COMPARING THREE FRAMEWORKS TO ASSESS INTERVENTION EFFECT ON EDUCATINOAL GROWTH

by

SEUNG MIN OH

(Under the Direction of Seock-Ho Kim)

ABSTRACT

Conducting intervention to promote educational growth is a typical topic in educational research. By comparing control versus treatment condition, educational researchers can gain if the intervention effect is significant. Classical test theory, item response theory and diagnostic classification model are widely used methods to evaluate intervention effect on ability growth. The three frameworks produce unique information that estimates item statistics, ability estimates, and intervention effects. While the three frameworks are widely used, the comparison among three frameworks on the same assessment data are less highlighted. By using three frameworks to the single randomized control-group pretest-posttest research design, this study will first aim to obtain item and ability estimates. Furthermore, this study will conduct analyses on how three frameworks produce unique information on intervention effects in math assessment. Lastly, this study will suggest theoretical and practical guidance towards which framework to assess intervention effect in educational growth.

INDEX WORDS: Intervention effect, classical test theory, item response theory, diagnostic classification model, math assessment, randomized control-group pretest-posttest research design

COMPARING THREE FRAMEWORKS TO ASSESS INTERVENTION EFFECT ON EDUCATINOAL GROWTH

by

SEUNG MIN OH

BA in Education, Kookmin University, 2021, South Korea

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF ARTS

ATHENS, GEORGIA

2023

COMPARING THREE FRAMEWORKS TO ASSESS INTERVENTION EFFECT ON EDUCATINOAL GROWTH

by

SEUNG MIN OH

Major Professor:	Seock-Ho Kim
Committee:	Zhenqiu Lu
	Matthew Madison

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
May 2023

DEDICATION

I would like to express my appreciation and indebtedness to my parents, Chang Boon Kang and Sang Kwan Oh, who always provided unchanging support for an immature me.

ACKNOWLEDGEMENTS

First, I would like to express gratitude to my academic advisor, Dr. Seock-Ho Kim who guided me thorough master's degree. Dr. Seock-Ho Kim supervised me from the first semester to the last semester in quantitative methodology program. The foundational courses in applied analysis of variance provided me with the stalwart statistical foundations. Furthermore, Dr. Seock-Ho Kim provided me the opportunity to join the research team, which even fostered me greatly throughout my master's degree.

Furthermore, I would like to express gratitude to the committee members, Dr. Zhenqiu Lu and Dr. Matthew Madison. The lectures in structural equation modeling and diagnostic classification methodology guided me to the new interests in quantitative methodology. The lectures that supported my knowledge and guided my future career to continue as a graduate student. I also want to express gratitude again for supporting my thesis project.

This thesis could not be completed without the support of Dr. Allan S. Cohen and Dr. Yasemin Copur-Gencturk who provided research assistantship opportunity and permission to use the dataset used in this research. Building experiences in research is much more bigger than tuition waivers and stipends. During the research assistantship, the help from Dr. Allan S. Cohen and Dr. Yasemin Copur-Gencturk fostered my theoretical backgrounds and applied experiences.

This paper is based on work supported by the Institute of Education Sciences under Grant Number R305A180392. All opinions and conclusions in this paper are those of the authors and do not necessarily reflect the views of the funding agency.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

Introduction

There are numerous trials to promote educational growth. A typical program to promote educational growth is to apply interventions to teachers or students. The intervention programs can be content-related (i.e., new instruction methods to teach ratios and proportional knowledge), while other intervention programs (i.e., motivational intervention programs) have been studied to have positive effect in educational growth. Regardless of intervention type, one goal of intervention in educational research is to promote educational growth.

Educational growth can be assessed by a variety of methods. A researcher may use quantitative, qualitative, or mixed methods for evaluation. If the researcher finds that the intervention promotes desired educational growth, the researcher can support the intervention program to be applied. In quantitative research methodology, comparing the significant differences between the control group without intervention to treatment group with intervention is a widely used research methodology. However, the decision of which quantitative methodology to use still depends on researcher's decision.

To evaluate the ability growth, educational researchers may use three frameworks to quantify the ability growth. First, classical test theory (CTT) uses the gain score as the criterion. Gain score is the difference between posttest score and pretest score. CTT provides continuous variables as an outcome. However, CTT has major limitations. The ability estimates provided by CTT are dependent on the sample. Moreover, the item statistics provided by CTT are also dependent on the sample. Furthermore, many researchers criticize CTT framework for having

lower reliability. Item response theory (IRT) is a logistic model that estimates item parameters and ability estimates from the item responses. IRT also provides continuous variables as an outcome. Item parameter estimates and ability estimates from IRT are sample independent, providing more sound information compared to CTT. However, IRT also has limitations. First, IRT requires a large number of participants to have reliable item statistics and ability estimates. Furthermore, IRT shows lower consistency compared to diagnostic classification model (DCM) when the number of items is small. DCM is also a logistic model that estimates item parameters and latent class. DCM also provides continuous outcomes for item statistics. However, unlike CTT and IRT, DCM provides categorical variables (i.e., latent class) for ability estimates. DCM uses latent transition analysis (LTA) to assess ability growth. DCM has unique characteristics by providing categorical variables for classification but is not suitable for providing continuous outcomes by its nature. Moreover, DCM is strongly impacted by the quality of the Q-matrix that defines attributes required for solving items. Furthermore, DCM still needs to investigate more indices to judge model and item fit.

All three frameworks can be applied to single-group, pretest-posttest research design. However, the research design cannot determine the exact cause of the ability growth. Due to single-group, pretest-posttest research design's limitation, educational researchers may want to compare differential growth between groups, known as control-group pretest-posttest research design. With properly randomized samples, the three frameworks can also be used in evaluating intervention effects. For CTT and IRT frameworks, analysis of variance (ANOVA) or analysis of covariance (ANCOVA) using continuous ability estimates is often implemented to quantify significant difference between groups. For DCM, educational researchers use transition diagnostic classification model (TDCM) to assess intervention effects.

This study's purpose is to compare three frameworks for assessing intervention effects in control-group pretest-posttest research design. More specifically, this study uses unidimensional, ten-item empirical data to assess intervention effect in students' math knowledge in ratios and proportions.

After the introduction, this study will conduct a literature review. Firstly, this study will review the brief history of educational intervention programs. Secondly, this research will review the history, theoretical backgrounds, and real-world usage of each of the three frameworks.

After the literature review session, this study will first describe the general process of discarding the specific item and population from the original data. The original assessment was an eleven-item assessment that consisted of nine multiple-choice items and two multiple-selection items. However, a specific item was too difficult and led this study to discard the item. Moreover, a total of 2174 students took the original assessment (either pretest or posttest) but included only 1727 students that had both the pretest and the posttest.

In the next chapter, this study will conduct applied study by using three frameworks from the ten-item assessment. First, this study will estimate item parameter estimates and ability estimates provided by CTT, IRT, and DCM. Second, this study will compare item parameter estimates and person parameter estimates obtained from three frameworks. Lastly, this study will propose practical theoretical and practical guidelines for selecting assessment models for future research. While CTT is the simplest framework, the need for alternative frameworks is desired due to major limitations of CTT. Theoretically, this study will provide how to compare and select sub-models in three frameworks. Furthermore, this study will provide how three frameworks yield unique information in assessing item parameters, person parameters, reliability

and intervention effects. Pragmatically, this study will interpret statistics estimated by each framework and propose implications and guidelines for future implementations.

In the next session, this study will open discussion about major limitations of the study. More specifically, the items were not originally developed to be used in DCM framework, lacking apt Q-matrix for use. As a result, the study assumed only one attribute (i.e., ratios and proportional relationship knowledge) as the only attribute. In this chapter, this study will also open discussion for how single factor, unidimensional approach can cause limitations in IRT framework. Moreover, this study will discuss major limitations of the study. The first limitation of this study is ignoring hierarchy of the data structure. The second limitation of this study, typically IRT and DCM estimation, is assumption of item invariance properties. Although IRT and DCM generally possess item invariance properties, the sample size cannot guarantee that the sample is fully representative of the population, resulting in possible item variance properties.

CHAPTER 2

Literature Review

Educational Intervention

There is copious research aimed to increase educational outcomes. A typical research interest is introducing interventions. Intervention programs can be aimed towards students and teachers. The intervention programs' goal can be related to educational contents directly. On the other hand, the intervention program can affect other constructs that produce indirect effects on educational outcomes. Furthermore, the intervention program can have face-to-face or asynchronous forms. This study will briefly introduce some educational intervention studies.

Research conducted by Lowrie and colleagues (2018) explored intervention programs on students in spatial reasoning ability. The research conducted a spatial training program students by introducing spatial reasoning instrument with trained teachers. The instrument consists of mental rotation, spatial orientation, and spatial visualization. After ten weeks of intervention, the study concluded that spatial intervention significantly increased the students' spatial ability throughout all spatial ranks (i.e., low, medium, and high groups).

Another research by Yeager and colleagues (2019) conducted a motivational intervention on students' growth mindset and quantified the effect of motivational intervention. This study had 12,490 ninth grade students collected from 65 regular public schools. Students in both control and treatment conditions received learning contents about brain function. However, only the treatment group received the idea of growth mindset (i.e., efforts and strategies can change the educational outcomes) for two sessions. The study concluded that the treatment group who

received growth mindset intervention scored significantly higher GPAs at the end of the academic year.

There is other intervention research that focuses on intervention programs for teachers. A study by Campbell and colleagues (2011) suggested that intervention on teachers that focused on content-focused professional development (PD) increased the quality of instruction and educational outcomes. This study conducted the randomized control study for three years. The PD program provided feedback on the teachers' content knowledge. The study concluded that the PD program significantly increased both teachers' instruction quality and educational outcomes compared to the control group. Campbell and colleagues suggested that the PD program produced significant difference for the second and third year but did not produce significant difference for the first year. The study also suggested that the effect size of the PD on educational outcomes increased in the second and the third year.

There is another research that focuses on online PD (OPD). Masters and colleagues (2010) constructed randomized control experiments to examine the effect of OPD on knowledge and instruction practices on vocabulary, reading, and writing. The research concluded that the OPD intervention produced small effects on vocabulary knowledge and medium effect on reading practices. On the other hand, OPD produced large effects on vocabulary practices, reading knowledge, writing knowledge, and writing practices. Regardless of the effect size, the OPD intervention had significant effect in all criteria.

Classical Test Theory

CTT framework is the most basic framework that provides continuous item parameters and person estimates. CTT framework provides the raw score that consists of the true score and measurement error. CTT assumes that the expected value of the measurement error is equal to

zero. As a result, CTT framework directly uses the raw score to represent latent ability. CTT framework assumes that the test is unidimensional, indicating that only one factor (or attribute) can be measured by an assessment. In order to assess multiple attributes, CTT framework requires at least an equivalent number of assessments. CTT items can be scored dichotomously, indicating that the correct answer is coded into 1 and wrong answers are coded into 0. On the other hand, CTT items can be scored on polytomous scale. CTT framework can be applied by having all items weighted equally or differently.

CTT framework does not typically require model estimation process. This is the reason that CTT models are the most basic and the most prevalent framework used in schools. However, CTT has undesired properties that researchers do not use the raw score to explore item characteristics and ability estimates. The first limitation of CTT framework is the lack of item invariance property. Item invariance refers to the characteristic of psychometric model that item parameters are independent from the group's characteristic. Mango (2009) conducted research by using chemistry test data to compare item characteristics derived from CTT and IRT frameworks. Mango suggested that IRT difficulty parameters were independent from the sample and had less measurement errors compared to CTT. Difficulty parameter, which is defined as the proportion of samples answering the answer correctly, differs by the level ability estimates. For example, if a moderate difficulty math test assessing sixth grades mathematics knowledge is given to third grade students, the difficulty parameter will indicate that the test is extremely hard. On the other hand, if the same test is given to mathematics major university students, the difficulty parameter is extremely high, indicating that the test is extremely easy. By using the different subsample from the population, difficulty parameter obtained from CTT framework

also varies. If a researcher divides the sample into lower, medium, and higher performing groups, the difficulty parameter will suggest difficult, moderate, and easy tests.

Furthermore, the discrimination parameter is obtained by subtracting the difficulty parameter of lower performing group from the higher performing group. Since the difficulty parameter in CTT framework is not consistent over samples, the discrimination obtained from the difficulty parameter is also not consistent over samples. Moreover, the definition of the lower and higher performing group is not consistent. The discrimination parameter can be obtained by splitting the sample into halves, while some divide the sample into three or four groups and compare the highest group to the lowest group.

Furthermore, the raw score obtained from CTT framework does not reflect the item characteristics. This is due to CTT framework providing equal raw scores without considering the difficulty of the item. CTT framework only uses the raw score containing only information about the right or wrong. In IRT framework, ability estimates reflect the difficulty of the item, indicating that the students will gain higher ability estimates by answering more difficult items than answering easier items. In DCM framework, students may be classified into different latent classes even though they have the same raw scores. Therefore, CTT framework cannot assure that participants with same raw score will have the same latent abilities.

The strength of CTT is that it's easy accessibility. Unlike IRT and DCM, CTT does not typically require model estimation process. Furthermore, the raw score provided by CTT is straightforward and easy-to-understand. As a result, the characteristics of CTT framework lead CTT framework to be the most popular and prevalent framework used in small scale and casual assessments. On the other hand, CTT framework have been removed from large scale or high-stake assessments.

Item Response Theory

IRT framework is a logistic framework that provides continuous item parameters and person estimates. IRT framework assumes that the latent trait can be represented into continuum. Depending on the model, IRT framework first fits the response pattern into the model and gains item parameters. From the item parameters and response patterns, IRT framework estimates individuals' latent ability estimates. Ability estimates, or IRT scores are used to represent the ability of the attribute. IRT framework can be applied to unidimensional dataset, indicating that only one attribute is associated with the dataset. The multidimensional IRT (MIRT) is the extension of the unidimensional IRT. MIRT provides item parameters that represent attributes associated with each item. Furthermore, IRT (or MIRT) models can be dichotomous scores (i.e., scored by either 0 or 1) or polytomous scores (i.e., 4-point scale rating). Depending on the researcher's purpose, the scoring of items and IRT models will defer.

IRT model has a number of variations. Rasch, 2-parameter logistic (2PL), and 3-parameter logistic (3PL) IRT models are the most well-known four IRT models despite there are other variations of IRT models. These IRT models are adjacent to each other. The first model, Rasch model, is developed by Georg Rasch (1960). Rasch model constrains the discrimination parameter to be 1 and estimates the difficult parameters for individual items. Rasch model does not include guessing parameters in the model. 2PL model freely estimates the discrimination parameters for each item. 2PL model also estimates difficulty parameters but does not include guessing parameters in the model. The last model, 3PL model estimates discrimination, difficulty, and guessing parameters. In other words, Rasch model is nested in 2PL model and 2PL model is nested in 3PL model. A researcher should select the model based on the purpose and the model fit indices.

In order to compare competing models, Cohen and Cho (2016) suggested information criteria and chi-square test based on -2 log likelihood difference can be used. The chi-square test based on -2 log likelihood test uses log likelihood values from the adjacent models. Since the difference between chi-square values follows chi-square distribution, the absolute deviance between log likelihood values from the models will follow chi-square distribution. For competing information criteria, Kang and Cohen (2007) conducted simulation study compared Akaike information criterion (AIC), Bayesian information criterion (BIC), deviance information criterion (DIC), and cross validation log-likelihood (CVLL) by different condition. The study simulated different test length, sample size and ability distribution to compare those criteria. Kang and Cohen suggested using CVLL when the true data is generated from 3PL model.

In comparison to CTT framework, IRT framework has item invariance property. Item invariance indicates that the characteristics of item (i.e., item parameters) will not be affected by the sample's ability level. As discussed above, CTT's item properties differ significantly by the characteristics of the sample. Furthermore, due to the item invariance, the ability estimates are also dependent from the sample. Unlike CTT, IRT models will provide constant ability estimates regardless of test levels. This invariance property can also be applied in longitudinal IRT models.

To assess ability growth, IRT ability estimates can be treated as a score allocated to the individuals. The continuous ability estimates enable IRT outcomes to be applied parallel to CTT framework's raw scores and gain scores. To test randomization, researchers may use ANOVA between groups with the ability estimates and conditions. To assess differential growth in ANCOVA framework, a researcher may use prior ability estimate as a covariate. Or, a researcher may also use mixed ANOVA, with both the within subject factor and between subject factors. Choosing between options depends on the researcher's decision.

One shortcoming of IRT framework is its required sample size. According to Templin and Bradshaw (2013), compared to DCM model, Rasch IRT model requires larger item numbers to gain consistent estimation. Under the simulation study, the research concluded that IRT Rasch model required 15 items to obtain the reliability of .66 assessing one latent trait. On the other hand, a parallel DCM required 10 items to reach the reliability of .80. The study concluded that DCM constantly showed higher reliability in all simulation conditions. However, IRT framework provides continuous latent trait that has been adopted in many high-stake tests such as Scholastic Assessment Test (SAT), American College Testing (ACT), and Graduate Record Examinations (GRE).

In total, IRT is a confirmatory factor analysis with categorical responses. IRT estimates continuous item parameters and ability estimates from either dichotomous or polytomous responses patterns. IRT has invariance characteristics that provides constant item parameters and ability estimates regardless of sample and item characteristics. Although IRT requires a sufficient number of items and sample size, the practical aspect of IRT has led IRT to be widely used in many large scale, high-stake assessments.

Diagnostic Classification Model

According to Rupp and Templin (2011), DCM framework is another logistic framework used to provide diagnosis and diagnostic feedback. DCM framework is a constrained and confirmatory latent class models. Item parameters, typically the first order main effect parameters, are constrained to be positive in order to ensure that having an attribute will increase the probability of getting to answer the item correctly. Moreover, participant's latent classes are known prior to the estimation of the model. In other words, researchers first set the number of latent classes of each attribute prior to the DCM model estimation. Lastly, Q matrix is developed

prior to the model estimation. Q matrix is a matrix that consists of items matched with parallel attributes. In educational assessment, the Q matrix defines the link between items and attributes. Q matrix is a crucial part of the DCM framework.

Similar to IRT framework having many different variations, DCM was also developed by many scholars with different variations. The different sub-types of DCM models were developed by constraining the main effects, interactions, and parameter restrictions. Deterministic inputs, noisy 'and' gate (DINA) was introduced by Haertel (1989), Junker and Sijtsma (1999) as a non-compensatory diagnostic model. The DINA model restricts the main effects to zero and the main effects to be positive. DINA also sets parameter restrictions across attributes. As a result, the examinees in the DINA model will have an increased probability of getting the item correct only if they have mastered all the attributes associated with the item. On the other hand, having only partial proficiency will not increase the probability of answering the item correctly. Deterministic inputs, noisy 'or' gate (DINO) was introduced by Templin and Henson in 2006. DINO is a compensatory diagnostic model that constrains the main effect parameters to be positive and equal. DINO also constrains interaction parameters to be negative value of the main effect estimates. As a result, DINO provides equal probability for getting the times correct if the examinees have at least one of the attributes. General diagnostic model (GDM) was developed by von Davier (2005). Log-linear cognitive diagnosis model (LCDM) was developed in 2009 (Henson, Templin, & Willse). Another general diagnostic model generalized DINA model (G-DINA) was developed by de la Torre (2011). Theoretically, LCDM is the more general form among competing DCM models.

DCM framework developed the number of fit statistics to determine the absolute fit and the relative fit of the model. For absolute fit statistics, mean absolute fit deviation (MADcor),

standardized mean square root of squared residuals (SRMSR), mean absolute deviation residual covariance multiplied by 100 (100MADRESIDCOV), and root mean squared error of approximation (RMSEA) are used to determine the absolute model fit. On the other hand, likelihood ratio test based on -2 loglikelihood can be conducted to compare the relative fit between nested models. For the non-nested models, DCM also uses AIC, BIC, and sample-size-adjusted BIC (SABIC) criteria as a relative fit measure to compare non-nested models. According to Sen and Bradshaw (2017), all three information criteria performed poorly when the item quality is medium and further suggested the patterns of information criteria on various simulation conditions. The authors further argue that information criteria cannot be trusted to test the relative fit of the competing models.

DCM framework also has item invariance properties. Unlike CTT, the item parameters are not dependent on the sample's latent abilities. The item invariance property of DINA model was tested by de la Torre and Lee (2010). This study confirmed that with the simulated data, the DINA model provided low mean absolute bias and obtained constant DINA item parameters. Furthermore, de la Torre and Lee used real assessment data with fifteen items and compared DINA item parameters across different groups with different levels of ability (i.e., low, average and high ability groups). The study suggested that even with the real assessment data, DINA showed almost perfect item invariance property across different ability groups. The item invariance property LCDM model was assessed by Bradshaw and Madison (2015). The study conducted a simulation study by adjusting the base rate, test difficulty and sample size in the first part of the research. Bradshaw and Madison suggested that the item parameter bias was low except for having less than 1000 samples or having an easy test with low base rate dataset. The study further suggested that the item invariance property of LCDM was strong when the sample

size was sufficient. However, with the smaller sample size, the median absolute bias for the interaction parameters varied the most. Therefore, this study concluded that the LCDM model possesses quality item invariance property with a sufficient number of sample sizes.

For DCM framework to estimate the ability growth, TDCM framework can be used to quantify educational growth and also significant effect of the intervention. Madison and Bradshaw provided TDCM that comprehended LTA and LCDM to evaluate the intervention effect in mathematics assessment with four attributes. The attributes were ratios and proportional relationships, measurement and data, number of systems (fractions), and geometry (graphing). The study suggested that the two attributes, ratios and proportional relationships and geometry (graphing) showed differential growth while other attributes did not show significant differences between conditions.

In total, DCM is a confirmatory and constrained latent class framework that provides diagnoses and diagnostic feedback. DCM provides categorical variables as an ability estimate, continuous variable as an item parameters with various parameter constraints. DCM has item invariance characteristics that provides more constant analyses in ability estimates and item parameters compared to CTT.

CHAPTER 3

The Current Study

Original Assessment

The same assessment was used for both the pretest and the posttest. The test was originally developed to assess students' mathematical knowledge in ratios and proportional knowledge. Ratios and proportional knowledge are math standards for sixth grade and seventh grade. According to Copur-Gencturk and colleagues, (2023) student test was developed by adopting items from the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), two Common Core-aligned national assessments, Partnership for Assessment of Readiness for College and Careers (PARCC), Smarter Balanced. The paper also included relevant items from previous research (Fisher, 1988; Van Dooren et al., 2005).

The original assessment consists of a total of eleven items. nine items were multiple choice items, and two items were multiple selection items. Multiple choice items required students to select only one option. Multiple selection items required students to select all the appropriate options. Furthermore, all multiple selection items have instructions in the item "Select ALL the statements that ~", for the clear instructions. The word 'ALL' is bolded and capitalized for easy recognition. Since the items are imported from various sources, the number of options varies. For the multiple choice items, one item had four options, five items had five options and three items had six options. For the multiple selection items, one item had five options while the other item had seven options. The multiple selection item with five options had

three options to have the answer correct. The second multiple selection item with seven options had two options to have the answer correct. A pseudo multiple choice item with the answer key is illustrated in Figure 1.

Students' assessments were scored dichotomously. For multiple choice items, students received zero if they had the wrong answer but received one if they had the correct answer. For multiple selection items, students received one only if their response contained only correct options. For example, if a student selected correct options with the last 'I don't know' option, the student will receive zero score for the item. For another example, if another student selected only a partial set of the correct options, the student would also receive zero score for the item. Blank answers were scored as zero for both multiple choice items and multiple selection items.

The number of options included the last option 'I don't know' option placed at the very end of each item. Before the pretest and posttest, students were notified that students should select 'I don't know' option if they do not know the answer. Students were also able to submit the answers multiple times via paper and Qualtrics, but only their first response was used in the research.

The Common Core State Standards (CCSS) defined ratios and proportional relationships knowledge for sixth grade and seventh grade students (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). For the sixth grade students' ratios and proportional relationships standards (i.e., 6.RP), CCSS defined three sub-attributes. The first attribute is understanding the concept of ratios and proportions. The second attribute is understanding the concept of the unit rate (i.e., $a{:}\,b$ and a/b). The third attribute is using the knowledge to solve real-world mathematical problems. For the seventh grade students' ratios and proportional relationships standards (i.e., 7.RP), CCSS defined three sub-attributes. The first

attribute is computing unit rates. The second attribute is recognizing and representing proportional relationships between quantities. The third attribute is using ratios and proportional relationships to solve multi-step real world problems.

Although sixth grade and seventh grade content standards did not match exactly, teachers used the same curriculum, in-class materials and instruction process for both sixth grade and seventh grade students. Some teachers who taught both sixth grade and seventh grade students used the same curriculum even if the grade level differed. This indicates that students regardless of grade level received the same instruction in ratios and proportional knowledge. Moreover, sixth grade students and seventh grade students received identical pretests and posttests to assess their mathematical knowledge in ratios and proportional knowledge.

The total number of students was 2147. Of the 2147 total population, 1944 students took the pretest, and 1947 students took the posttest. Among the 1944 students who had the pretest, 1124 students were randomly assigned in the treatment group while 820 students were randomly assigned in the control group. Among the 1947 students who had the posttest, 1112 students were in the treatment group while 835 students were in the control group. Only 1727 students completed both the pretest and the posttest. Among the 1727 students, 1010 students were in the treatment group while 717 students were in the control group.

Partial Data Selection

Partial Item Selection

Although the original assessment consists of eleven items and 2147 number of participants. This research decided to exclude a specific item, item two, from the assessment. The pseudo item two is described in Figure 2. Item two is a multiple selection item that had five options including the ‘I don’t know’ option placed at the end of the item. Item two had five

possible options, and students had to select three correct options out of five possible options. Item three, which is another multiple selection item, had a total of seven options. Item three also had 'I don't know' option at the end of the item. Despite students having to select two correct options out of seven possible options, item statistics obtained from CTT, IRT, and DCM frameworks indicate that item three is a typical and acceptable item.

The main rationale for removing item two is that item two is an atypically difficult item that produced minimum information about students' ability estimates. To begin with, CTT analysis indicated that only 1 percent of the students on pretest (i.e., after instruction) answered item two correctly. The item descriptive statistics for the original posttest assessment items are described in Table 1. Table 1 presents the proportion of the correct answers to each item on the test.

Typically, items with percentage correct of 30 or less indicate that the item is too difficult for the average level of the sample. Items with percentage correct of 80 or more indicate that the item is too easy for the average level of the sample. Items between 30 to 80 indicate the item has desired difficulty level for the average level of the sample. The items with highest percentage correct, items eight and seven have percentage correct of 78 and 77. On the other hand, the items with second to third lowest percentage correct, items ten and three, have percentage correct of 36 and 39. Without item two, the mean difficulty for the ten-item assessment is .60 which is a typical difficulty. This indicates that from CTT framework, item two is an atypical item with too high difficulty.

Coefficient alpha, which is a lower-bound for the internal consistency measure is also used to justify the removal of item two. Coefficient alpha if item deleted for item two increased, which indicates that removing item two yielded higher internal consistency. Higher internal

consistency is a desired characteristic of a reliable assessment. The coefficient alpha for the original eleven-item assessment is .747, which is an acceptable reliability. However, if item two is removed the coefficient alpha for the ten-item test is increased to .757, indicating that assessment including item two is less reliable. Furthermore, item-rest correlation indicating the biserial correlation between having the specific item correct with the total score is used. Except for item two, all 10 items had positive item-rest correlation. On the other hand, item two had item-rest correlation of -.002, which is close to 0. This item-rest correlation for item two indicates that item two score is not significantly correlated to the total score, indicating that removing item two does not significantly deteriorate the information provided by the assessment. The item-rest correlation for the original posttest is described in Table 2. Coefficient alpha if item deleted for the original posttest is also described in Table 2.

From IRT perspective, item two is also too difficult for the test takers. To estimate IRT model from the full test, the R package "mirt" written in R 4.1.3 (Chalmers, 2012) was used to first determine the best fitting model. There are a number of IRT models commonly used to assess item statistics and ability estimates. Since 2PL model is nested in 3PL model, the statistical comparison between two models is conducted by chi-square test using -2 log likelihood value differences from two models suggested by Cohen and Cho (2016). From the original eleven-item assessment, 3PL model was significantly better than 2PL model ($\chi^2(df = 10) = 211.686, p < .001$). The -2 log likelihood values are presented in Table 4. After selecting the model, the three item parameters are estimated. The item parameters for the original test are described in Table 5.

In Table 5, the difficulty parameter for item two is 112.302, which is not a typical number for log-linear models. Furthermore, the discrimination for item two is 0.04, indicating

that item two is almost not discriminating. By considering the very high difficulty parameter and very low discrimination parameters, this study concluded that item two is removeable from the eleven-item assessment in IRT framework. Moreover, the item statistics obtained from the original eleven-item assessment and item statistics obtained from the reduced ten-item assessment showed very high correlation. The item parameters for the reduced, ten-item test are described in Table 6. Even after removing item two, discrimination and difficulty parameters showed above .999 correlation. Guessing parameters showed .998 correlation. Item statistics not significantly impacted by item two indicate that removing item two does not have a significant impact in item statistics.

By using DCM framework, DCM analysis indicates that item two is too difficult for the test takers. To estimate TDCM from the full test, Mplus version 8.4 (Muthén, L.K. & Muthén, B.O., 2017), was used for the analysis. The detailed procedure for analyzing TDCM model with pseudo Q-matrix is described in the next chapter. DCM fits data confirmatory latent class with the Q matrix. DCM with adjunct numbers of classes is a nested model. To compare DCM with adjunct numbers of class, an adjusted chi-square test provided by Mplus is suggested (Muthén, L.K. & Muthén, B.O., 2017). By comparing the 4-class model to 3-class model, 4-class model was significant $\chi^2(df = 18) = 247.237, p < .001$, but the 4-class model had spurious class that had little practical implication of the latent class. By comparing the 3-class model to the 2-class model, the 3-class model was significant $\chi^2(df = 16) = 825.690, p < .001$, without having spurious class. The 3-class model showed sound interpretation results. The -2 log likelihood values and adjustment numbers provided by Mplus are illustrated in Table 7.

From the 3-class TDCM model, the item two parameters also indicate extreme difficulty. The item parameters obtained from the eleven-item assessment are presented in Table 8. Item

two's parameters indicate that the low class and medium class both had about 1 percent chance of getting the item two correct while high class had 2 percent chance of getting the item two correct. This also indicates that item two is too difficult and also non-discriminating. Similarly, the item statistics obtained from the original eleven-item assessment and item statistics obtained from the reduced ten-item assessment showed very high correlation even in DCM framework. The item parameters for the reduced, ten-item test are described in Table 9. All 3 parameters including the main effect, λ_1 and λ_2 indicate that parameters had correlation above .999. Even after removing item two, DCM item estimates not significantly impacted by item two indicate that removing item two has not significant impact in item statistics.

Partial Participant Selection

Moreover, this research also decided to only use 1727 students who completed both the pretest and the posttest. This decision was due to using ANCOVA for CTT and IRT analysis. Since students possess different mathematical knowledge before instruction which may impact their further learning (i.e., posttest raw scores or ability estimates), pretest raw scores and pretest ability estimates are used as a covariate to assess intervention effect on educational growth. As a result, students with missing either pretest or posttest will be removed while assessing the intervention effect because ANCOVA requires both pretest and posttest results. Therefore, this research chose to have consistent number of participants from the start to the end. All the analysis in this study is based on 1727 participants (717 from the treatment group and 1010 from the control group).

In conclusion, this study will use the partial dataset that consists of 10 items (i.e., by removing item two which is too difficult) from the pretest and posttest. Furthermore, this study

used the partial participants of 1727 students who completed both the pretest and posttest and removed other students who did not complete either pretest or posttest.

Methods

From the original data, teachers were randomly assigned to the BAU group or PD group. According to Copur-Gencturk and colleagues (2023), teachers in the BAU group did not receive any treatment while the teachers in the PD group received online professional development (OPD). OPD programs are developed to provide interactive and personalized computer-based teacher PD program. Before the instruction, teachers were assessed in two tests. The first test assessed the teachers' content knowledge measures. The items were imported from the Learning Mathematics for Teaching (LMT) Project (Hill et al., 2004). The first measure consists of 22 multiple choice items. Teachers' responses were scored dichotomously. The second test assessed the teachers' pedagogical content knowledge. The second test was imported from the prior item pool developed by Hill and Kersting and colleagues (Hill et al., 2004; Kersting et al., 2010). The second measure included open-text responses with multiple choice items. The teacher's responses were scored on a 4-point scale. The open-text items were graded by two raters with Cohen's kappa statistics of .81, indicating strong agreement between ratings. The ability estimates of the teachers' content knowledge and pedagogical content knowledge were calculated by using generalized partial credit model (GPCM). However, this study's goal is to compare three major frameworks without hierarchy. As a result, the hierarchy of the data (i.e., students are nested in teachers) is simplified and students were treated as if they were in treatment or control conditions.

Students received the same control group or treatment group coding from their teachers before the pretest. Since some areas and schools were not allowed to collect other than math

assessment data, students were first allocated with a unique identification to match their pretest with the posttest results. Students first completed the pretest to assess prior knowledge in ratios and proportional knowledge. After the pretest, students had instruction from teachers who were either in BAU or PD condition. After the instruction, students were asked again to take the posttest. Therefore, the researchers could get the data with students' identification numbers, pretest responses, posttest responses and teacher's identification numbers bonded. However, the hierarchy of the data structure is simplified for this research's goal.

After collecting the data, grading was conducted by the predefined rules. Students' responses were scored dichotomously for both multiple choice items and multiple selection items. For multiple selection items, students were required to have the exact match with the answer code to receive credits. After the grading procedure, the dataset with students' identification numbers, pretest response patterns, pretest raw scores, posttest response patterns and posttest raw scores were used for the analyses.

While the typical CTT uses the gain score (i.e., posttest raw score – pretest raw score) to assess educational growth, this research also implemented ANCOVA. ANCOVA is conducted with the pretest raw score as a covariate, posttest score as a dependent variable and treatment condition as a fixed factor. The ANOVA methodology can also be applied in IRT framework, enabling this study to compare the intervention effect in CTT and IRT. In IRT framework, researchers can use pretest ability estimate as a covariate, posttest ability estimate as a dependent variable and treatment condition as a fixed factor. By comparing CTT to IRT, CTT used the raw score while IRT used ability estimate obtained from IRT models. However, the ANCOVA approach cannot be soundly applied to DCM (or TDCM) methodology. According to Madison and Bradshaw (2018), Wald test and odds-ratio are typical statistics to assess the intervention

effect in DCM framework. This research will present the odds-ratio to assess intervention effect in DCM framework.

CHAPTER 4

Analyses Results and Interpretations

Reliability Analysis

First, the reliability analysis was conducted to assess the quality of the ten-item partial assessment. As discussed in the last chapter, including item two decreased the coefficient alpha statistics. Also, other rationale for removing item two had been analyzed. Coefficient alpha, which is a lower bound measure for internal statistics has been measured by statistical software IBM SPSS (Version 28). Students' posttest response pattern was used to estimate coefficient alpha. Coefficient alpha for the ten-item test was .757, also indicating acceptable internal consistency.

Furthermore, the item-rest correlation between items and the posttest raw scores were acceptable. Typically, item-rest correlation above .3 is accepted. It indicates that all items have positive correlation with the total score. Moreover, none of the items had higher coefficient alpha if item deleted compared to the coefficient alpha for the ten-item assessment. If an item has higher coefficient alpha if item deleted than coefficient alpha, this indicates that removing the item will increase the internal consistency of the assessment. For the partial, ten-item assessment, removing any items will decrease the internal consistency of the assessment, indicating that all items are positively affecting the reliability of the assessment.

In conclusion, all three criteria indicate that the reliability of the assessment is acceptable. The coefficient alpha of the assessment is .757, which is in the acceptable range. Second, the item-rest correlations of all ten items are positive, indicating that items in the assessment are

reliable. Third, coefficient alpha if item deleted is lower than .757, indicating that removing any item will degrade the internal consistency of the measure.

Dimensionality Analysis

CTT and unidimensional IRT assume that the assessment is unidimensional. In other words, the number of possible factor(s) associated with the assessment is one. To determine dimensionality, an exploratory factor analysis (EFA) was conducted by using IBM SPSS Statistics (Version 28). SPSS is a widely used software for various statistical analysis. EFA was conducted without rotation.

To determine the number of factors in EFA, there are various criteria for obtaining the number of factors. The first criterion is Reckase's criterion (Reckase, 1979) by checking the percentage of variance explained. Reckase's criterion suggests that if the percentage of variance explained exceeds 20 percent, the dataset can be interpreted as having only one factor. The posttest assessment data has the percentage of variance explained by the first factor over 32 percent, suggesting that the dataset passed the Reckase's criterion. Eigenvalues, percentage of variance explained, and cumulative percentage of variance explained by the given factor structure are illustrated in Table 10. The second criterion is looking at the scree plot of eigenvalues. The scree plot for the posttest assessment data is illustrated in Figure 3. The researchers search for the specific point in the scree plot where the eigenvalues drop significantly and start decreasing at a slower rate. The scree plot in Figure 3 indicates that the eigenvalues decrease largely in the second factor. It is an indication that the dataset is unidimensional.

The two commonly used criteria in EFA suggest that the posttest response data is unidimensional. As a result, this research will assume that the assessment is unidimensional. From CTT perspective, unidimensionality indicates that all raw scores consist of the one true

score on ratios and proportional knowledge and error. From IRT perspective, dimensionality analysis provides rationale for using unidimensional IRT models instead of multidimensional IRT models.

This assessment was not built under DCM framework and does not have a priori Q matrix. Therefore, this research supposed a pseudo Q matrix for the analysis. Although not typical, this research used a pseudo Q-matrix that consists of only one attribute full of ones. On the other hand, a typical Q matrix will have multiple attributes and indicate which item is associated with specific attributes. In other words, every item in this assessment only includes one attribute (i.e., ratios and proportional knowledge). The pseudo Q-matrix is described in Table 11. With the pseudo Q-matrix, DCM analysis will provide classifications (i.e., 2-class, 3-class, or larger) based on only one attribute. Since the pseudo Q matrix is unidimensional, there was no interaction between attributes.

Classical Test Theory Framework

CTT Model Estimation

CTT framework does not typically transform raw responses into different variables. Instead, CTT assumes that the true score is equal to the raw score with errors. CTT also assumes that the mean of error is equal to zero. Therefore, CTT framework directly uses the raw score for the model. Dichotomous variables obtained from students' responses with pre-defined scoring rule can be used for the analysis without altercations. Instead, IRT fits the response data with the competing IRT models (i.e., 2PL and 3PL) while DCM fits the response data with the Q matrix to the model (i.e., DINA or full LCDM). As a result, there is no need to do further model estimation in CTT framework.

Without the model estimation process, CTT framework can be directly applied to obtain the raw scores. Students' pretest raw scores can be obtained by adding all sub-scores obtained from ten pretest items. For example, a student who scored items one, three and eight correct on the pretest will have three ones and seven zeros coded. The student will have a pretest raw score of three. Similarly, students' posttest raw scores can be obtained by adding all sub-scores obtained from ten posttest items. For another example, if a student mentioned above scored items one, three, five, eight and nine correct on the posttest will have five ones and five zeros coded. The student will have a posttest raw score of five. Gain score, which is another variable to assess the educational growth in CTT framework, can be obtained by subtracting pretest raw score from posttest raw score. From the example above, the student's gain score will be two (i.e., subtracting three, the pretest score, from five, the posttest score).

From CTT framework, there are two typical item parameters. The first item parameter is the difficulty parameter. The difficulty parameter in CTT is equal to the ratio of sample had the correct answer for the item. Therefore, the minimum value for the item difficulty in CTT is equal to 0 and the maximum value is 1. The item with higher difficulty parameter means that the item is relatively easier than another item with lower difficulty parameter. The difficulty parameter below .30, indicating less than 30% of the sample answering the item correct, indicates that the item is too difficult for the sample. On the other hand, the difficulty parameter over .80 indicates that the item is too easy for the sample. The item difficulty parameter for the posttest is described in Table 1. From Table 1, this research concluded that the overall difficulty of the posttest was reasonable.

The second item parameter used in CTT framework is the discrimination parameter. From the total group, CTT framework splits the group into halves. The discrimination parameter

for each item can be obtained by subtracting the difficulty parameter from the lower performing group from the difficulty parameter from the higher performing group. The item with higher discrimination parameter indicates that the higher performing group has more chance to answer correctly. On the other hand, the item with lower discrimination parameter indicates that there is little difference between higher and lower performing group to answer correctly. From CTT perspective, having many low discriminating items is not a desired characteristic of an assessment. The item discrimination parameter for the posttest is described in Table 12. From Table 12, this research concluded that the overall discrimination of the posttest was reasonable.

Figure 4 illustrates the distribution of posttest raw scores. The histogram indicates that the minimum score is 0, indicating the students who did not answer all ten items correctly. On the other hand, the maximum score is ten, indicating the students who answered all ten items correctly. The distribution of the posttest raw scores is negatively skewed. The skewness statistic obtained from IBM SPSS (Version 28) is -0.364 with the standard error of 0.059. This indicates that the posttest scores are significantly negatively skewed. However, this is a typical distribution in assessment result since some students could not get higher score than the number of items. In other words, students have their posttest score bounded by ten, which is the maximum possible score that students can obtain. Therefore, this research will use parametric methods for CTT, IRT, and DCM analysis.

CTT Intervention Effect Analysis

For CTT analysis, an ANOVA was first conducted to compare pretest raw scores for students in the BAU condition and students in the treatment condition. A prior ANOVA was conducted to determine the randomization process of the dataset. If the ANOVA on the pretest raw scores is not significant, this indicates that the randomization process is successful. On the

other hand, if the ANOVA result on the pretest raw scores is significant, this indicates that the randomization process was not successful. If the randomization is not successful, the researchers should go back to the recruiting and data collection process and search for possible biases that affected the randomization process. Table 13 presents means, and standard deviations of ability estimates of pretest and posttest for each condition.

This ANOVA result was not statistically significant ($F(1, 1725) = 2.973, p = .085, \eta^2 = .002$). This indicates that there was no significant difference between the BAU and treatment groups on the pretest, indicating a good randomization process. The ANOVA results are presented in Table 14. Although there is a pretest raw score difference by group illustrated in Figure 5, ANOVA result suggests that the two groups do not possess significant ability differences in ratios and proportional knowledge.

Next, an ANCOVA in which pretest raw score was used as a covariate, was conducted to investigate whether the two groups differed on posttest ability estimates. The effect of the pretest was significant, ($F(1, 1724) = 1376.812, p < .001, \eta^2 = .444$), indicating that student's posttest raw score was significantly influenced by their pretest raw score. The main effect of treatment was also significant, ($F(1, 1724) = 12.735, p < .001, \eta^2 = .007$) indicating that there were significant differences between the BAU and treatment groups on the adjusted posttest raw score while controlling for their pretest raw score. This indicates that the PD instruction taken by the teachers significantly improved their students' math knowledge in ratios and proportions. Although the partial η^2 for the treatment is small, this is due to that all students regardless of condition showed large educational growth. The pretest raw scores and posttest raw scores by group are described in Figure 5. The ANCOVA results are presented in Table 15.

Additionally, a gain score approach can also be conducted in CTT framework. Although gain score approach cannot be directly transformed into other frameworks, it is a widely used methodology in CTT framework to assess the intervention effect. The ANOVA is applied to the gain score with the treatment variable as a fixed factor. This ANOVA result was statistically significant ($F(1, 1725) = 15.650, p < .001, \eta^2 = .009$), indicating that there was significant difference between the BAU and treatment groups on the gain scores. This is an indication the intervention was significant. The ANOVA results on gain score are presented in Table 16.

Item Response Theory Framework

IRT Model Estimation

An item response theory model was used to estimate the best fitting model and item parameters. In order to determine the best fitting model, this research first compared three item response theory models, Rasch, 2-PL, and 3-PL models. The model comparison was conducted using the chi square test based on the difference between -2 log-likelihood estimates by Cohen and Cho (2016).

IRT framework fits the response pattern to the desired model. From this study, there are two possible response patterns for use. The first response pattern is the students' pretest responses. The second response pattern is the students' posttest responses. In this study, students' posttest response data were used to determine model fit and to estimate item and ability parameters. This is due to the fact that posttest responses had more variability. In other words, students before instruction will have less variability because they had learned less. To sum up, this study used students' posttest response data to compare competing IRT models.

The R package "mirt" written in R 4.1.3 (Chalmers, 2012) was used to estimate each IRT model. The R package "mirt" is a reliable package used in a variety of research. After

determining the best fitting models, item parameter estimates and students' posttest ability estimates are also obtained. Furthermore, this research fixed posttest item parameters to obtain students ability estimates for pretest. As a result, IRT model estimation process suggested the best fitting model, IRT item parameters, students' pretest ability estimates and students' posttest ability estimates as the outcomes.

Different IRT models estimate different numbers of item parameters. The three item parameters are discrimination (*a*), difficulty (*b*), and guessing (*c*). The discrimination parameter represents how well the item discriminates against the applicants. An item with a higher discrimination parameter (i.e., 1.5) will have more difference of probability of answering the item correct for two individuals with different ability estimates than another item with lower discrimination parameter (i.e., 0.1). A typical discrimination parameter ranges from 0.5 to 2. The second parameter, the difficulty parameter, is also called the location parameter. The difficulty parameter determines the item location, which maximizes the slope of the item characteristic curve (ICC). In other words, if the person's ability estimate is equal to the difficulty parameter, the person will have .5 chance of getting the item correct. Typically, having too high a difficulty parameter (i.e., 3.0) or too low a difficulty parameter (i.e., -4.0) will not provide sufficient information about the ability estimates. A typical difficulty parameter ranges from -2 to 2. The third parameter is the guessing parameter. The guessing parameter is the probability of answering the item correctly even if the participant is located at the negative infinity of the continuous ability distribution. Since the guessing parameter represents the chance of getting the item correct, the guessing parameter ranges from 0 to 1. Typically, some multiple choice items may have guessing parameters, but the guessing parameters do not always match the inverse of

the number of choices. The options may have unintended clues, or they may have different attractiveness for test takers with different ability levels.

Rasch model assumes discrimination parameter to be fixed at 1 and guessing parameters to be fixed at 0. This effectively makes Rasch model equivalent to a 1PL model. 2PL model estimates the difficulty and discrimination parameters and assumes the guessing parameters to be fixed at 0. 3PL estimates all three item parameters: difficulty, discrimination and guessing. The equation for 3PL IRT model is presented below:

$$p_i(\theta) = c_i + (1 - c_i)(\frac{1}{1 + e^{-a_i(\theta - b_i)}}) \quad (1)$$

Table 17 presents model fit results for these three models. Akaike information criterion (AIC), Baysian information criterion (BIC) and -2 log likelihood values are denoted in Table 17. A likelihood ratio test, calculated as the difference in -2 log likelihood values for two adjacent IRT models, was used to determine the best fitting IRT model. The comparison between Rasch model and the 2 PL model was significant ($\chi^2(df = 10) = 211.686, p < .001$). Both AIC and BIC suggested 2PL model over Rasch model. This suggested that the 2 PL fit the data better than Rasch model. Next, 2PL model and 3PL model were compared. 3PL model fit significantly better than 2PL ($\chi^2(df = 11) = 40.144, p < .001$). Although both AIC and BIC suggested 2PL model, this study chose -2 log likelihood test suggested by Kang and Cohen (2007). This suggested that 3PL model was the better fitting model for these data.

For 3PL model, the posttest item parameter estimates are given in Table 18. These same item parameters were used for estimating ability for both the pretest and posttest data. This use of the item parameter estimates from the posttest placed the ability estimates for the pretest on the same scale as the posttest.

From the item parameters described in Table 18, the item parameters obtained from the posttest response do not have extreme outliers. The minimum discrimination parameter is 0.974 and the maximum discrimination parameter is 2.621. There are 4 items (i.e., items 5, 6, 7 and 9) that exceed the discrimination parameter of 2. However, from the item descriptive statistics and 3PL model's difficulty parameters, the items with discrimination parameter above 2 have typical difficulty. This indicates that the four items are good items that are highly discriminating with reasonable difficulty levels. There is no item that had atypical difficulty parameters. For the guessing parameters, item one had a relatively high guessing parameter of .392. Item one has only four options with 'I don't know' option at the end, having three possible options for guessing. The guessing parameter of .392 is close to the inverse of the number of options other than the last option (i.e., 1/3 = .333). This indicates that although the guessing parameter of item one is quite high, it can be interpreted as a typical assessment item.

Since IRT model fitted the model with the posttest response patterns, the mean of the ability estimates of the posttest is 0. The minimum ability estimates were -1.790, indicating the ability estimates of the students who had no correct answers. The maximum ability estimates were 1.454, which are the ability estimates of students who had all items correct. As discussed in CTT sessions, the negatively skewed distribution of the assessment results is typical in educational research. The simple histogram of the ability estimates from the posttest is illustrated in Figure 6. The skewness statistics for posttest ability estimates is -0.364 with the standard error of .059. This indicates that the posttest ability estimates are negatively skewed. Parallel to the previous statements, having a negatively skewed distribution is typical in educational research. Therefore, this study also decided to use Gaussian statistics for IRT analysis.

IRT framework does not typically use reliability estimate to assess the consistency of the measure. Instead, IRT framework uses the information function. The information function measures the amount of information obtained from each item at each level of the ability estimates. The test information function is the total sum of the information function of each item. The test information function is used on the total assessment level while the information function is used on the individual item level. In IRT framework, researchers typically use the information function to assess the accuracy of the test scores at the different continuum of ability estimates.

IRT Intervention Effect Analysis

The students in the study were in classes that were taught by teachers who had been randomly assigned to one of two groups in the study. In one group, 717 students were taught by teachers who had taken the PD, and thus were classified as in the treatment condition. In the second group, 1010 students were taught by teachers who had not taken the PD, and thus were classified as in the BAU condition. Table 19 presents means, and standard deviations of ability estimates of pretest and posttest for each condition. The ability estimates from the BAU and treatment groups for pretest and posttest are illustrated in Figure 7 below.

An ANOVA was first conducted to compare pretest ability estimates for students in the BAU condition and students in the treatment condition. The ANOVA results are presented in Table 20. This ANOVA result was not statistically significant ($F(1, 1725) = 1.340, p = .247, \eta^2 = .001$) indicating that there was no significant difference between the BAU and treatment groups on the pretest. This is an indication that the randomization of teachers into the treatment and control groups was successful.

Next, an ANCOVA, in which pretest was used as a covariate, was conducted to investigate whether the two groups differed on posttest ability estimates. The ANCOVA results

are described in Table 21. The effect of the pretest was significant, $(F(1, 1724) = 1527.480, p < .001, \eta^2 = .470)$, indicating that student's posttest ability was significantly influenced by their pretest ability. The main effect of treatment was also significant, $(F(1, 1724) = 13.327, p < .001, \eta^2 = .008)$ indicating that there were significant differences between the BAU and treatment groups on the posttest while controlling for their pretest ability. This is an indication that the PD instruction taken by the teachers significantly improved their students' math knowledge in ratios and proportions.

Diagnostic Classification Model Framework

DCM Model Estimation

A diagnostic classification model was used to estimate the best fitting model and estimate item parameters. To determine the best fitting model, this paper first compared three DCM models: 2-class model, 3-class model, and 4-class model. Model comparison was conducted by using a chi-square likelihood test based on -2 log likelihood values with adjusted scaling factor suggested by Satorra and Bentler (Satorra & Bentler, 2010). Based on DCM framework's item invariance property suggested by Bradshaw and Madison (Bradshaw & Madison, 2016), item parameters were estimated from using both pretest and posttest response data. After determining the best fitting model, latent transitioning between groups were obtained. Odds-ratio were given to assess the significance of the intervention effect. However, the Wald test used by Bradshaw and Madison article (Bradshaw & Madison, 2018) was not conducted. As discussed above, the DCM model estimation was conducted by using the pseudo Q matrix in Table 11.

DCM item parameters and classifications were obtained by using Mplus version 8.4. Mplus is a reliable software used in DCM framework. The possible item parameters are

intercepts and main effects. Due to different difficulty levels obtained from CTT and IRT, the cPDCM, which constrains the main effect coefficients the same, model was not tested.

To compare 3 DCM models, model fit indices were compared between models. AIC, BIC and a chi-square test based on adjusted -2 log likelihood values was used to evaluate two adjacent DCM models recommended by Satorra and Bentler (Satorra & Bentler, 2010). 3-class model had both lower AIC and BIC compared to 2-class model. The comparison between the 2-class model and the 3-class model was significant ($\chi^2(df = 15) = 821.621, p < .001$). This suggested that the 3-class model fit the data better than the 2-class model. Next, the 3-class model and 4-class model were compared. Both AIC and BIC suggested that 4-class model fitted better than 3-class model. For the chi-square test, 4-class model fit significantly better than the 3-class model, ($\chi^2(df = 17) = 279.074, p < .001$). This suggested that the 4-class model was the best fitting model for this data. Table 21 presents the AIC, BIC, -2 log likelihood values, and adjustment values of the TDCM models.

Although the 4-class model was statistically significant, the classification result obtained from the 4-class model was not interpretable. The 4-class model proposed a spurious class with a class membership of about 5% of students. From the pretest group as a total, only 2 students (about 0.1%) were allocated to the third class. Furthermore, the 4-class model allocated 695 students (about 40.2%) as the third class for the posttest. However, even if the instruction is effective, having more than 40% of the students allocated at the second highest class is an atypical result. Furthermore, students it is highly unlikely for students to have sufficient knowledge to be allocated at the third or the fourth class since students' pretest was assessed before instruction.

The interpretation of the 4-class model is also problematic. In order to obtain a valid latent class, researchers should be able to provide the group names. Compared to 3-class classification that can be easily translated into low, medium and high latent classes, 4-class classification is hardly interpretable. Due to the impractical interpretation, this research chose to reject the 4-class model that provided non-interpretable results. In other words, although the 4-class model was statistically significant, the 4-class model was practically not significant. Table 23 presents classification results obtained from the 2-class to 4-class models.

After rejecting the 4-class model, this research also investigated the interpretability of classification results from the 3-class model. Table 23 presents classification results obtained from the 3-class model. The 3-class model did not have a spurious class with a class membership of less than 5% of students. The first latent class, which can be interpreted as a low, had 715 members (about 41.4%) from the pretest. The second latent class, which can be interpreted as medium, had 762 members (about 44.1%) from the pretest. The third latent class, which can be interpreted as high, had 250 members (about 14.5%) from the pretest. This may be due to the mixed sample consisting of sixth grade students with seventh grade students.

The posttest classification, which was obtained after instruction, was also interpretable. The first class had 444 members (about 25.7%) and the second class had 635 members (about 36.8%). The third class had 648 members (about 37.5). In total, students were allocated less to the lower class in the pretest results. Also, students were allocated more to the higher class in the posttest results. This is a typical students' ability before and after instruction. Students tend to have lower ability before instruction. On the other hand, students tend to have higher ability after instruction. However, not all students gain high, or proficient ability after instruction. The 3-class model provided the typical before instruction, and after instruction results observed in a typical

classroom. As a result, the 3-class model is both statistically significant and also practically significant. Although the 2-class model is not used, Table 22 presents classification results obtained from the 2-class model.

Mplus (Version 8.4) used for the analysis also provides item statistics with the 3-class models. In polytomous DCM framework, there are multiple item parameters obtained from the model. The first parameter is the intercept, and the second and third parameter are main effect coefficients, written as λs. The probability of each latent class member to have the item correct is obtained by those parameters.

As a result, DCM parameters are constrained to have higher probability of getting the item correct for the higher class members. In the 3-class model, the low group will have the lowest probability of getting each item (i.e., from item one to item eleven) correct. The high group will have the highest probability of getting each item correct. The medium group will have the probability of getting each item correct between the low group and the high group. As a result, the DCM in educational research constrains all λs to be positive. The intercept parameters can be either positive or negative but any value below -3 or above 3 is not a typical value. The equation for obtaining the probability of correct answer for trichotomous attributes (i.e., the 3-class model) is presented in equation below:

$$P(X_{ri} = 1|\alpha_r) = \frac{e^{\lambda_{i,0}+\lambda_{i,1}(\alpha^1)+\lambda_{i,2}(\alpha^2)}}{1 + e^{\lambda_{i,0}+\lambda_{i,1}(\alpha^1)+\lambda_{i,2}(\alpha^2)}} \quad (2)$$

Table 9 presents intercept and parameters obtained from the 3-class model. Table 24 presents probability of an item correct based on 3-class model. Since DCM constrains λs to be positive, students in the higher class shows higher probability for answering the items correctly. The intercept, λs, and probability of an item correct all indicate that the 3-class model provides interpretable result. In order to obtain discrimination parameters in DCM framework, researchers

can subtract the probability of correct answer from each group (i.e., high group's probability – low group's probability of the correct answer) for each item.

In DCM framework, researchers imported a number of reliability measures. DCM reliability measures assess the consistency of the classifications. The three typical reliability measures are test-retest reliability measures suggested by Templin and Bradshaw (2013). This reliability measure assumes no memory and proficiency effect between assessment practices and provides the tetrachoric correlation of the classifications. The second and the third reliability measures in DCM framework are point biserial and information gain reliability statistics suggested by Johnson and Sinharay (2020). Point biserial correlation provides reliability measures similar to CTT reliability. On the other hand, information gain provides reliability measures based on the entropy function.

Madison (2019) suggested polytomous TCDM (pTDCM) measurement based on the tetrachoric DCM reliability provided by Templin and Bradshaw (2013) extended for multiple time points. According to pTDCM reliability measure by Madison (2019), the longitudinal reliability of 3-class TDCM model was .917 indicating good consistency.

DCM Intervention Effect Analysis

In DCM framework, Wald test or odds-ratio is used to evaluate the differential growth between groups. Wald statistics are used to quantify the significance of the difference of the proportion of the latent transition. The 3-class TDCM classifications by condition are described in Table 25. The row indicates the pretest classification group, and the column indicates the posttest classification group. However, researchers can not directly use ANOVA or ANCOVA for the classification variables provided by TDCM framework because the classifications are categorical variables. Odds-ratio is defined by the ratio between proportion of growth divided by

proportion of non-growth (i.e., 1 − proportion of growth) from the treatment group to the control group. In other words, the ratio of growth in the PD group is divided by the ratio of growth in the BAU group. This research defined the growth as latent class transition to the higher class. Transition from low to medium, low to high, and medium to high is defined as growth. The proportion of each transition and odds-ratio between PD and BAU groups are presented in Table 26. The odds-ratio between treatment condition and control condition was 1.175, indicating that there is a very weak or no intervention effect.

Another odds-ratio was obtained based on the students who were classified to the low group before instruction. The odds-ratio between treatment condition and control condition for the limited sample was 1.465, indicating a moderate intervention effect. The next odds-ratio obtained from the medium group before instruction was .921, indicating no intervention effect. Odds-ratio from the high pretest group was not conducted since the high group could not move over to any higher class. The odds-ratios obtained from the total sample, low pretest group, and medium pretest group indicate that there is no significant intervention effect.

Results and Interpretations

This study's aim is to apply three statistical frameworks into real empirical data. The first part of the analysis was the reliability analysis. Generally, coefficient alpha is used to measure the internal consistency of the assessment. Since coefficient alpha is the lower bound, having acceptable value indicates that other reliability measures will produce at least equivalent results. On the other hand, IRT and DCM frameworks may present their own reliability measures. IRT framework uses information function instead of reliability. On the other hand, DCM framework uses test-retest reliability, point biserial reliability, and information gain reliability functions.

Therefore, it is difficult to directly compare reliability (or information function) obtained from three frameworks.

Moreover, three frameworks produced different item parameters and person parameters. Based on the person parameters, randomization analysis was conducted, CTT and IRT frameworks suggest good randomization. In CTT framework, ANOVA between BAU and treatment group on pretest score difference was not significant. In IRT framework, the pretest estimates instead of pretest raw scores were used, but the ANOVA results were also not significant. CTT and IRT frameworks used continuous person parameters to evaluate the randomization process, while DCM framework used categorical person parameters. Regardless of the type of person parameters and the values, CTT and IRT frameworks indicated that the randomization was successful.

Third, CTT and IRT frameworks suggested similar results for the intervention effect and the significance of prior knowledge. For the ANCOVA on posttest raw scores with the pretest raw scores as a covariate, the ANCOVA indicated that the pretest raw scores significantly inferenced the posttest scores. Similarly, ANCOVA on posttest ability estimates with the pretest ability estimates as a covariate produced similar results. The η^2 value for the covariates and treatment was also similar. This first indicates that the educational growth that PD produced relatively small, but still significant changes in posttest results. Furthermore, this also indicates that the students' knowledge in ratios and proportional knowledge before instruction has a huge impact on their ratios and proportional knowledge after instruction. This is parallel to previous findings on longitudinal growth models in mathematics. In DCM framework, direct ANCOVA cannot be directly applied since DCM provides categorical outcomes for the person parameters.

Fourth, only two frameworks may suggest significant instructional differences by instructional condition. In CTT and IRT frameworks, ANCOVA results were presented to show that the intervention effect is significant. In DCM frameworks, the odds-ratio was obtained to show that the students in the treatment condition transitioned similar to the students in the control condition. The three frameworks provided different estimates in item estimates and ability estimates, and also produced different results in assessing the intervention effect on students' growth in ratios and proportional knowledge.

There are possible reasons for DCM framework to suggest non-significant intervention effect. One possibility is that the effect size of the intervention was relatively small. As shown in CTT and IRT analyses, students regardless of condition improved greatly by instruction. Another reason may be that DCM framework does not provide growth if the students stayed in the original group. In the BAU condition, about 66% of students stayed in the original group. In the PD condition, 59% of students stayed in their pretest classification. Furthermore, since this study used atypical unidimensional Q matrix, further study should be conducted to examine the intervention effect while both BAU and PD provide huge growth.

Next, all three frameworks suggested general knowledge growth after the instruction. In other words, regardless of instruction type, students all developed their knowledge on ratios and proportions. The simplest way to find the overall knowledge growth is to see the descriptive statistics before and after instruction. Figure 5 indicates that regardless of instruction type, all students improved their knowledge by having higher average posttest raw scores compared to average pretest raw scores. Parallel to CTT's raw score approach, Figure 7 also suggests that regardless of instruction type, all students showed educational growth by having higher average posttest ability estimates compared to average pretest ability estimates. For DCM framework,

latent class transition suggested parallel results, indicating that students developed their knowledge in ratios and proportions after any type of instruction. TDCM suggests regardless of instruction type, 1070 students did not show any transition while only five students transitioned to the lower class (i.e., from high to low, from high to medium, and from medium to low). TDCM also suggests that 672 students transitioned to the higher class. The reason for TDCM to have many students to show no transition may be due to 273 students who already had high knowledge before instruction, and also stayed in the same group after the instruction. Out of 711 students who were classified into the low group on the pretest, 275 students transitioned to the higher groups. On the other hand, 758 students were originally classified into the medium group. Out of 758 students, 397 students actually transferred to the high group. The transition results are presented in Table 25. The latent class transition also suggests parallel results that regardless of the instruction types, students showed educational growth after instruction.

On the other hand, if the researcher's intention is to provide classifications, CTT and IRT frameworks cannot provide good results. CTT only provides raw score that does not provide enough information about the different cognitive process that different item requires. IRT can provide apt information, but the cut-off point for the classification is not clear. A researcher can set a norm-referenced guideline for the classification, but the norm-referenced classification cannot suggest the exact ability estimate required that indicates attribute mastery. For CTT and IRT person estimates, the researcher can be less sure if the cut-off point represents valid criterion-referenced ability difference. On the other hand, DCM provides the best classification since the framework itself is intended to provide classification. With the sufficient number of items and populations, DCM is the best framework for the classification.

Three frameworks produced information about individual students' ability growth, but the information was different. CTT provided the gain score, which can be interpreted as the individual's ability growth. However, the gain score is less reliable than other ability growth estimates. CTT does not provide any information about item difficulty. Even if the two individuals had the pretest, posttest and gain score, their latent ability and latent ability growth may differ since each item possess different characteristics. For example, a student with the same posttest score of nine by missing one easier item may have higher knowledge than another student who had the same posttest score by not knowing how to work with multiple steps applied process. IRT framework provides ability estimates which is similar to Z scores. Since the ability estimates are obtained by the response pattern and the item parameters, ability estimates in IRT framework can solve the problem that CTT has. From the example above, although two students have the same raw score, they will receive different ability estimates. In reality, the first student who had missed the easier item will get higher ability estimate than the later student. This result is more valid since the first student is more likely to have higher latent knowledge and the ability to solve applied multiple steps than the second student. The DCM framework provides classifications, but DCM framework does not provide enough information for the individual difference in same groups. For the example of the two students above, having 9 out of 10 items will classify two students in the high group, indicating that two students will be treated equally. However, in DCM frameworks, some students with same raw scores can be classified into different groups. It is similar to IRT frameworks that the student who scored the more difficult item will have a higher chance to be classified into higher latent class than the other student who scored the easier item.

All three frameworks produced unique information about item statistics. In CTT framework, difficulty parameters and discrimination parameters are provided. In IRT framework, 3PL model provided discrimination, difficulty and guessing parameters. The 3-class TDCM model provided the intercept and two λs that is the threshold for the cut-off line for the medium and the high class. From the three parameters DCM provided, probability of getting each item correct can be obtained. The Discrimination parameters can also be defined by the group difference between probability of answering items correctly.

All three frameworks provide discrimination parameters. CTT and DCM frameworks define discrimination as a difference in probability of answering the item correctly. On the other hand, IRT provides discrimination parameter based on the slope. CTT framework cannot guarantee if the students can be divided into only two groups. In this research, researchers have to set an arbitrary cut-off criterion to classify students into three groups only based on the raw scores. Even if the samples belong to two groups, CTT framework cannot guarantee the sample size for each group. This indicates that DCM frameworks' discrimination parameter is more reliable and practical compared to CTT framework's discrimination parameter. From IRT framework, items five, six, seven and nine have relatively high discrimination parameter above 2. This is not perfectly parallel to CTT discrimination parameters since item seven has discrimination parameter of .380 while the other three items have discrimination parameters higher than .450. This may be due to the fact that item seven has low difficulty so that even the lower group in CTT framework had more than 50 percent chance of getting item seven correctly. When comparing IRT and DCM framework, they provide relatively similar results. For items five, six, seven and nine, the items listed above have less than 33% chance for the low group to

answer the items correctly while the highest group have higher than 88% chance for the high group to answer the items correctly.

For the difficulty parameter, CTT provides an overall percentage of the correct answer while IRT and DCM provides log-linear parameters based on their models. While the difficulty parameter provided by CTT is the easiest for interpretation, the percentage of the correct answer is both inferenced by the discrimination and the difficulty of the item. For example, item five in the posttest possesses high item discrimination above .50, indicating high discrimination function of the item. For item five, the lower performing group has a difficulty parameter of .32 while the higher group has difficulty parameter of .82. This indicates that item five is a difficult item for the lower group and an easy item for the higher group. However, the total difficulty parameter is .57, which may be interpreted as a moderate difficulty item. Furthermore, the item parameter provided in CTT is dependent on the sample, which suggests that the item parameter is likely to change as the sample changes. As a result, CTT difficulty parameter itself cannot be interpreted directly and reliably.

On the other hand, difficulty parameters in IRT framework are typically ranged between -2 to 2. A negative difficulty parameter indicates that the item is relatively an easy item while a positive difficulty parameter indicates a harder item. From IRT analysis, 3PL indicates that items three and ten are the two hardest items in the assessment and items seven and eight are the two easiest items. This result is parallel to CTT difficulty parameters. In DCM frameworks, the model calculates the intercept parameter constrained λs that function as a threshold. In educational research, a typical intercept and λs range between -2 to 2. However, DCM does not provide direct difficulty parameter, but the researchers have to comprehensively look for intercept and λs estimated for each item. The lower the intercept parameters indicate harder

items while the higher intercept parameters indicate easier items. For this assessment, items three and ten have intercept parameters lower than -2. The first item has the highest intercept, but it does not indicate the item one has the highest probability of getting the answer correctly between the three groups.

For the guessing parameter, only 3PL IRT model provides the relevant parameters. In CTT framework, researchers may develop possible guessing probabilities based on the content experts' opinion. For example, some theories simply assume that the guessing parameter for each item is the inverse of each option available. However, the assessment used for this study has the 'I don't know' option that hampers this research to use the guessing parameter above. DCM framework does not provide any guessing parameters. Instead, DCM provides the probability of getting the item correct for the lowest group. However, this probability cannot be treated as the guessing parameter since the students classified as the lowest group will still have some latent ability in ratios and proportional knowledge. As discussed in the literature review session, there is a debate in IRT framework if the guessing parameters truly represent the true guessing process. Some researchers use other IRT models, such as Rasch or 2PL, that fix the guessing parameters to 0. However, if the researcher wants to obtain the guessing parameters for the assessment, only IRT framework provides relevant information.

CHAPTER 5

Discussions

Exploring Multidimensionality

An assessment or measure can be inferenced by multiple factors. In CTT, the most measures assume that the measures assess only one factor. If a researcher wants to assess multiple factors in CTT framework, the researcher should propose a number of assessments. Even though a number of assessments are presented, CTT framework cannot quantify the relevant information about the structure of the attributes. Attributes may have hierarchies, that indicates that an attribute must be obtained prior to obtaining another attribute. In mathematics, a common attribute with hierarchy is the relationship between factorization and second order equations. On the other hand, CTT framework has limitation in exploring the interaction between attributes. In other words, some items require multiple attributes to get the answer correct. However, CTT cannot provide theoretically sound, and practical testing guidelines to deal with interaction terms. As a result, CTT cannot provide ample information about multidimensionality and hierarchy of attributes for an assessment, that IRT and DCM frameworks can provide.

In IRT frameworks, multidimensional item response theory (MIRT) is developed to explore the multiple factor structure of the assessment. First, IRT framework can be either unidimensional or multidimensional by its' nature. In unidimensional IRT framework, the assessment fits item parameters for a single factor and also provides ability estimates for the single factor. On the other hand, MIRT framework fits item parameters for the number of factors

and also provides the matching number of ability estimates. As a result, MIRT framework will provide item parameters and ability estimates for multiple factors.

In DCM framework, DCM framework is intended to deal with complex structures of the attributes. According to Rupp and Templin (2008), DCM model itself is multidimensional and concerns interaction between attributes. The unidimensional DCM model that this study proposed will be discussed in the limitations of this study. On the other hand, exploring multiple attributes in one assessment has its cost in DCM framework. First, a valid and reliable Q matrix should be established by the content and psychometrics experts. By using full LCDM model, the higher-order interaction terms increase significantly. Even if the researchers want to use restricted DCM models (i.e., DINA or DINO), the increase in latent variables lead may cause other problems. If a researcher fits the model with increased latent attributes with the item numbers fixed, the reliability of the assessment may decrease. In other words, researchers have to increase the number of items when the researcher introduces more complex structure of the data in order to obtain similar reliability. Thus, it may degrade DCM framework's advantages of having higher reliability compared CTT and IRT.

Exploratory Q-matrix Analysis

As discussed above, DCM framework was originally developed to deal with multidimensionality of the data. Therefore, it is uncommon for DCM researchers to consider the unidimensional and uniform Q matrix. In this session, this research will conduct exploratory Q matrix analysis with three different latent attributes with hierarchies.

The CCSS defined three latent attributes for sixth grade students' mathematics knowledge in ratios and proportional knowledge. The first attribute is acknowledging the concept and the second attribute is to acknowledge the concept of the unit rate. The third

attribute is to use relevant knowledge in ratios and proportions to solve the real world problem. Since the prior attribute(s) should be mastered prior to the proceeding attribute, this paper used attribute structure with hierarchy without interaction terms. From the ten-item assessment, this study defined an exploratory Q matrix. All items required students to acknowledge that the items were related to the ratios and proportions. Therefore, this research assumed that all items required the first attribute. Second, the ten items required the concept of the unit rate to solve the problems. Similar to the first attribute, all ten items were assigned to be inferenced by the second attribute. The third attribute, real world problem solving, was given to the items that included real world problems. The exploratory pseudo Q matrix is described in Table 27.

Since this session's purpose is to provide further research questions, this study suggested an exploratory pseudo Q matrix in dichotomous classes. Since the attributes were constrained to have hierarchies, there are only four classifications. For the first class, students were classified to have mastered none of the attributes. For the second class, students were classified to have mastered only the first attribute. For the third and the fourth class, students either mastered the first two or all attributes. To estimate the DCM model and item parameters with the exploratory pseudo Q matrix, the R package "CDM" written in R 4.1.3 (George et al., 2016) was used. This exploratory process used the posttest response data.

Table 28 describes exploratory hierarchical diagnostic classification model (HDCM) classification results. Since the model constrained the hierarchy between attributes, there are four latent class profiles. 305 students were classified into the first latent class, indicating that 305 students possess no proficiency for three attributes. 480 students were classified into the second latent class, which may be interpreted as having only basic knowledge. For the third class, 243 students were classified for being proficient in both realizing the concept and understanding the

concept of the unit rate. 699 students were classified into the highest class that has overall proficiency in all three attributes.

The three correlations between attributes were .99, which is not typical for HDCM analysis. Although researchers typically expect high correlation between attributes with hierarchies, the correlation of .99 indicates that the attribute structure needs to be further analyzed. From the 1727 number of students for the posttest, HDCM model suggests 82%, 55% and 40% proficiency rates for the three attributes. Despite extremely high correlation, exploratory HDCM results produced interpretable classifications.

To suggest the hierarchy among attributes, DCM model without hierarchy should show the proportion of attribute profiles supporting possible hierarchy. LCDM model without restriction should have low proportion of attribute profiles. To compare the possible hierarchy of the dataset, LCDM model without interaction was conducted. The exploratory LCDM classification proportions are described in Table 29. However, the skill pattern probabilities failed to indicate possible hierarchy of the data. Failure to capture the hierarchy between the attributes may have a number of reasons. The exploratory pseudo Q matrix may not represent proper attributes since the exploratory pseudo Q matrix is not developed a priori to the research. Second, high multicollinearity between attributes may cause estimation problems since DCM typically uses clearly distinctive knowledge or skills as attributes. The third possible consideration is the grain size of the assessment. Considering constant ones allocated to the first attribute, changing the grain size may produce better results. The two exploratory models had different tetrachoric reliability measures suggested by Templin and Bradshaw (2013). The exploratory HDCM had tetrachoric reliability measures of .898, .894, and .852, suggesting good consistency. The reliability measure over .8 suggests good reliability, reliability measure

between .6 to .8 suggests acceptable reliability, and reliability measure below .6 suggests not acceptable reliability. The exploratory LCDM had tetrachoric reliability measures of .809, .719, and .418 suggesting not acceptable reliability for the third attributes.

HDCM item parameters are illustrated Table 30. All the intercept parameters and main effect parameters are typical except for the λ_1 of the first item and λ_3 of the of the item 7. The main effect of parameter zero indicates that there is no gain of the probability of getting the item correct by mastering the attribute. Furthermore, item 7's third main effect parameter is above 3.0, suggesting that further item level analysis should be conducted. Problems related to HDCM model indicate further adjustments to Q matrix structure.

Limitations

The first limitation of this study is that this study ignored the nested structure of the data. The first nested structure simplified in this study is that the treatment has been applied to the teachers, not students. Therefore, directly applying ANOVA and ANCOVA with only the students' ability estimates (i.e., raw scores, ability estimates, and latent classifications) may not produce proper information about the treatment effect. In the original research by Copur-Gencturk and colleagues (2023), there is another set of assessments that quantified the teachers' content knowledge and pedagogical knowledge. There is possible hierarchical structure in the data that with teachers' content and pedagogical knowledge differ by BAU and PD condition that caused the indirect intervention effect of students' mathematics knowledge in ratios and proportions. The second nested structure of the data is the students' grade. Although the teachers reported that they provided the same instruction and curriculum for sixth grade and seventh grade students, the average prior mathematics knowledge may differ by age since students have

been exposed to mathematics curriculum for different time periods. Therefore, the two nested structures should also be considered in further research.

The next limitation of this study is that this study applied DCM framework without predetermined Q matrix. Prior dimensionality analysis was conducted to justify the single factor structure of the data. However, DCM framework clearly requires a valid Q matrix defined by the experts. The exploratory study to assess multidimensionality used the pseudo Q matrix with three attributes, but the quality of the pseudo Q matrix is still questionable. Mathematics researchers, teachers and other experts' comprehensive decisions to confirm a valid Q matrix. Specifically, some multiple choice in this assessment was in the grey area between the second attribute and the third attribute. In other words, some multiple choice items were applied, but the cognitive process required to solve some multiple choice items did not require multiple cognitive processes. On the other hand, the exploratory DCM analysis also defined three attributes based on the CCSS standards. However, there is no guarantee that actual content experts and DCM experts would also agree with three attributes defined in CCSS standards. Also, hierarchical structures between attributes should be also discussed. In the exploratory pseudo Q matrix, this research assumed that the first attribute, acknowledging the ratio and proportion, should be mastered prior to the understanding of the unit rate. This study also assumed that the second attribute should be mastered prior to real world problem solving attribute. There should be further research conducted to explore the hierarchy of the data. Also, this study proposed that there is no interaction between attributes. This all indicates that the Q matrix defined for the analyses in this study possesses possible problems. This opens space for further research in content attributes and the relationship between attributes with multicollinearity and hierarchies

that will lead to the better understanding of the theoretical and practical implications of DCM frameworks in mathematics assessment.

The third limitation of this research is the generalizability of the sample and results. The original study by Copur-Gencturk and colleagues (2023) was permitted to gather students' assessment data with unique identification numbers only. Since the original data could not gather information about students' ethnicity, socioeconomic status (SES), and other characteristics, this study cannot assure that the 1727 samples can represent the US sixth and seventh grade students. Furthermore, even though IRT and DCM frameworks possess item invariance properties, the specific item parameters may change significantly if the sample used in this study is biased. Therefore, further study is required to assure the generalizability.

Further research should also consider not only test-level intervention effect, but also item-level intervention effect of PD. This study mainly focused on the overall ability, the ratios and proportional knowledge. However, PD may also have differential effects on different cognitive processes. Differential effects may take place on certain attributes. For example, some intervention may significantly develop cognitive processes that require higher depth of knowledge (DOK). In this study, another research may be conducted to see if PD significantly improved students' ability to solve the problems that require multi-step real world problems. To assess differential effect on different cognitive processes, the multidimensionality or the complex cognitive process required for the assessment should be further assessed by the content and psychometrics experts.

Despite limitations, this study proposed the simplified structure of the data to compare three different frameworks in mathematics assessment for sixth and seventh grade students. With

the analyses, implications and further suggestions of the study, further research to consider the complex structure of the data, attributes, items, and cognitive processes should be considered.

References

Bradshaw, L., & Madison, M. C. (2016). Invariance properties for general diagnostic classification models. *International Journal of Testing*, *16*(2), 99–118. https://doi.org/10.1080/15305058.2015.1107076

Campbell, P. S., & Malkus, N. (2011). The impact of elementary mathematics coaches on student achievement. *Elementary School Journal*, *111*(3), 430–454. https://doi.org/10.1086/657654

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the *R* Environment. *Journal of Statistical Software*, *48*(6). https://doi.org/10.18637/jss.v048.i06

Cohen, A. S., & Cho, S.-J. (2016). Information criteria. In W. J. van der Linden (Ed.), *Handbook of item response theory: Statistical tools* (pp. 363-378). Boca Raton, FL: Chapman & Hall/CRC.

Copur-Gencturk, Y., Li, J. X., Oh, S. M., Cohen, A. S., & Chandra, O. H. (2023). *The impact of an interactive, personalized computer-based teacher professional development program on student performance: A randomized controlled trial* [Manuscript submitted for publication]. Department of Education, University of Southern California.

De La Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179–199. https://doi.org/10.1007/s11336-011-9207-7

De La Torre, J., & Lee, Y. (2010). A note on the invariance of the DINA model parameters. *Journal of Educational Measurement*, *47*(1), 115–127. https://doi.org/10.1111/j.1745-3984.2009.00102.x

Fisher, L. (1988). Strategies used by secondary mathematics teachers to solve proportion problems. *Journal for Research in Mathematics Education*, *19*(2), 157–168. https://doi.org/10.5951/jresematheduc.19.2.0157

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The *R* package CDM for cognitive diagnosis models. *Journal of Statistical Software*, *74*(2). https://doi.org/10.18637/jss.v074.i02

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

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

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal*, *105*(1), 11–30. https://doi.org/10.1086/428763

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

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement*, *31*(4), 331–358. https://doi.org/10.1177/0146621606292213

Kersting, N. B., Givvin, K. B., Sotelo, F. L., & Stigler, J. W. (2010). Teachers' analyses of classroom video predict student learning of mathematics: Further explorations of a novel measure of teacher knowledge. *Journal of Teacher Education*, *61*(1–2), 172–181. https://doi.org/10.1177/0022487109347875

Lowrie, T., Logan, T., Harris, D. A., & Hegarty, M. (2018). The impact of an intervention program on students' spatial reasoning: Student engagement through mathematics-enhanced learning activities. *Cognitive Research: Principles and Implications*, *3*(1). https://doi.org/10.1186/s41235-018-0147-y

Madison, M. J. (2019). Reliably assessing growth with longitudinal diagnostic classification models. *Educational Measurement: Issues and Practice*, *38*(2), 68–78. https://doi.org/10.1111/emip.12243

Madison, M. C., & Bradshaw, L. (2018a). Assessing growth in a diagnostic classification model framework. *Psychometrika*, *83*(4), 963–990. https://doi.org/10.1007/s11336-018-9638-5

Madison, M. J., & Bradshaw, L. (2018b). Evaluating intervention effects in a diagnostic classification model framework. *Journal of Educational Measurement*, *55*(1), 32–51. https://doi.org/10.1111/jedm.12162

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *Social Science Research Network*. http://files.eric.ed.gov/fulltext/ED506058.pdf

Masters, J., De Kramer, R. M., O'Dwyer, L. M., Dash, S., & Russell, M. J. (2010). The effects of online professional development on fourth grade English language arts teachers' knowledge and instructional practices. *Journal of Educational Computing Research*, *43*(3), 355–375. https://doi.org/10.2190/ec.43.3.e

Muthén, L.K. & Muthén, B.O. (2017). Mplus version 8 user's guide. Los Angeles, CA: Muthén & Muthén

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common core state standards for mathematics.* Washington D.C.: Author. http://corestandards.org/

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* Nielsen & Lydiche.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*(3), 207–230. https://doi.org/10.3102/10769986004003207

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

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

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

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79*(2), 317–339. https://doi.org/10.1007/s11336-013-9362-0

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

Van Der Linden, W. J. (2017). *Handbook of Item Response Theory: Volume 2: Statistical Tools*. CRC Press.

Van Dooren, W., De Bock, D., Hessels, A., Janssens, D., & Verschaffel, L. (2005). Not everything is proportional: Effects of age and problem type on propensities for overgeneralization. *Cognition and Instruction*, *23*(1), 57–86. https://doi.org/10.1207/s1532690xci2301_3

Von Davier, M. (2005). A general diagnostic model applied to language testing data. *ETS Research Report Series*, *2005*(2), i–35. https://doi.org/10.1002/j.2333-8504.2005.tb01993.x

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C., Paunesku, D., Romero, C., Flint, K., Roberts, A., Trott, J., Iachan, R., Buontempo, J., Yang, S. M., De Carvalho, C. R. R., . . . Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, *573*(7774), 364–369. https://doi.org/10.1038/s41586-019-1466-y

Table 1

Item Descriptive Statistics of the Original Posttest

Item Number	Correct Answer (%)
Item 1	67
Item 2	1
Item 3	39
Item 4	61
Item 5	57
Item 6	66
Item 7	77
Item 8	78
Item 9	71
Item 10	36
Item 11	53

Table 2

Item-Rest Correlation and Coefficient Alpha if Deleted of the Original Posttest

Item Number	Item-Rest Correlation	Coefficient Alpha if Item Deleted
Item 1	.328	.737
Item 2	-.002	.757
Item 3	.323	.739
Item 4	.433	.723
Item 5	.415	.726
Item 6	.521	.710
Item 7	.493	.716
Item 8	.445	.722
Item 9	.524	.711
Item 10	.387	.730
Item 11	.336	.737

Note. Coefficient Alpha of the original posttest with 11 items was .747.

Table 3

Item-Rest Correlation and Coefficient Alpha if Deleted of the Posttest

Item Number	Item-rest Correlation	Coefficient Alpha if Item Deleted
Item 1	.321	.750
Item 3	.333	.749
Item 4	.431	.735
Item 5	.411	.738
Item 6	.525	.721
Item 7	.499	.727
Item 8	.451	.733
Item 9	.527	.722
Item 10	.387	.741
Item 11	.332	.750

Note. Coefficient Alpha of the posttest with 10 items was .757.

Table 4

-2 Log-likelihood of the Original Posttest IRT Models

IRT Model	-2 Log-likelihood
Rasch	19726.770
2PL	19515.084
3PL	19474.940

Table 5

3PL Item Parameters of the Original Posttest

Item Number	Discrimination	Difficulty	Guessing
Item 1	1.841	0.125	.393
Item 2	0.040	112.302	.001
Item 3	0.977	0.628	.034
Item 4	1.336	-0.444	.005
Item 5	2.359	0.192	.233
Item 6	2.009	-0.548	.000
Item 7	2.618	-0.746	.158
Item 8	1.794	-1.034	.059
Item 9	2.504	-0.578	.118
Item 10	1.618	0.789	.094
Item 11	0.950	0.004	.069

Table 6

3PL Item Parameters of the Posttest

Item Number	Discrimination	Difficulty	Guessing
Item 1	1.837	0.124	.392
Item 3	0.974	0.625	.033
Item 4	1.350	-0.425	.014
Item 5	2.357	0.192	.233
Item 6	2.016	-0.543	.003
Item 7	2.621	-0.744	.160
Item 8	1.768	-1.064	.038
Item 9	2.510	-0.575	.120
Item 10	1.620	0.790	.095
Item 11	0.959	0.022	.076

Table 7

TDCM Model Comparisons of the Original Assessment

DCM Model	-2 Log-likelihood	Scaling Correction Factor
2-Class	39734.834	1.140
3-Class	38941.15	1.070
4-Class	38727.83	1.007

Table 8

3-Class Item Parameters of the Original Assessment

Item Number	Intercept	λ_1	λ_2
Item 1	-0.374	0.741	1.814
Item 2	-5.026	0.033	0.852
Item 3	-2.281	0.884	1.515
Item 4	-1.089	1.387	1.620
Item 5	-1.098	0.801	2.377
Item 6	-1.470	2.122	2.479
Item 7	-0.777	2.730	1.959
Item 8	-0.625	1.952	2.792
Item 9	-1.124	2.806	1.673
Item 10	-2.027	1.084	1.467
Item 11	-0.884	1.145	0.812

Table 9

3-Class TDCM Item Parameters

Item Number	Intercept	λ_1	λ_2
Item 1	-0.374	0.737	1.815
Item 3	-2.280	0.878	1.519
Item 4	-1.089	1.383	1.618
Item 5	-1.098	0.798	2.366
Item 6	-1.47	2.116	2.486
Item 7	-0.779	2.731	1.938
Item 8	-0.625	1.947	2.804
Item 9	-1.126	2.807	1.661
Item 10	-2.028	1.084	1.463
Item 11	-0.885	1.145	0.809

Table 10

Eigenvalues, Percentage of Variance Explained and Cumulative Percentage

Number of Factors	Eigenvalues	% of Variance explained	Cumulative %
1	3.220	32.204	32.204
2	1.013	10.135	42.338
3	0.917	9.165	51.504
4	0.867	8.670	60.174
5	0.760	7.596	67.770
6	0.743	7.430	75.199
7	0.665	6.650	81.849
8	0.659	6.586	88.436
9	0.596	5.955	94.391
10	0.561	5.609	100.000

Table 11

Pseudo Q Matrix for the DCM Analysis

Item Number	Attribute
	Ratios and Proportional Knowledge
Item 1	1
Item 3	1
Item 4	1
Item 5	1
Item 6	1
Item 7	1
Item 8	1
Item 9	1
Item 10	1
Item 11	1

Table 12

Item Discrimination Parameters in CTT Analysis

Item Number	Correct Answer (%)		
	Lower Group	Upper Group	Item Discrimination
Item 1	46	89	.426
Item 3	20	59	.382
Item 4	37	85	.473
Item 5	32	82	.507
Item 6	40	92	.514
Item 7	58	96	.380
Item 8	61	96	.353
Item 9	49	95	.467
Item 10	14	58	.437
Item 11	34	73	.395

Note. The lower group has 864 students. The Upper group has 863 students.

Table 13

Descriptive Statistics of Pretest and Posttest Raw Scores by Condition

	Pretest		Posttest	
Conditions	*M*	*SD*	*M*	*SD*
BAU	4.828	2.529	5.939	2.697
Treatment	4.617	2.517	6.133	2.568

Note. BAU group's $N = 717$, Treatment group's $N = 1010$

Table 14

Means, Standard Deviations, and One-Way Analysis of Variance of Pretest Scores

Measure	BAU		Treatment		$F(1, 1725)$	η^2
	M	*SD*	*M*	*SD*		
Pretest Raw Score	4.828	2.529	4.617	2.517	2.973	.002

Note. $p = .085$. BAU group's $N = 717$, Treatment group's $N = 1010$

Table 15

One-Way Analysis of Covariance for Intervention Effects of Raw Scores

Source of Variation	Sums of Squares	*df*	Mean Squares	*F*	*p*-value	η^2
Intercept	2897.541	1	2897.541	757.213^{***}	<.001	.305
Pretest Raw Score	5268.588	1	5268.488	1376.812^{***}	<.001	.444
Treatment	48.732	1	48.732	12.735^{***}	<.001	.007
Error	6597.034	1724	3.827			
Total	75138.000	1727				

Note. $^{***}p < .001$. BAU group's $N = 717$, Treatment group's $N = 1010$

Table 16

Means, Standard Deviations, and One-Way Analysis of Variance of Gain Scores

Measure	BAU		Treatment		$F(1, 1725)$	η^2
	M	*SD*	*M*	*SD*		
Gain Scores	1.116	1.972	1.522	2.189	15.650^{***}	.009

Note. $^{***}p < .001$. BAU group's $N = 717$, Treatment group's $N = 1010$

Table 17

-2 Log-likelihood of the Posttest IRT Models

IRT Model	AIC	BIC	-2 Log-likelihood
Rasch	19506.60	19566.60	19484.604
2PL	19337.05	19446.14	19297.054
3PL	19316.96	19480.59	19256.962

Table 18

Descriptive Statistics of Pretest and Posttest Ability Estimates by Condition

	Pretest		Posttest	
Conditions	*M*	*SD*	*M*	*SD*
BAU	-0.390	0.813	-0.049	0.899
Treatment	-0.435	0.796	0.037	0.837

Note. BAU group's $N = 717$, Treatment group's $N = 1010$

Table 19

Means, Standard Deviations, and One-Way Analysis of Variance of Pretest Ability Estimates

Measure	BAU		Treatment		$F(1, 1725)$	η^2
	M	*SD*	*M*	*SD*		
Pretest Ability Estimates	-0.390	0.813	-0.435	0.796	1.340	.001

Note. $p = .247$. BAU group's $N = 717$, Treatment group's $N = 1010$

Table 20

One-Way Analysis of Covariance for Intervention Effects of Ability Estimates

Source of Variation	Sums of Squares	*df*	Mean Squares	*F*	*p*-value	η^2
Intercept	118.437	1	118.437	298.581^{***}	<.001	.148
Pretest Ability	605.901	1	605.901	1527.480^{***}	<.001	.470
Treatment	5.286	1	5.286	13.327^{***}	<.001	.008
Error	683.854	1724	0.397			
Total	1285.054	1727				

Note. $^{***}p < .001$. BAU group's $N = 717$, Treatment group's $N = 1010$

Table 21

TDCM Model Fit Indices

DCM Model	AIC	BIC	-2 Log-likelihood	Scaling Correction Factor
2-Class	39428.295	39553.741	39382.296	1.148
3-Class	38667.479	38874.736	38591.48	1.075
4-Class	38223.791	38741.935	38362.922	0.9958

Table 22

TDCM Classifications

	2-Class Model		3-Class Model		4-Class Model	
Latent Class	Pretest	Posttest	Pretest	Posttest	Pretest	Posttest
1	956	629	715	444	691	477
2	771	1098	762	635	790	176
3			250	648	2	695
4					244	379

Note. The black values indicate that the 2-class model did not allocate students in the third and fourth latent classes, and 3-class model did not allocate students in the fourth latent classes.

Table 23

3-Class TDCM Classification by Condition

Conditions	Pretest			Posttest		
	Low	Medium	High	Low	Medium	High
BAU	296	306	115	119	241	277
Treatment	415	452	143	242	392	376

Note. BAU group's $N = 717$, Treatment group's $N = 1010$

Table 24

Probability Correct of Items by 3-Class TDCM

	Correct Answer (%)		
Item Number	Class 1	Class 2	Class 3
Item 1	41	59	90
Item 3	9	20	53
Item 4	25	57	87
Item 5	25	43	89
Item 6	19	66	96
Item 7	31	88	98
Item 8	35	79	98
Item 9	24	84	97
Item 10	12	28	63
Item 11	29	56	74

Note. Item discrimination parameters can be obtained by subtracting the probability of the correct answer by group.

Table 25

Latent Class Transition of the 3-Class TDCM by Condition

	BAU			Treatment		
	Low	Medium	High	Low	Medium	High
Low	197	99	0	239	173	3
Medium	0	142	164	0	219	233
High	2	0	113	3	0	140

Note. The row indicates the pretest classification results while the column indicates the posttest classification results. For example, in the BAU group, 99 students were first classified into the low group for the pretest and transitioned into the medium group for the posttest. For another example, 219 students in the treatment group were first classified into the medium group in both pretest and the posttest.

Table 26

Proportion of Transitions of the 3-Class TDCM by Condition

Classifications		Condition	
Pretest	Posttest	BAU	PD
Low	Low	.275	.234
Low	Medium	.138	.171
Low	High	0	.003
Medium	Low	0	0
Medium	Medium	.198	.217
Medium	High	.223	.231
High	Low	.003	.003
High	Medium	0	0
High	High	.158	.139

Table 27

Exploratory Pseudo Q Matrix Attribute Hierarchies

Item Number	Attributes		
	Understanding the Concept	Understand the Concept of a Unit Rate	Solving Real World Problems
Item 1	1	1	0
Item 3	1	1	0
Item 4	1	1	0
Item 5	1	1	1
Item 6	1	1	1
Item 7	1	1	1
Item 8	1	1	0
Item 9	1	1	1
Item 10	1	1	0
Item 11	1	1	1

Note. Exploratory Pseudo Q Matrix is based on attribute description in CCSS.

Table 28

Exploratory HDCM Classifications

Class	Attributes			Number of Students
	Attribute 1	Attribute 2	Attribute 3	
1	0	0	0	305
2	1	0	0	480
3	1	1	0	242
4	1	1	1	699

Note. 0 indicates non-proficiency and 1 indicates proficiency. The first attribute refers to the understanding of the concept of ratios and proportions. The second attribute refers to the understanding of the concept of the unit rate. The last attribute refers to solving real world problems.

Table 29

Exploratory LCDM Classifications

Attributes			
Attribute 1	Attribute 2	Attribute 3	Proportions (%)
0	0	0	1
0	0	1	21
0	1	0	5
0	1	1	21
1	0	0	4
1	0	1	20
1	1	0	4
1	1	1	25

Note. 0 indicates non-proficiency and 1 indicates proficiency. The first attribute refers to the understanding of the concept of ratios and proportions. The second attribute refers to the understanding of the concept of the unit rate. The last attribute refers to solving real world problems.

Table 30

Exploratory HDCM Item Parameters

Item Number	Intercept	Main Effects		
		λ_1	λ_2	λ_3
Item 1	-0.313	0.000	1.737	
Item 3	-1.802	0.652	1.372	
Item 4	-1.337	1.146	1.717	
Item 5	-1.084	0.544	0.180	2.544
Item 6	-1.932	2.059	1.224	1.359
Item 7	-1.113	2.324	0.212	4.153
Item 8	-0.653	1.746	1.796	
Item 9	-1.462	2.047	0.845	2.200
Item 10	-1.967	0.218	1.923	
Item 11	-1.114	0.848	0.473	0.728

Note. Black parameters indicate that the attribute is not associated with the items.

Figure 1

A Pseudo Multiple Choice Item

7. There are a total of 24 apples in a basket. The ratio of green apples to red apples is 1:3. How many green apples are in the basket?

1. 4
2. 6
3. 7
4. 9
5. I don't know.

Note. The answer for pseudo multiple choice item is the second option.

Figure 2

A Pseudo Multiple Selection Item 2

2. When playing darts, John makes 3 out of every 8 shots he takes. Select **ALL** the statements that describe John's situation

1. The ratio of the number of shots John makes to the number of shorts he takes is 3:8.
2. The ratio of the number of shots John makes to the number of shorts he fails is 3:5.
3. The equation $3x = 5y$ shows the relationship between x, the number of shots John makes, and y, the number of shorts he did not make.
4. The equation $8x = 3z$ shows the relationship between x, the number of shots John makes, and z, the number of shorts he takes.
5. I don't know.

Note. The answer for pseudo item two is the first, third and fourth options. All multiple selection items had the statement to select all the correct answers. The word 'ALL' was both capitalized and bolded.

Figure 3

Scree Plot for Dimensionality Analysis

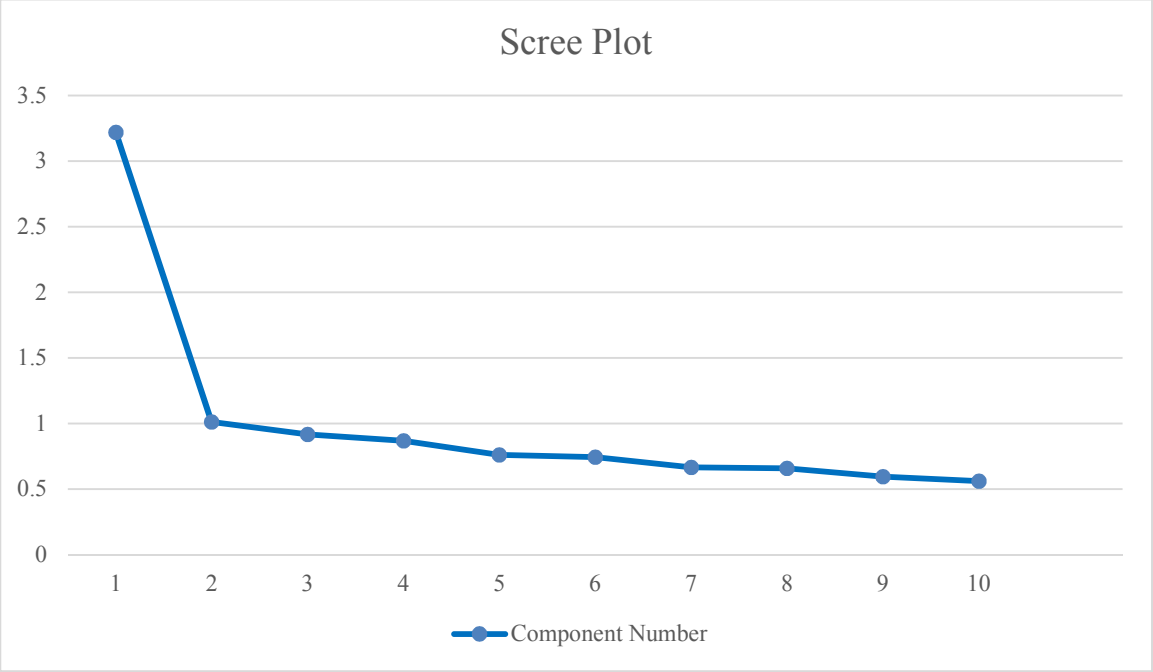
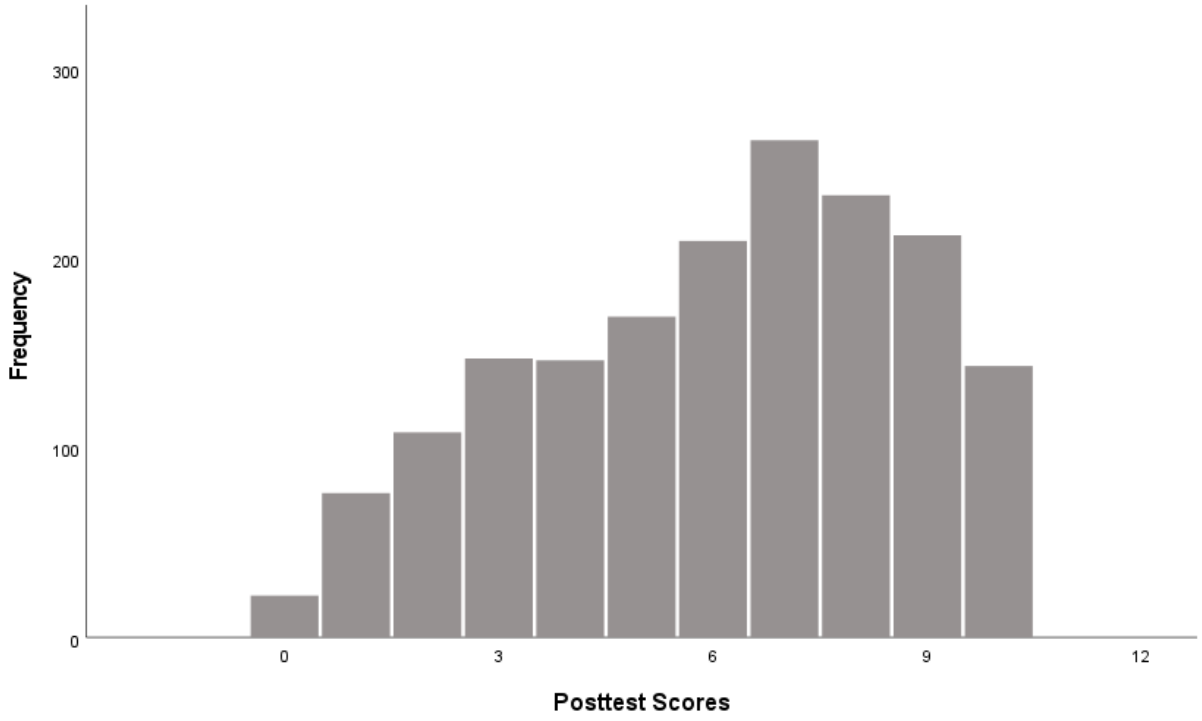


Figure 4

Simple Histogram of Posttest Raw Scores



Note. The minimum posttest raw score is 0 and the maximum posttest raw score is 10.

Figure 5

Plot of Raw Scores for BAU and Treatment Groups

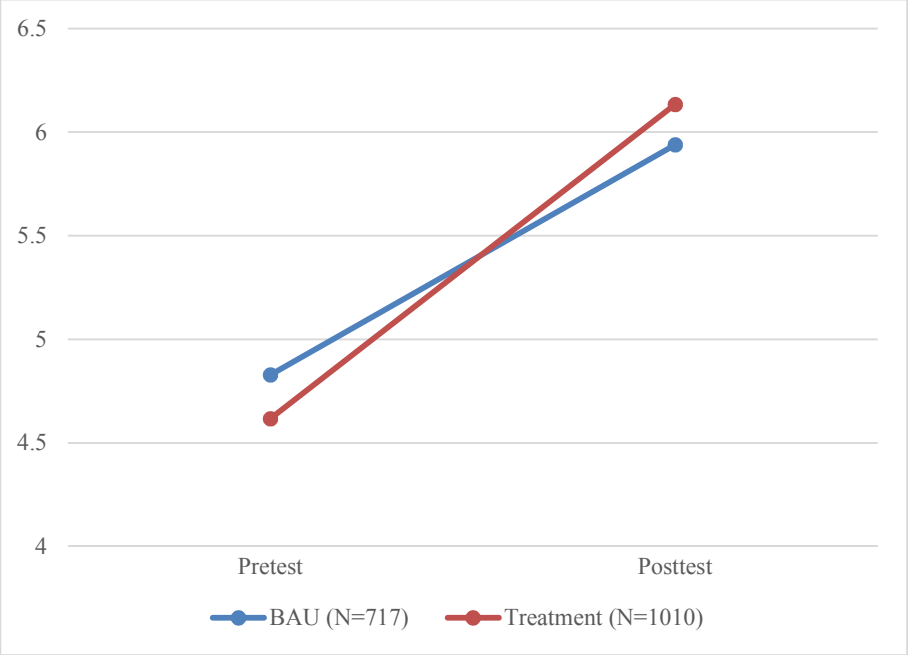
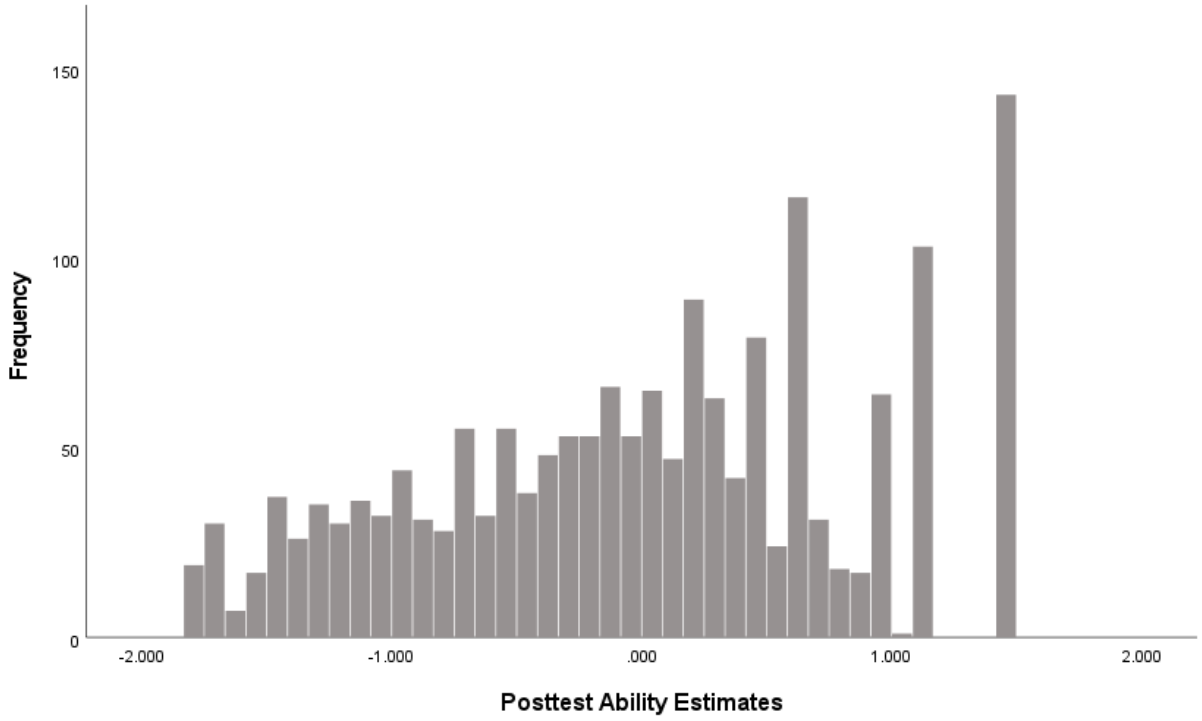


Figure 6

Simple Histogram of Posttest Ability Estimates



Note. The minimum ability estimate is -1.790 and the maximum ability estimate is 1.454.

Figure 7

Plot of Ability Estimates for BAU and Treatment Groups