0.027	0.045	0.026	**SD**	0.014	0.078	0.074	0.078	0.085	0.076
300Cov.6	0.945	0.719	0.720	0.719	0.666	0.751	**300Cov.6**	0.991	0.719	0.729	0.719	0.589	0.737
SD	0.017	0.030	0.029	0.030	0.038	0.026	**SD**	0.010	0.078	0.073	0.078	0.069	0.069
300Cov.9	0.960	0.717	0.718	0.717	0.800	0.809	**300Cov.9**	0.995	0.720	0.730	0.720	0.729	0.776
SD	0.012	0.027	0.027	0.027	0.025	0.021	**SD**	0.006	0.078	0.074	0.078	0.052	0.053
1200Cov1	0.966	0.717	0.719	0.717	0.846	0.846	**1200Cov1**	1.000	0.723	0.733	0.723	0.863	0.867
SD	0.006	0.016	0.016	0.016	0.011	0.011	**SD**	0.000	0.063	0.060	0.063	0.026	0.026
1200Cov.3	0.920	0.718	0.721	0.718	0.532	0.726	**1200Cov.3**	0.989	0.722	0.733	0.722	0.489	0.726
SD	0.013	0.017	0.017	0.017	0.027	0.016	**SD**	0.012	0.063	0.059	0.063	0.062	0.061
1200Cov.6	0.931	0.715	0.718	0.715	0.663	0.748	**1200Cov.6**	0.991	0.723	0.733	0.723	0.588	0.738
SD	0.010	0.015	0.014	0.015	0.018	0.012	**SD**	0.009	0.063	0.059	0.063	0.051	0.056
1200Cov.9	0.953	0.716	0.719	0.716	0.801	0.808	**1200Cov.9**	0.995	0.723	0.733	0.723	0.730	0.776
SD	0.008	0.016	0.016	0.016	0.014	0.011	**SD**	0.005	0.063	0.059	0.063	0.034	0.041

Appendix 15 Correlation Attributes 1 and 2

Correlation Attributes 1 and 2

	Rasch Data						LCDM Data						
	LCDM	Raw	Dif	H2	H3	H4		LCDM	Raw	Dif	H2	H3	H4
	CorA1A2	CorA1A2	CorA1A2	CM2A1A2	CM3A1A2	CM4A1A2		CorA1A2	CorA1A2	CorA1A2	CM2A1A2	CM3A1A2	CM4A1A2
300Cov1	0.981	0.578	0.578	0.578	1.000	0.994	**300Cov1**	1.000	0.625	0.625	0.625	1.000	0.985
SD	0.017	0.043	0.043	0.043	0.000	0.007	**SD**	0.000	0.085	0.085	0.085	0.000	0.016
300Cov.3	0.418	0.171	0.171	0.171	1.000	0.399	**300Cov.3**	0.327	0.128	0.128	0.128	1.000	0.289
SD	0.165	0.058	0.058	0.058	0.000	0.095	**SD**	0.134	0.056	0.056	0.056	0.000	0.110
300Cov.6	0.775	0.344	0.344	0.344	1.000	0.765	**300Cov.6**	0.630	0.261	0.261	0.261	1.000	0.577
SD	0.099	0.045	0.045	0.045	0.000	0.059	**SD**	0.105	0.060	0.060	0.060	0.000	0.098
300Cov.9	0.968	0.523	0.523	0.523	1.000	0.978	**300Cov.9**	0.918	0.448	0.448	0.448	1.000	0.884
SD	0.028	0.041	0.041	0.041	0.000	0.018	**SD**	0.036	0.068	0.068	0.068	0.000	0.054
1200Cov1	0.993	0.578	0.578	0.578	1.000	0.999	**1200Cov1**	1.000	0.600	0.600	0.578	1.000	0.999
SD	0.005	0.023	0.023	0.023	0.000	0.001	**SD**	0.000	0.083	0.083	0.023	0.000	0.001
1200Cov.3	0.431	0.172	0.172	0.172	1.000	0.399	**1200Cov.3**	0.316	0.113	0.113	0.172	1.000	0.399
SD	0.085	0.027	0.027	0.027	0.000	0.045	**SD**	0.059	0.031	0.031	0.027	0.000	0.045
1200Cov.6	0.783	0.350	0.350	0.350	1.000	0.764	**1200Cov.6**	0.632	0.245	0.245	0.350	1.000	0.764
SD	0.061	0.025	0.025	0.025	0.000	0.031	**SD**	0.054	0.041	0.041	0.025	0.000	0.031
1200Cov.9	0.968	0.518	0.518	0.518	1.000	0.981	**1200Cov.9**	0.922	0.430	0.430	0.518	1.000	0.981
SD	0.023	0.022	0.022	0.022	0.000	0.008	**SD**	0.022	0.059	0.059	0.022	0.000	0.008

Appendix 16 Correlation Attributes 1 and 3

Correlation Attributes 1 and 3

	Rasch Data						LCDM Data						
	LCDM CorA1A3	Raw CorA1A3	Dif CorA1A3	H2 CM2A1A3	H3 CM3A1A3	H4 CM4A1A3		LCDM CorA1A3	Raw CorA1A3	Dif CorA1A3	H2 CM2A1A3	H3 CM3A1A3	H4 CM4A1A3
300Cov1	0.980	0.615	0.615	0.615	1.000	0.995	**300Cov1**	1.000	0.642	0.642	0.642	1.000	0.987
SD	0.021	0.036	0.036	0.036	0.000	0.006	**SD**	0.000	0.081	0.081	0.081	0.000	0.018
300Cov.3	0.439	0.183	0.183	0.183	1.000	0.423	**300Cov.3**	0.345	0.135	0.135	0.135	1.000	0.313
SD	0.162	0.056	0.056	0.056	0.000	0.101	**SD**	0.128	0.060	0.060	0.060	0.000	0.126
300Cov.6	0.764	0.366	0.366	0.366	1.000	0.776	**300Cov.6**	0.645	0.274	0.274	0.274	1.000	0.602
SD	0.116	0.057	0.057	0.057	0.000	0.069	**SD**	0.095	0.068	0.068	0.068	0.000	0.104
300Cov.9	0.957	0.553	0.553	0.553	1.000	0.981	**300Cov.9**	0.917	0.458	0.458	0.458	1.000	0.889
SD	0.035	0.041	0.041	0.041	0.000	0.016	**SD**	0.039	0.071	0.071	0.071	0.000	0.056
1200Cov1	0.993	0.613	0.613	0.613	1.000	0.999	**1200Cov1**	1.000	0.634	0.634	0.613	1.000	0.999
SD	0.006	0.021	0.021	0.021	0.000	0.002	**SD**	0.000	0.077	0.077	0.021	0.000	0.002
1200Cov.3	0.445	0.186	0.186	0.186	1.000	0.425	**1200Cov.3**	0.312	0.123	0.123	0.186	1.000	0.425
SD	0.088	0.029	0.029	0.029	0.000	0.048	**SD**	0.059	0.032	0.032	0.029	0.000	0.048
1200Cov.6	0.776	0.370	0.370	0.370	1.000	0.777	**1200Cov.6**	0.629	0.260	0.260	0.370	1.000	0.777
SD	0.054	0.025	0.025	0.025	0.000	0.032	**SD**	0.049	0.044	0.044	0.025	0.000	0.032
1200Cov.9	0.966	0.548	0.548	0.548	1.000	0.983	**1200Cov.9**	0.921	0.453	0.453	0.548	1.000	0.983
SD	0.016	0.023	0.023	0.023	0.000	0.009	**SD**	0.022	0.060	0.060	0.023	0.000	0.009

Appendix 17 Correlation Attributes 2 and 3

Correlation Attributes 2 and 3

	Rasch Data						LCDM Data						
	LCDM CorA2A3	Raw CorA2A3	Dif CorA2A3	H2 CM2A2A3	H3 CM3A2A3	H4 CM4A2A3		LCDM CorA2A3	Raw CorA2A3	Dif CorA2A3	H2 CM2A2A3	H3 CM3A2A3	H4 CM4A2A3
300Cov1	0.986	0.682	0.682	0.682	1.000	0.997	**300Cov1**	1.000	0.708	0.708	0.708	1.000	0.993
SD	0.016	0.032	0.032	0.032	0.000	0.004	**SD**	0.000	0.065	0.065	0.065	0.000	0.008
300Cov.3	0.409	0.204	0.204	0.204	1.000	0.410	**300Cov.3**	0.308	0.135	0.135	0.135	1.000	0.276
SD	0.173	0.054	0.054	0.054	0.000	0.091	**SD**	0.111	0.060	0.060	0.060	0.000	0.112
300Cov.6	0.761	0.407	0.407	0.407	1.000	0.759	**300Cov.6**	0.618	0.291	0.291	0.291	1.000	0.563
SD	0.087	0.048	0.048	0.048	0.000	0.062	**SD**	0.079	0.055	0.055	0.055	0.000	0.085
300Cov.9	0.959	0.611	0.611	0.611	1.000	0.976	**300Cov.9**	0.909	0.505	0.505	0.505	1.000	0.875
SD	0.027	0.039	0.039	0.039	0.000	0.019	**SD**	0.036	0.062	0.062	0.062	0.000	0.052
1200Cov1	0.991	0.680	0.680	0.680	1.000	0.999	**1200Cov1**	1.000	0.698	0.698	0.680	1.000	0.999
SD	0.006	0.018	0.018	0.018	0.000	0.001	**SD**	0.000	0.056	0.056	0.018	0.000	0.001
1200Cov.3	0.398	0.203	0.203	0.203	1.000	0.407	**1200Cov.3**	0.305	0.134	0.134	0.203	1.000	0.407
SD	0.071	0.028	0.028	0.028	0.000	0.045	**SD**	0.054	0.031	0.031	0.028	0.000	0.045
1200Cov.6	0.743	0.407	0.407	0.407	1.000	0.759	**1200Cov.6**	0.611	0.283	0.283	0.407	1.000	0.759
SD	0.053	0.024	0.024	0.024	0.000	0.029	**SD**	0.042	0.035	0.035	0.024	0.000	0.029
1200Cov.9	0.960	0.613	0.613	0.613	1.000	0.980	**1200Cov.9**	0.911	0.498	0.498	0.613	1.000	0.980
SD	0.019	0.020	0.020	0.020	0.000	0.009	**SD**	0.016	0.043	0.043	0.020	0.000	0.009

Appendix 18 Raw Method

Raw Method - Reliability Estimates

Rasch Data

	RelA1	RelA2	RelA3	RelTot
300Cov1	0.510	0.640	0.719	0.844
SD	0.046	0.035	0.023	0.013
300Cov.3	0.516	0.640	0.713	0.728
SD	0.042	0.032	0.027	0.023
300Cov.6	0.513	0.637	0.719	0.792
SD	0.048	0.035	0.030	0.020
300Cov.9	0.515	0.644	0.717	0.833
SD	0.044	0.037	0.027	0.016
1200Cov1	0.515	0.642	0.717	0.844
SD	0.027	0.021	0.016	0.008
1200Cov.3	0.518	0.642	0.718	0.732
SD	0.028	0.019	0.017	0.012
1200Cov.6	0.524	0.643	0.715	0.793
SD	0.025	0.022	0.015	0.010
1200Cov.9	0.514	0.644	0.716	0.833
SD	0.027	0.020	0.016	0.007

Raw Method - Reliability Estimates

LCDM Data

	RelA1	RelA2	RelA3	RelTot
300Cov1	0.537	0.663	0.719	0.857
SD	0.139	0.088	0.079	0.030
300Cov.3	0.539	0.662	0.719	0.719
SD	0.139	0.089	0.078	0.049
300Cov.6	0.538	0.662	0.719	0.774
SD	0.138	0.090	0.078	0.042
300Cov.9	0.539	0.662	0.720	0.824
SD	0.137	0.090	0.078	0.035
1200Cov1	0.525	0.645	0.723	0.853
SD	0.130	0.093	0.063	0.027
1200Cov.3	0.524	0.645	0.722	0.713
SD	0.129	0.093	0.063	0.039
1200Cov.6	0.526	0.645	0.723	0.768
SD	0.129	0.093	0.063	0.035
1200Cov.9	0.525	0.645	0.723	0.820
SD	0.130	0.094	0.063	0.030

Raw Method - Correlation Estimates

Rasch Data

	CorA1A2	CorA1A3	CorA2A3
300Cov1	0.578	0.615	0.682
SD	0.043	0.036	0.032
300Cov.3	0.171	0.183	0.204
SD	0.058	0.056	0.054
300Cov.6	0.344	0.366	0.407
SD	0.045	0.057	0.048
300Cov.9	0.523	0.553	0.611
SD	0.041	0.041	0.039
1200Cov1	0.578	0.613	0.680
SD	0.023	0.021	0.018
1200Cov.3	0.172	0.186	0.203
SD	0.027	0.029	0.028
1200Cov.6	0.350	0.370	0.407
SD	0.025	0.025	0.024
1200Cov.9	0.518	0.548	0.613
SD	0.022	0.023	0.020

Raw Method - Correlation Estimates

LCDM Data

	CorA1A2	CorA1A3	CorA2A3
300Cov1	0.625	0.642	0.708
SD	0.085	0.081	0.065
300Cov.3	0.128	0.135	0.135
SD	0.056	0.060	0.060
300Cov.6	0.261	0.274	0.291
SD	0.060	0.068	0.055
300Cov.9	0.448	0.458	0.505
SD	0.068	0.071	0.062
1200Cov1	0.600	0.634	0.698
SD	0.083	0.077	0.056
1200Cov.3	0.113	0.123	0.134
SD	0.031	0.032	0.031
1200Cov.6	0.245	0.260	0.283
SD	0.041	0.044	0.035
1200Cov.9	0.430	0.453	0.498
SD	0.059	0.060	0.043

Appendix 19 a priori Differential Method

Differential Method - Reliability Estimates
Rasch Data

	RelA1	RelA2	RelA3	RelTot
300Cov1	0.513	0.643	0.721	0.846
SD	0.045	0.035	0.024	0.013
300Cov.3	0.517	0.642	0.715	0.734
SD	0.045	0.032	0.026	0.022
300Cov.6	0.515	0.640	0.720	0.795
SD	0.047	0.034	0.029	0.019
300Cov.9	0.516	0.646	0.718	0.835
SD	0.045	0.037	0.027	0.016
1200Cov1	0.519	0.644	0.719	0.845
SD	0.025	0.022	0.016	0.008
1200Cov.3	0.520	0.645	0.721	0.738
SD	0.027	0.019	0.017	0.012
1200Cov.6	0.527	0.645	0.718	0.796
SD	0.025	0.022	0.014	0.010
1200Cov.9	0.517	0.646	0.719	0.835
SD	0.026	0.020	0.016	0.008

Differential Method - Reliability Estimates
LCDM Data

	RelA1	RelA2	RelA3	RelTot
300Cov1	0.547	0.672	0.728	0.863
SD	0.136	0.086	0.075	0.028
300Cov.3	0.549	0.672	0.728	0.731
SD	0.135	0.087	0.074	0.046
300Cov.6	0.548	0.672	0.729	0.783
SD	0.135	0.088	0.073	0.040
300Cov.9	0.549	0.671	0.730	0.831
SD	0.134	0.088	0.074	0.033
1200Cov1	0.535	0.657	0.733	0.859
SD	0.127	0.090	0.060	0.025
1200Cov.3	0.535	0.657	0.733	0.726
SD	0.126	0.090	0.059	0.037
1200Cov.6	0.536	0.657	0.733	0.778
SD	0.126	0.090	0.059	0.033
1200Cov.9	0.536	0.656	0.733	0.828
SD	0.127	0.090	0.059	0.029

Differential Method - Correlation Estimates
Rasch Data

	CorA1A2	CorA1A3	CorA2A3
300Cov1	0.578	0.615	0.682
SD	0.043	0.036	0.032
300Cov.3	0.171	0.183	0.204
SD	0.058	0.056	0.054
300Cov.6	0.344	0.366	0.407
SD	0.045	0.057	0.048
300Cov.9	0.523	0.553	0.611
SD	0.041	0.041	0.039
1200Cov1	0.578	0.613	0.680
SD	0.023	0.021	0.018
1200Cov.3	0.172	0.186	0.203
SD	0.027	0.029	0.028
1200Cov.6	0.350	0.370	0.407
SD	0.025	0.025	0.024
1200Cov.9	0.518	0.548	0.613
SD	0.022	0.023	0.020

Differential Method - Correlation Estimates
LCDM Data

	CorA1A2	CorA1A3	CorA2A3
300Cov1	0.625	0.642	0.708
SD	0.085	0.081	0.065
300Cov.3	0.128	0.135	0.135
SD	0.056	0.060	0.060
300Cov.6	0.261	0.274	0.291
SD	0.060	0.068	0.055
300Cov.9	0.448	0.458	0.505
SD	0.068	0.071	0.062
1200Cov1	0.600	0.634	0.698
SD	0.083	0.077	0.056
1200Cov.3	0.113	0.123	0.134
SD	0.031	0.032	0.031
1200Cov.6	0.245	0.260	0.283
SD	0.041	0.044	0.035
1200Cov.9	0.430	0.453	0.498
SD	0.059	0.060	0.043

Appendix 20 Haberman 2 Method

Haberman 2 Method - Reliability Estimates
Rasch Data

	RM2A1	RM2A2	RM2A3	RTot
300Cov1	0.510	0.640	0.719	0.844
SD	0.046	0.035	0.023	0.013
300Cov.3	0.516	0.640	0.713	0.728
SD	0.042	0.032	0.027	0.023
300Cov.6	0.513	0.637	0.719	0.792
SD	0.048	0.035	0.030	0.020
300Cov.9	0.515	0.644	0.717	0.833
SD	0.044	0.037	0.027	0.016
1200Cov1	0.515	0.642	0.717	0.844
SD	0.027	0.021	0.016	0.008
1200Cov.3	0.518	0.642	0.718	0.732
SD	0.028	0.019	0.017	0.012
1200Cov.6	0.524	0.643	0.715	0.793
SD	0.025	0.022	0.015	0.010
1200Cov.9	0.514	0.644	0.716	0.833
SD	0.027	0.020	0.016	0.007

Haberman 2 Method - Reliability Estimates
LCDM Data

	RM2A1	RM2A2	RM2A3	RTot
300Cov1	0.537	0.663	0.719	0.857
SD	0.139	0.088	0.079	0.030
300Cov.3	0.539	0.662	0.719	0.719
SD	0.139	0.089	0.078	0.049
300Cov.6	0.538	0.662	0.719	0.774
SD	0.138	0.090	0.078	0.042
300Cov.9	0.539	0.662	0.720	0.824
SD	0.137	0.090	0.078	0.035
1200Cov1	0.525	0.645	0.723	0.853
SD	0.130	0.093	0.063	0.027
1200Cov.3	0.524	0.645	0.722	0.713
SD	0.129	0.093	0.063	0.039
1200Cov.6	0.526	0.645	0.723	0.768
SD	0.129	0.093	0.063	0.035
1200Cov.9	0.525	0.645	0.723	0.820
SD	0.130	0.094	0.063	0.030

Haberman 2 Method - Correlation Estimates
Rasch Data

	CM2A1A2	CM2A1A3	CM2A2A3
300Cov1	0.578	0.615	0.682
SD	0.043	0.036	0.032
300Cov.3	0.171	0.183	0.204
SD	0.058	0.056	0.054
300Cov.6	0.344	0.366	0.407
SD	0.045	0.057	0.048
300Cov.9	0.523	0.553	0.611
SD	0.041	0.041	0.039
1200Cov1	0.578	0.613	0.680
SD	0.023	0.021	0.018
1200Cov.3	0.172	0.186	0.203
SD	0.027	0.029	0.028
1200Cov.6	0.350	0.370	0.407
SD	0.025	0.025	0.024
1200Cov.9	0.518	0.548	0.613
SD	0.022	0.023	0.020

Haberman 2 Method - Correlation Estimates
LCDM Data

	CM2A1A2	CM2A1A3	CM2A2A3
300Cov1	0.625	0.642	0.708
SD	0.085	0.081	0.065
300Cov.3	0.128	0.135	0.135
SD	0.056	0.060	0.060
300Cov.6	0.261	0.274	0.291
SD	0.060	0.068	0.055
300Cov.9	0.448	0.458	0.505
SD	0.068	0.071	0.062
1200Cov1	0.578	0.613	0.680
SD	0.023	0.021	0.018
1200Cov.3	0.172	0.186	0.203
SD	0.027	0.029	0.028
1200Cov.6	0.350	0.370	0.407
SD	0.025	0.025	0.024
1200Cov.9	0.518	0.548	0.613
SD	0.022	0.023	0.020

Appendix 21 Haberman 3 Method

Haberman 3 Method - Reliability Estimates
Rasch Data

	RM3A1	RM3A2	RM3A3
300Cov1	0.862	0.848	0.848
SD	0.061	0.030	0.018
300Cov.3	0.263	0.384	0.527
SD	0.061	0.054	0.045
300Cov.6	0.499	0.575	0.666
SD	0.066	0.047	0.038
300Cov.9	0.767	0.778	0.800
SD	0.059	0.036	0.025
1200Cov1	0.852	0.845	0.846
SD	0.028	0.015	0.011
1200Cov.3	0.261	0.381	0.532
SD	0.030	0.032	0.027
1200Cov.6	0.502	0.578	0.663
SD	0.032	0.024	0.018
1200Cov.9	0.756	0.778	0.801
SD	0.030	0.018	0.014

Haberman 3 Method - Reliability Estimates
LCDM Data

	RM3A1	RM3A2	RM3A3
300Cov1	0.926	0.880	0.867
SD	0.134	0.037	0.033
300Cov.3	0.219	0.335	0.484
SD	0.067	0.070	0.085
300Cov.6	0.388	0.476	0.589
SD	0.074	0.063	0.069
300Cov.9	0.646	0.681	0.729
SD	0.071	0.054	0.052
1200Cov1	0.895	0.871	0.863
SD	0.061	0.030	0.026
1200Cov.3	0.196	0.320	0.489
SD	0.051	0.066	0.062
1200Cov.6	0.362	0.460	0.588
SD	0.050	0.057	0.051
1200Cov.9	0.629	0.671	0.730
SD	0.048	0.037	0.034

Haberman 3 Method - Correlation Estimates
Rasch Data

	CM3A1A2	CM3A1A3	CM3A2A3
300Cov1	1.000	1.000	1.000
SD	0.000	0.000	0.000
300Cov.3	1.000	1.000	1.000
SD	0.000	0.000	0.000
300Cov.6	1.000	1.000	1.000
SD	0.000	0.000	0.000
300Cov.9	1.000	1.000	1.000
SD	0.000	0.000	0.000
1200Cov1	1.000	1.000	1.000
SD	0.000	0.000	0.000
1200Cov.3	1.000	1.000	1.000
SD	0.000	0.000	0.000
1200Cov.6	1.000	1.000	1.000
SD	0.000	0.000	0.000
1200Cov.9	1.000	1.000	1.000
SD	0.000	0.000	0.000

Haberman 3 Method - Correlation Estimates
LCDM Data

	CM3A1A2	CM3A1A3	CM3A2A3
300Cov1	1.000	1.000	1.000
SD	0.000	0.000	0.000
300Cov.3	1.000	1.000	1.000
SD	0.000	0.000	0.000
300Cov.6	1.000	1.000	1.000
SD	0.000	0.000	0.000
300Cov.9	1.000	1.000	1.000
SD	0.000	0.000	0.000
1200Cov1	1.000	1.000	1.000
SD	0.000	0.000	0.000
1200Cov.3	1.000	1.000	1.000
SD	0.000	0.000	0.000
1200Cov.6	1.000	1.000	1.000
SD	0.000	0.000	0.000
1200Cov.9	1.000	1.000	1.000
SD	0.000	0.000	0.000

Appendix 22 Haberman 4 Method

Haberman 4 Method - Reliability Estimates
Rasch Data

	RM4A1	RM4A2	RM4A3
300Cov1	0.869	0.852	0.849
SD	0.070	0.032	0.019
300Cov.3	0.543	0.654	0.721
SD	0.039	0.030	0.026
300Cov.6	0.618	0.693	0.751
SD	0.038	0.029	0.026
300Cov.9	0.780	0.792	0.809
SD	0.050	0.027	0.021
1200Cov1	0.853	0.846	0.846
SD	0.030	0.015	0.011
1200Cov.3	0.543	0.655	0.726
SD	0.025	0.018	0.016
1200Cov.6	0.623	0.696	0.748
SD	0.019	0.018	0.012
1200Cov.9	0.767	0.789	0.808
SD	0.024	0.014	0.011

Haberman 4 Method - Reliability Estimates
LCDM Data

	RM4A1	RM4A2	RM4A3
300Cov1	0.950	0.886	0.871
SD	0.193	0.041	0.034
300Cov.3	0.557	0.669	0.723
SD	0.122	0.083	0.076
300Cov.6	0.600	0.689	0.737
SD	0.100	0.075	0.069
300Cov.9	0.715	0.747	0.776
SD	0.060	0.054	0.053
1200Cov1	0.905	0.876	0.867
SD	0.074	0.033	0.026
1200Cov.3	0.537	0.652	0.726
SD	0.124	0.090	0.061
1200Cov.6	0.578	0.672	0.738
SD	0.103	0.080	0.056
1200Cov.9	0.692	0.734	0.776
SD	0.059	0.052	0.041

Haberman 4 Method - Correlation Estimates
Rasch Data

	CM4A1A2	CM4A1A3	CM4A2A3
300Cov1	0.994	0.995	0.997
SD	0.007	0.006	0.004
300Cov.3	0.399	0.423	0.410
SD	0.095	0.101	0.091
300Cov.6	0.765	0.776	0.759
SD	0.059	0.069	0.062
300Cov.9	0.978	0.981	0.976
SD	0.018	0.016	0.019
1200Cov1	0.999	0.999	0.999
SD	0.001	0.002	0.001
1200Cov.3	0.399	0.425	0.407
SD	0.045	0.048	0.045
1200Cov.6	0.764	0.777	0.759
SD	0.031	0.032	0.029
1200Cov.9	0.981	0.983	0.980
SD	0.008	0.009	0.009

Haberman 4 Method - Correlation Estimates
LCDM Data

	CM4A1A2	CM4A1A3	CM4A2A3
300Cov1	0.985	0.987	0.993
SD	0.016	0.018	0.008
300Cov.3	0.289	0.313	0.276
SD	0.110	0.126	0.112
300Cov.6	0.577	0.602	0.563
SD	0.098	0.104	0.085
300Cov.9	0.884	0.889	0.875
SD	0.054	0.056	0.052
1200Cov1	0.999	0.999	0.999
SD	0.001	0.002	0.001
1200Cov.3	0.399	0.425	0.407
SD	0.045	0.048	0.045
1200Cov.6	0.764	0.777	0.759
SD	0.031	0.032	0.029
1200Cov.9	0.981	0.983	0.980
SD	0.008	0.009	0.009

Appendix 23 LCDM Method

LCDM Method - Reliability Estimates
Rasch Data

	RelA1	RelA2	RelA3
300Cov1	0.943	0.962	0.969
SD	0.023	0.012	0.010
300Cov.3	0.810	0.889	0.929
SD	0.045	0.025	0.019
300Cov.6	0.853	0.908	0.945
SD	0.036	0.022	0.017
300Cov.9	0.926	0.947	0.960
SD	0.028	0.017	0.012
1200Cov1	0.943	0.956	0.966
SD	0.012	0.007	0.006
1200Cov.3	0.774	0.870	0.920
SD	0.027	0.018	0.013
1200Cov.6	0.835	0.894	0.931
SD	0.019	0.015	0.010
1200Cov.9	0.910	0.937	0.953
SD	0.017	0.010	0.008

LCDM Method - Reliability Estimates
LCDM Data

	RelA1	RelA2	RelA3
300Cov1	0.999	1.000	1.000
SD	0.004	0.001	0.000
300Cov.3	0.911	0.973	0.988
SD	0.062	0.025	0.014
300Cov.6	0.930	0.979	0.991
SD	0.049	0.020	0.010
300Cov.9	0.968	0.990	0.995
SD	0.024	0.009	0.006
1200Cov1	0.999	1.000	1.000
SD	0.003	0.000	0.000
1200Cov.3	0.890	0.964	0.989
SD	0.081	0.030	0.012
1200Cov.6	0.915	0.972	0.991
SD	0.060	0.024	0.009
1200Cov.9	0.964	0.986	0.995
SD	0.025	0.011	0.005

LCDM Method - Correlation Estimates
Rasch Data

	CorA1A2	CorA1A3	CorA2A3
300Cov1	0.981	0.980	0.986
SD	0.017	0.021	0.016
300Cov.3	0.418	0.439	0.409
SD	0.165	0.162	0.173
300Cov.6	0.775	0.764	0.761
SD	0.099	0.116	0.087
300Cov.9	0.968	0.957	0.959
SD	0.028	0.035	0.027
1200Cov1	0.993	0.993	0.991
SD	0.005	0.006	0.006
1200Cov.3	0.431	0.445	0.398
SD	0.085	0.088	0.071
1200Cov.6	0.783	0.776	0.743
SD	0.061	0.054	0.053
1200Cov.9	0.968	0.966	0.960
SD	0.023	0.016	0.019

LCDM Method - Correlation Estimates
LCDM Data

	CorA1A2	CorA1A3	CorA2A3
300Cov1	1.000	1.000	1.000
SD	0.000	0.000	0.000
300Cov.3	0.327	0.345	0.308
SD	0.134	0.128	0.111
300Cov.6	0.630	0.645	0.618
SD	0.105	0.095	0.079
300Cov.9	0.918	0.917	0.909
SD	0.036	0.039	0.036
1200Cov1	1.000	1.000	1.000
SD	0.000	0.000	0.000
1200Cov.3	0.316	0.312	0.305
SD	0.059	0.059	0.054
1200Cov.6	0.632	0.629	0.611
SD	0.054	0.049	0.042
1200Cov.9	0.922	0.921	0.911
SD	0.022	0.022	0.016

Appendix 24 LCDM Analysis Ouptput

LCDM Item Parameters: In this simulation the Q-matrix is a simple Q-matrix, one item for each attribute. The item parameters estimated are the intercept and main effect for each item.

Attribute Reliability: The reliability for each attribute using the polychoric correlation method of calculation.

Marginal Estimates: MAP (Maximum a posteriori). One method of assigning latent class attribute probability profiles. Can be difficult to interpret due to no direct probability estimates for each individual attribute for each person. These estimates are reported in the form of 0s and 1s. These are a rounded version of the EAP Marginal Estimates (Rupp et al., 2010).

MAP Profile Estimates: The estimate of which profile the participant belongs based on the MAP output. In this simulation there were eight profiles. [000] [001] [010] [100] [011] [110] [101] [111]

EAP Marginal Estimates: EAP (Expected a posteriori). A second method of assigning latent class attribute probabilities. These are estimates probabilities for each participant, for each individual attribute. This is more suitable due to the estimated expected value for each individual attribute for each person (Rupp et al., 2010).

EAP Profile Estimates: The estimate of which profile the participant belongs based on the EAP output.