Computerized Formative Assessment with the Item Digraph

by

Stephanie T. Varga

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Education

in

Measurement, Evaluation, and Cognition

Department of Educational Psychology

University of Alberta

Abstract

This paper introduces and tests a new approach I have designed for computerized formative assessment in education. The assessment, called GRAPH-CAT, is developed in the programming language of Python and a simulation is performed to measure its effectiveness. GRAPH-CAT is a cognitively diagnostic computerized adaptive test (CD-CAT) that estimates mastery of an attribute hierarchy using a directed acyclic graph and traditional computer adaptive testing (CAT). GRAPH-CAT reports ability based on both attribute mastery and traditional item response theory (IRT) ability. In this study, a Monte Carlo simulation of student responses is performed using a simulated item bank. Previous CD-CATs have generally not relied on traditional CAT based on the IRT framework. Instead, a single performance measure, such as θ, is usually replaced by classification into knowledge state as defined by mastered attributes. GRAPH-CAT provides an arguably more robust measure of performance as it is based on both attribute mastery and IRT ability. The introduction of the *item digraph* realizes the potential held between the connectivity of attributes to create an efficient test. A strength of GRAPH-CAT is that it departs from the CD-CAT reliance on stochastic item administration. It is found that GRAPH-CAT is able to estimate attribute mastery with 92% accuracy in twenty items with a standard error of 0.33 and achieves 83% accuracy in ten items with a standard error of 0.38. These results demonstrate how ordering items with an item digraph may help provide the needed structure for item administration in CD-CAT.

Preface

This thesis is an original work by Stephanie Varga. No part of this thesis has been previously published.

Dedication

This is dedicated to my late father whose conversation and wisdom helped shape my educational path. This is also dedicated to my friends and family who have given me endless support throughout my education. A final dedication goes to all teachers who find meaning in strengthening the mental agency of their students. Your work is appreciated.

Everything is connected.

Acknowledgements

I would like to thank my supervisor, Dr. Mark Gierl, for his guidance, encouragement, and support.

Table of Contents

I. Introduction 1

I. A Computer Adaptive Testing 4

I. A. i Evaluating a CAT with Monte Carlo Simulation 12

I. B Cognitively Diagnostic Assessment with the Item Digraph 13

I. B. i Required Background on The Rule Space Method. 16

I. B. ii Required Background on Graph Theory. 20

I. B. iii The Item Digraph. 25

I. B. iv The Theory of Difficulty to Attribute Alignment. 31

I. B. v New Definitions Central to GRAPH-CAT. 34

I. B. i Partially Ordered Set Theory. 41

I. B. ii Cascading Mastery Estimation. 44

I. C Concluding Remarks on CD-CAT with the Item Digraph 51

II. Problem Statement 52

III. Research Questions 53

IV. Thesis Statement 54

V. Methods 55

V. A Simulation Variables 56

V. B Simulating and Estimating the Learner 57

V. C Simulating Population Parameters 59
V. D Measuring Accuracy...64
V. E Simulation Conditions 70

VI. Results 71

VI. A Simulation 1: Test Stopping Rule of Complete Outcome Mastery Assignment 71
VI. B Simulation 2: Test Stopping Rule of Maximum Ten Items 78

VII. Discussion 82

VII. A Highlight of Major Findings...82
VII. A. i Standard Error of Measurement (SE)...84
VII. A. ii Maximum Test Length. 84
VII. A. iii Number of Items in the Bank. 85
VII. A. iv Test Scale...85
VII. B Limitations of the Study 86
VII. C Directions for Future Research...87
VII. C. i DIFFR: Equity. 87
VII. C. ii DIFFR: Security. 91
VII. C. iii DIFFR: Interface Design. 94

VIII. Conclusion 96

Bibliography 98

List of Tables

Table 1 Example of 2×3 Q Matrix 17

Table 2 An Example of Items that Might have the Q-Matrix in Table 1 18

Table 3 An Example of Items that Might be Represented by the Q-Matrix in Table 1 18

Table 4 Population Size of Each Simulated Population (N=538) 60

Table 5 Parameters of Simulated Item Bank 62

Table 6 A Description of the Frequency Match Categories 69

Table 7 The Attribute Frequency Matching Values for the Yamada (2008) Item Bank Using a Test Stopping Rule of Complete Outcome Mastery Assignment 72

Table 8 The MLE Outcome Frequency Matching Values for the Simulated Yamada (2008) Item Bank Using a Test Stopping Rule of Complete Outcome Mastery Assignment 74

Table 9 The Average Difference Between Estimation Grade and Real Grade for the Simulated Yamada (2008) Item Bank Using a Test Stopping Rule of Complete Outcome Mastery Assignment...76

Table 10 The Attribute Frequency Matching Values for the Yamada (2008) Item Bank Using a path-CAT Stopping Rule of Maximum Ten Items 78

Table 11 The MLE Outcome Frequency Matching Values for the Simulated Yamada (2008) Item Bank Using a Test Stopping Condition of Maximum Ten Items 79

Table 12 The Average Difference Between Estimation Grade and Real Grade for the Simulated Yamada (2008) Item Bank Using a Test Stopping Condition of Maximum Ten Items 80

List of Figures

Figure 1. The probability, P, of a correct response according to the Item Characteristic Curve (2-PL) (Thompson, 2009)..8

Figure 2. The probability, Q=1-P, of an incorrect response according to the Item Characteristic Curve (2-PL) (Thompson, 2009)..8

Figure 3. L= Likelihood Function (Thompson, 2009)..9

Figure 4. An example of attributes for fraction subtraction problems (Tatsuoka, 2009, pg 41) 14

Figure 5. An example of attributes for the specific outcome of Percents from the Grade 6 mathematic program of studies (Education, 2007)....................15

Figure 6. An undirected graph with 5 nodes..21

Figure 7. A directed graph (digraph) with 3 nodes..21

Figure 8. A tree graph with 5 nodes..23

Figure 9. A height-balanced tree...24

Figure 10. A rooted digraph with one sink 24

Figure 11 A diagrammatic representation of the prerequisite relationship 25

Figure 12. An item tree for fraction subtraction problems (Tatsuoka, 2009, page 130) 27

Figure 13. A rooted item digraph depicting the item tree in Figure 12 rotated by 180^0 28

Figure 14. An item digraph derived from The Alberta Mathematics Kindergarten to Grade 12 Scope and Sequence Document (Education, 2017)................29

Figure 15. A conceptual diagram of the hierarchical path.....................................35

Figure 16. An example of a mastery path..36

Figure 17. Ability versus mastery in a mastery path 37

Figure 18. An example of a mastery graph...38

Figure 19 An example of a hierarchical path in the Grade 6: Number-Specific Outcome 6 (Percents) item digraph 39

Figure 20 A mastery path associated with the hierarchical path in Figure 19 for a student with an IRT ability of 0..40

Figure 21. A mastery graph for a student with an IRT ability of 0 in every hierarchical path of the Grade 6-Percents item digraph in Figure 19 40

Figure 22. A Hasse diagram representation of the Q matrix in Table 1 (Yamada, 2008) 43

Figure 23. Illustration of an upwards subgraph 46

Figure 24. Illustration of a downwards subgraph 46

Figure 25. Ability and mastery of a student with IRT ability of 0 in path 1 of the Grade 6-Percents item digraph 49

Figure 26. Mastery node labelling according to Cascading Mastery Estimation (CME) as determined by the ability and mastery in Figure 25 50

Figure 27. A diagrammatic representation of the Zone of Proximal Development adapted from Kym Buchanan 51

Figure 28 The item digraph derived from the Yamada (2008) item bank. The items and attributes for each learning outcome are indicated within each node. 63

Figure 29. An item digraph with imbalanced attribute assignment.........................66

Figure 30. Diagrammatic representation of attbygr and ocbygr.............................68

Figure 31 Number of paths administered for a test stopping rule of complete outcome mastery assignment. Note that grDIFF indicates mastery level where grDIFF=1 is high ability and grDIFF=4 is low ability (Table 4). 77

Figure 32 Number of paths administered for a test stopping rule of maximum ten items. Note that grDIFF indicates mastery level where grDIFF=1 is high ability and grDIFF=4 is low ability (Table 4) 81

I. Introduction

The underlying theory behind any computerized formative assessment tends to focus on the estimation of *attribute* mastery. An *attribute* is a cognitive skill or process required to correctly answer a problem (Yamada, 2008). Traditionally, items have been associated with new attributes introduced by a lesson, while assuming that the necessary prerequisite attributes have been mastered. Computerized formative assessment, on the other hand, tests both newly introduced and prerequisite attributes.

Formative assessment provides a diagnostic measure whereas the traditional ability test provides a summative measure. A key difference between diagnostic versus ability testing is that considerably more skills and knowledge outcomes are measured in diagnostic testing. The magnitude of the difference can be illustrated by considering the topic of *Percents* from Grade 6 Mathematics. In mastery testing, items are designed to only test attributes from the outcome of Grade 6 *Percents*. In contrast, diagnostic testing requires that every prerequisite outcome from Kindergarten to Grade 5 must also be tested. According to Alberta Education

(2017), the magnitude of difference between mastery and diagnostic testing for Grade 6 *Percents* is a factor of approximately nineteen.

This leap in magnitude of tested skills and knowledge is normal for any subject and makes a computer adaptive test (CAT) necessary for cognitively diagnostic assessment. A CAT personalizes a test to the ability of the student so that every item does not have to be administered. A CAT is generally much shorter than the full-length test and has traditionally been used for ability—summative—testing.

The instrument most commonly used for computerized formative assessment is the cognitively diagnostic computer adaptive test (CD-CAT). CD-CATs generally have not been systematic in certain factors that provide intelligibility to item sequencing by content or difficulty (Yamada, 2008; Falmagne, Cosyn, Doignon, & Thiéry, 2006; Wang, Chang, & Huebner, 2011; McGlohen, & Chang, 2008; Gierl, Alves, & Majeau, 2010). This is due to item selection rules that rely on stochastic item administration for ability measures. Item selection rules for CD-CAT attempt to maximize information about the location of the examinee in an ability space or minimize error in the estimation of ability. The Kullback-Leibler function has widely been the approach taken for CD-CAT as it allows for a measure of attribute mastery that does not rely on a continuous ability scale. This has proved useful for providing a discrete measure of attribute mastery.

This paper offers the following salient contributions to previous implementations of CD-CAT. To provide structure and sequencing to item administration, a graph data structure, referred to as the *item digraph*, is introduced. The paths of the item digraph induce an organization on the attributes by common dimensionality. This organization allows for item administration based on the maximum likelihood estimation function. A test shortening technique based on the item digraph, referred to as *cascading mastery estimation*, is proposed and proved. A theory for the estimation of prior item difficulty based on the item digraph is also proposed, however, practical verification is left as a direction for future research. A measure of student progress is made available by monitoring the rate of expected mastery towards the learning goal. Meanwhile, the longitudinal monitoring of the item digraph offers a graph-theory based interpretation for the unification of formative and summative assessment. Monitoring of the item digraph also suggests an approach for constructing the ideal item digraph based on student responses. This is discussed in detail under Directions for Future Research.

GRAPH-CAT is a novel CD-CAT that uses the *item digraph* to estimate mastery of attributes. Attribute mastery using GRAPH-CAT is dependent on a *well-structured hierarchical curriculum*. A *well-structured hierarchical curriculum* is one that has been reviewed and designed by subject matter experts, so that certain properties of the attributes and the attribute hierarchy are standardized. The well-structured curriculum groups the units of latent skills, known as attributes, into larger sets referred to here as *learning outcomes*.

GRAPH-CAT reports back to the teacher and student about prerequisite learning outcomes that have not yet been mastered by the student. Supplementary learning materials are provided based on those learning outcomes that are not yet mastered. This type of embedded formative assessment acknowledges that many students are deficient in certain required knowledge and skills before learning a new topic.

Embedded formative assessment respects and enhances the structure of the traditional lesson-cycle (William, 2011). As an embedded assessment GRAPH-CAT keeps the teacher in-the-loop and provides them with a curriculum-companion. For students, the test helps locate their ability in an otherwise impersonalized curriculum. GRAPH-CAT aims to improve the learning experience for both teacher and student by tailoring the curriculum to individual learning needs.

I. A Computer Adaptive Testing

A CAT helps lower or completely reduce the levels of confusion and frustration by ensuring the student is constantly receiving items that are neither too easy nor too difficult (Maravic, Cesar et al., 2010). This adapts the test items to the examinee to ensure they stay in what Vigotsky called the Zone of Proximal Development (ZPD)(Shabani, Mohamad, & Saman, 2010). The ZPD holds the idea that individuals learn best when given tasks that are slightly more difficult than what they can do, such that they will need to work together with another to finish the task (Shabani, Mohamad, & Saman, 2010). Overall, the CAT maximizes both accuracy and efficiency, reducing test-fatigue (Čisar, Radosav, Markoski, Pinter, & Čisar, 2010). By creating a unique testing experience, CAT increases security and decreases the risk of cheating. The advantages to test administrators include a reduced testing time and increased reliability (Kantrowitz, Dawson, & Fetzer, 2011). The advantages to examinees include a shorter test length and a sense of interactivity and personalization (McGlohan & Chang, 2008).

The maximum likelihood estimation (MLE) function is an ability measure that is widely used for implementing a CAT (Hsu, Wang, & Chen, 2013). According to Hsu, Wang, and Chen (2013) the MLE measures the amount of information an observable random variable X carries about an unknown parameter θ. MLE has not generally been used in CD-CAT as the conditional distribution of X *—learner responses*—must be continuous with respect to θ. To understand how the MLE can be used for CD-CAT, a brief theoretical background of CAT is first introduced. Note that the formulas in this section are intended for interest only and are not necessary for understanding the theory.

The following notation will be adopted from van der Linden and Pashley (2000). Items in the item pool, I, are denoted by $i_k \in I$ where the index of administration order is denoted by k = 1...K, and K is the length of the test. The probability of a correct response to item *i* is given by

$$P_i(\theta)=\frac{\left(e^{a*(\theta-b)}\right)}{1+e^{a*(\theta-b)}}=\frac{1}{1+e^{-a*(\theta-b)}}$$

where a is the discrimination, *b* is the difficulty of the item i, and θ is the ability of the learner (Lord, 1980).

The formula for $P_i(\theta)$ is grounded in the item response theory (IRT) framework. IRT is a paradigm for using multi-item scales to determine ability of hypothetical constructs (Embretson & Reise, 2013). Traditional computer adaptive testing is based in IRT and depends on ability*,* θ*,* and item difficulty*, b,* being on the same scale. The values of θ and *b* are continuous and are usually found in the range of -3 to 3, although values beyond this range are also possible (Baker& Kim, 2004). The alignment of ability and difficulty scales allows a CAT to administer items according to ability (Linacre, 2000).

According to the assumptions of IRT, the probability of answering item i_m correctly is considered independent from the probability of answering item i_n correctly for any $m \neq n$, $i_m, \ i_n \in I$ (Lord, 1980). This allows for construction of the maximum likelihood function as follows. Given the response pattern, u, to the first N items of a test, a likelihood function provides the probability that the examinee has a specific ability for the construct being measured. The likelihood function L is defined as:

$$\mathrm{L}\ (u|\theta_j) = \prod_{k=1}^{N} P_{kj}^{u_{kj}} Q_{kj}^{1-u_{kj}}$$

where P_{kj} is the probability that an examinee with ability θ_j has given a correct response to item i_k, u_k is the correctness of the response to item i_k, and Q_{k_j} = 1 - P_{kj} is the probability that an incorrect response is given to item i_k. Notice that the exponent of P_{kj} is 1 when the correct answer is given and it is 0 when the incorrect answer is given. This multiplicand disappears—equals 1—if the answer is incorrect. Similarly, the exponent of Q_{kj} is 0 when the correct answer is given, in which case this multiplicand is 1.

To visualize what is occurring, let us consider an example as provided by Thompson (2009). Let let i_m and i_n be two items with the same difficulty and discrimination. The item characteristic curve for the item can be seen in Figure 1 below. This two-parameter logistic (2-PL) item characteristic curve (ICC) gives the probability that a learner with ability theta (θ) correctly answers an item with given difficulty and discrimination. Figure 2 displays the probability of an examinee with ability theta (θ) giving an incorrect response to an item with the ICC in Figure 1.

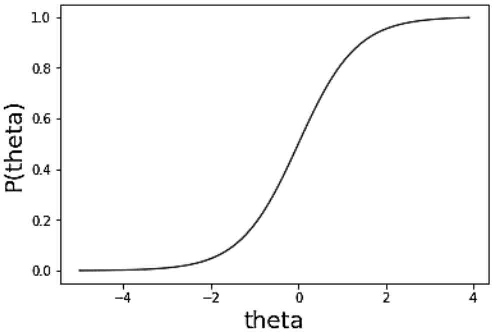


Figure 1. The probability, P, of a correct response according to the Item Characteristic Curve (2-PL) (Thompson, 2009)

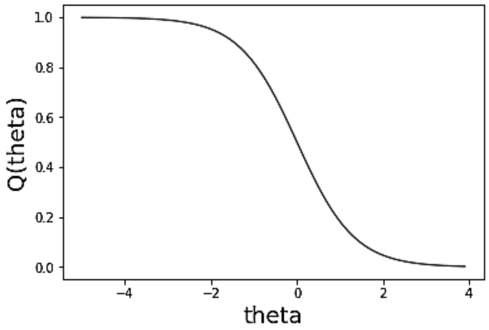


Figure 2. The probability, Q=1-P, of an incorrect response according to the Item Characteristic Curve (2-PL) (Thompson, 2009)

If an examinee has given a correct response to item i_m and an incorrect response to item i_n, then the two curves in Figure 1 and Figure 2 are multiplied to get the likelihood function in Figure 3. The task is to find the maximum point on the likelihood curve. In this case the maximum point is located at theta= 0 (Figure 3). This theta value represents the most likely ability of the examinee, given their response pattern so far.

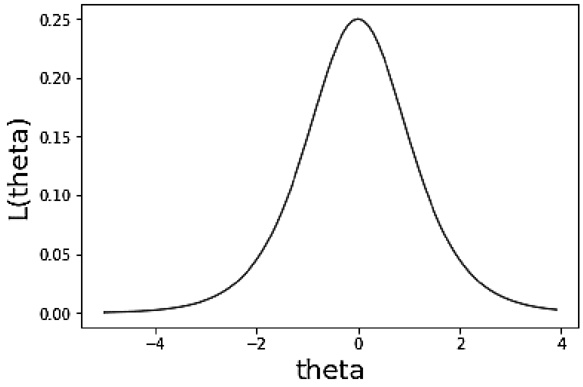


Figure 3. L= Likelihood Function (Thompson, 2009)

The formula

$$I(\theta) = \sum I_i$$

is used to calculate the test information where I_i is the item information function such that,

$$I_i = \frac{P'(\theta)}{P(\theta)Q(\theta)} \text{ (Ackerman, 1989)}$$

The item information tells us the amount of information an item provides about a student with ability θ. Similarly, the test information function provides a measure of the amount of information a test provides about a student with ability θ.

The maximum likelihood function also produces an index of error, called the standard error of measurement (SEM) (Thompson, 2009). According to Thompson (2009) the SEM is measured by the spread of the curve, where a wider curve indicates more error. The SEM is the square root of the inverse of the test information function (Lord, 1980).

$$SEM = \frac{1}{\sqrt{I}}$$

There are three common approaches used to estimate the real ability, θ, of a learner. The most commonly used is the maximum likelihood approach described above (Figure 3). The likelihood function need not have a maximum, and if it does, the maximizer is not necessarily unique (Geyer, 2003).

A variant of the MLE is the maximum a posteriori or MAP for short. In this case the likelihood function is multiplied by a *posteriori* curve that represents the probability distribution of the response data from a test (Yan & Magis, 2016). The MAP, unlike MLE, accommodates for the response patterns that are all correct or all incorrect. Such uniform response patterns reveal a weakness of the MLE as they cause the function to continually increase or decrease, so that a true maximum does not exist.

A variant of the MAP Bayesian adaptive algorithm is the *expectation a posterior* (EAP) (Yan & Magis, 2016). The EAP further accommodates for asymmetrical likelihood functions. Rather than finding a single maximum point, the EAP takes the average likelihood value as weighted by a posterior function.

Unlike the EAP, both the MLE and MAP functions require that a maximum point be found. Three common approaches are the Bisection Method, the Newton-Raphson method, and the Brute Force Method. The Bisection Method divides the range of the likelihood function into two sections and evaluates the derivative of the midpoint of these sections (Thompson, 2009). According to Thompson (2009), if the derivative is negative, the next iteration eliminates a section of possible theta values. The Newton-Raphson method, on the other hand, uses both the first and second derivative (Thompson, 2009). It determines the rate at which the slope is increasing or decreasing to find the maximum. The signs of the first and second derivate act as indicators as to where θ is located relative to the maximum.

The most straightforward approach to finding the maximum is to evaluate the MLE for every value of θ in the given range. This is called the Brute Force Method and, traditionally, it has been viewed as undesirable as it is computationally expensive. For example, 6000 iterations are required to calculate and compare every θ at 0.001 increments between -3 and 3 (Thompson, 2009). Advances in computational speed have overcome this barrier so that the Brute Force Method is now a viable option for determining the maximum of the MLE and MAP.

I. A. i Evaluating a CAT with Monte Carlo Simulation The quality of a computer adaptive test (CAT) depends on quantities of interest such as test length, accuracy, and significance. A Monte Carlo simulation (MCS) can be used to evaluate the quality of a CAT without having to create test items (van der Linden & Pashley, 2010). MCS is a method of generating large amounts of data from a given distribution so that the quality of a CAT can be explored using methods of statistical inference (Kroese, Brereton, Taimre, & Botev, 2014). MCS has traditionally been considered a last resort when other methods are not feasible, but it is now widely regarded as a highly useful method of experimentation (Kroese, Brereton, Taimre, & Botev, 2014). According to Krose et al. (2014) MCS is frequently used in applications of science, finance, and engineering due to its wide-ranging applicability and simplicity.

MCS has its roots in World War II from a group of scientists working at Los Alamos Scientific Laboratory (LSASL). It was first used to model neutron diffusion in nuclear fission where the path of a neutron was traced as it underwent various interactions (Gass & Assad, 2005). According to Gass, Arjun, and Assad (2005), the fission researchers were able to determine actual outcomes by sampling from known distributions of neutron collision-types.

MCS allows for the evaluation of CAT-related issues by sampling learner responses from a given distribution of ability (van der Linden & Pashley, 2010).

According to van der Linden and Pashley (2010) MCS is able to evaluate item exposure, size of the item bank, and precision of examinee scores. MCS simulates the administration of a CAT under varying conditions for a large number of examinees. Using a random number generator, the entire data set of binary values representing correct and incorrect responses is generated given a sample of simulated ability (θ) values. This approach is highly valuable when real student responses are not available.

I. B Cognitively Diagnostic Assessment with the Item Digraph

The described approach to CAT is generally used for ability testing as it produces a continuous measure of ability (θ). For use in diagnostic testing, the continuous measure of ability produced by CAT must be reconciled with a discrete measure of attribute mastery. An example of attributes with a hierarchical structure can be seen in Figure 4 below (Tatsuoka, 2009).

Attributes
• Converting a whole number to a fraction or mixed number
• Separating a whole-number part from a fraction part
• Simplifying before getting the common denominator
• Finding the common denominator
• Borrowing one from the whole-number part
• Column borrowing for subtraction of the numerators
• Reducing answer to the simplest form
• Subtracting numerators

Figure 4. An example of attributes for fraction subtraction problems (Tatsuoka, 2009, pg 41)

An example of a well-structured curriculum of attributes is offered by the *achievement indicators* from the Grade 6 Specific Learning Outcome of *Percents* outlined by the Alberta Program of Studies in Grade 6 Mathematics (Education, 2007) (Figure 5). These attributes belong to the same learning outcome hence they are not ordered by a hierarchy as none is a prerequisite to the other.

Specific Outcome	Achievement Indicators
Demonstrate an understanding of percent (limited to whole numbers), concretely, pictorially and symbolically.	• Explain that "percent" means "out of 100." • Explain that percent is a ratio out of 100. • Use concrete materials and pictorial representations to illustrate a given percent. • Record the percent displayed in a given concrete or pictorial representation. • Express a given percent as a fraction and a decimal. • Identify and describe percents from real-life contexts, and record them symbolically. • Solve a given problem involving percents.

Figure 5. An example of attributes for the specific outcome of Percents from the Grade 6 mathematic program of studies (Education, 2007)

The cognitive model used by GRAPH-CAT describes attribute mastery around a curriculum structure. According to Roussos, DiBello, Stout, Hartz, Henson, and Templin (2007), the notion of describing the cognitive model is central to all diagnostic tests. The GRAPH-CAT cognitive model combines elements of Graph Theory, the Rule Space Method (RSM)(Tatsuoka, 2009), and Partially Ordered Set Theory (POSET). The *item digraph*, introduced and described herein, plays a central role as it allows for a discrete measure of attribute mastery that is based on a continuous measure of IRT ability (θ).

As of yet, Shadow Testing is one of the few successful attempts at implementing a diagnostic test using both IRT and attribute mastery (McGlohen &

Chang, 2008). McGlohen and Chang (2008) conducted a study on diagnostic testing with Shadow Testing where the item parameters were precalibrated on the basis of a simple random sample of state examinees who wrote a state-mandated large-scale assessment. The item bank contained 396 items for the math portion of the test. The estimated parameters of 3000 of the 6000 examinees who wrote the test were used in the simulation. They found that this model achieved, on average, 81.7% accuracy across all attributes. This value is considered ‘good’ as the accuracy is high enough to provide useful formative feedback. To understand how a graph data structure can be used to estimate ability using both IRT and attribute mastery, let us first look into the required theories of the RSM.

I. B. i Required Background on The Rule Space Method. Tatsuoka and Tatsuoka’s (1983) Rule Space Method (RSM) is one of the founding models for cognitive diagnosis (Tatsuoka, 2009). The RSM is a probabilistic model designed to determine mastery of *attributes*. The RSM gives the probability of a response pattern, expressed by an ordered pair

$$(\theta, \zeta)$$

where θ represents ability and ζ represents a cautionary index (Tatsuoka, 2009). The cautionary index (ζ) provides a measure of response unusualness and it can act as an indicator of cheating or random responses (Yamada, 2008). In a sense, the

RSM is analogous to classification in machine learning where input is classified into output classes based on observable features. An example of this is the classification of handwritten numbers into text-based numbers. The issue with using attributes as features is that often they are not directly observable. This latency is a problem when trying to classify response patterns into states of attribute mastery.

The Q-matrix was introduced to account for precedence relationships—pre-requisite relationships—that manifest as latent attributes. The Q-matrix describes the precedence relationship in detail so that the RSM is able to extract information about the student (Tatsuoka, 2009). The Q matrix represents the items and the attributes that they measure. If an attribute is needed to solve the item, then the item × attribute entry is set to 1. If the attribute is not needed by the item, then the entry is set to 0. An example of a Q matrix, taken from Yamada (2008) can be seen below:

Table 1 *Example of 2×3 Q Matrix*

		Items			
		1	2	3	4
Attributes	1	1	1	0	0
	2	0	1	1	0

Note: Yamada, 2008

In the matrix configuration above, a 1 in the column of an item indicates that the item provides a measure for mastery of the corresponding row attribute. Similarly, a 0 indicates that the item does not measure the corresponding row attribute. For example, in column 1, the vector $\begin{bmatrix}1\\0\end{bmatrix}$ indicates that item 1 measures attribute 1 and does not measure attribute 2. Concretely, an example of items and attributes that might be represented by the Q-matrix in Table 1 above can be seen in Table 2 and Table 3 below.

Table 2 *An Example of Items that Might have the Q-Matrix in Table 1*

#	**Attributes**
1	Adding Numerators
2	Reducing answer to simplest form

Table 3 *An Example of Items that Might be Represented by the Q-Matrix in Table 1*

#	**Items**
1	3/4 + 3/4= ?
2	What is ¾ + ¾ in reduced mixed form?
3	Convert 6/4 to reduced mixed form
4	4 × 4= ?

The mapping from response patterns to attribute mastery takes into account *ideal responses*. The *ideal response pattern* would occur if no *slips* have been made. *Slips* are clerical errors that stem from inattention or carelessness rather than deficiencies in understanding or knowledge (Tatsuoka, 2009). They lead to incorrect responses and response patterns that are not ideal.

The ideal response pattern can be inferred from attribute mastery patterns. Attribute mastery patterns, also referred to as *knowledge states*, are the true set of attributes that have been mastered by the learner (Tatsuoka, 2009). The mapping from attribute mastery pattern to ideal response pattern is determined by the Boolean Descriptive Function (BDF)(Tatsuoka, 2009).

The BDF is used to derive the ideal response patterns from a knowledge state. Tatsuoka (2009) states the definition of the BDF as follows. Given an item X and the set of attributes A = $\{a_1, a_2, \ldots a_L\}$ required to solve X:

$$\text{BDF}(A)=\begin{Bmatrix} 1 & \quad if\ a_i\ is\ mastered\ for\ all\ a_i \in A \\ 0 & \quad if\ there\ exists\ a_i \in A\ such\ that\ a_i is\ not\ mastered \end{Bmatrix}$$

The proposed graph theory approach to CD-CAT makes a slight adaptation to the BDF as follows:

$$\text{BDF}(A)=\begin{Bmatrix} 1 & \quad iff\ a_i\ is\ mastered\ for\ all\ a_i \in A \\ 0 & \quad iff\ there\ exists\ a_i \in A\ such\ that\ a_i is\ not\ mastered \end{Bmatrix}$$

The *iff* (if and only if) indicates a two-way implication. Therefore, if all attributes of a set A are mastered then BDF(A)= 1 *and also* if BDF(A)=1 then all attributes of A are mastered. Similarly, if there exists at least one attribute of set A that is not mastered, then BDF(A)=0 *and also* if BDF(A)=0 then there is at least one attribute of set A that is not mastered. This updated definition of the BDF is essential for understanding *assignment of mastery* to learning outcomes (see New Definitions Central to GRAPH-CAT).

I. B. ii Required Background on Graph Theory. Several attempts have been made to base a CD-CAT on the RSM (Tatsuoka, 2009). A common problem in previous attempts is that item selection is random until a general estimation can be inferred about the knowledge state of the learner (Yamada, 2008). In practice, this could potentially cause confusion for the examinee, as the difficulty and content of the administered items is unstructured. As demonstrated in the following sections, a *graph* can provide order to item administration.

Graph theory is a branch of mathematics that employs graphs to facilitate the understanding of relationships between objects of interest. A *graph* is composed of nodes, also called vertices or points. In practice, the nodes are associated with the object of interest and the edges are associate with a relationship between the objects (Tatsuoka, 2009). In modelling a well-structured curriculum with a graph, the nodes are associated with a learning outcome and the edges represent a prerequisite relationship between the learning outcomes.

The edges of a graph can be undirected or directed (Figure 6 and Figure 7). In an undirected graph, there is no difference between the two vertices connected by an edge (Voloshin, 2009). Conversely, in a directed graph, the edges are directed. The nodes and edges of a graph can be represented as sets of unordered or ordered pairs. For an undirected graph, the pairs in the set are unordered and are represented using curly brackets:

E(G)= { {1,2}, {1,5}, {2,3}, {2,4}, {2,5}, {3, 1} }

For a directed graph, also referred to as a *digraph*, the pairs in a set are ordered and are represented using parentheses: E(G)= { (1,2), (2,3), (3,1) }

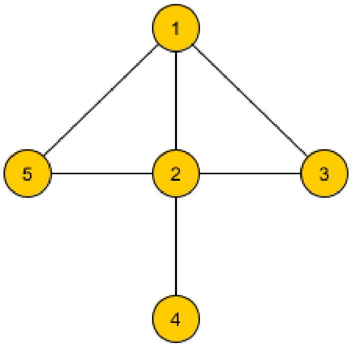


Figure 6. An undirected graph with 5 nodes

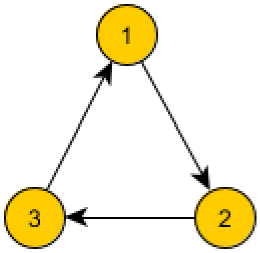


Figure 7. A directed graph (digraph) with 3 nodes

The nodes are connected by edges, otherwise called arcs or lines. In an undirected graph, two nodes *x and y* of *G* are *adjacent*, or *neighbours*, if *{x, y}* is an edge of *G* (Diestel, 2017). In a digraph, two nodes x, y of G are adjacent if (x, y) is an edge of G. A vertex v is *incident* with an edge e if $v \in e$ (Diestel, 2017). The degree of a node is the number of its neighbors (Diestel, 2017).

A *walk* of length k in a graph is an alternating sequence of vertices and edges which begins and ends with vertices:

$$(v_0, e_0, v_1, e_1, \dots v_{k-1}, e_{k-1}, v_k)$$

The edges and vertices need not be distinct in a walk. If the graph is directed then e_i is an edge from v_i to v_{i+1} (Diestel, 2017). A *trail* is a walk in which all edges are distinct. A *path* is a trail in which all vertices, except possibly the first and last vertices, are distinct. A path in a graph is denoted by a sequence of edges that connect a sequence of adjacent vertices. In a digraph, movement between any vertices in a path is unidirectional (Diestel, 2017).

A *cycle* is a directed path beginning and ending at the same vertex which passes through at least one other vertex (Figure 7) (Aho, Garey, & Ullman, 1972). A *loop* is an edge of the form (v, v), meaning it connects a vertex to itself (Aho, Garey, & Ullman, 1972). A *simple graph* contains no loops and an *acyclic graph* contains no cycles (Aho, Garey, & Ullman, 1972).

A structure that will help introduce the item digraph, is the *tree* (Figure 8). In a mathematical context, a *tree* is an undirected graph in which any two nodes are connected by exactly one path (Diestel, 2017).

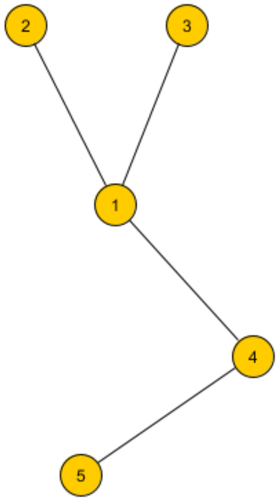


Figure 8. A tree graph with 5 nodes

A *rooted tree* is a tree in which a special node is singled out, as can be seen by node 1 in Figure 8. As a data structure, the edges of the tree are implicitly directed away from the root (Shaffer, 1997). The *level* of a node in a rooted tree is its distance as determined by the number of edges in the path from the root to the node (Diestel, 2017). The *child* of a node is an adjacent node at a higher level. A leaf of a rooted tree is a node with no children. A rooted tree data structure where each leaf is at the same level is often referred to as being *height balanced* (Figure 9) (Shaffer, 1997).

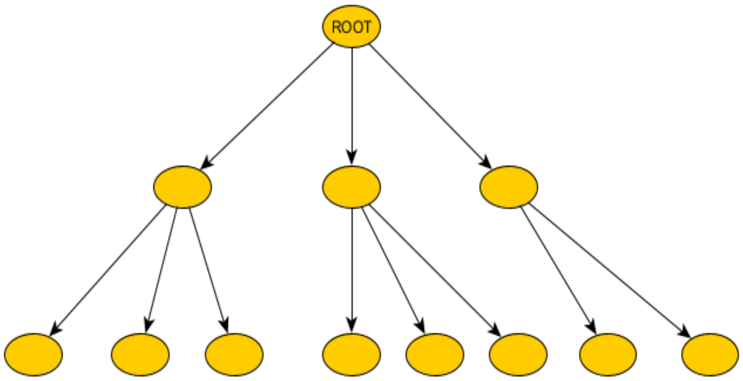


Figure 9. A height-balanced tree

A *rooted digraph* is a directed graph in which a special node is singled out as the root (Figure 10)(Harary, 1955). A *sink* of a rooted digraph is a node with no outgoing edges (Figure 10)(Margoliash, 2010). As with the rooted tree, the *level* of a node in a rooted digraph is the number of edges in the path from the root to the node.

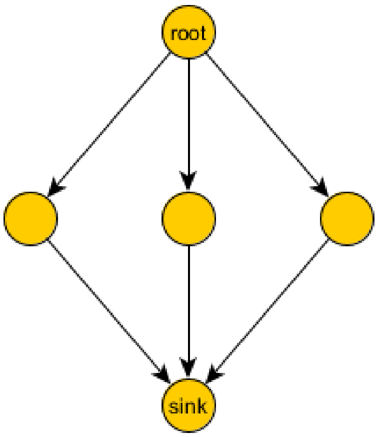


Figure 10. A rooted digraph with one sink

The next section combines these graph theory concepts with the RSM to define the item digraph.

I. B. iii The Item Digraph. The relation "A is harder than B," will be expressed using a directed edge (Figure 11).

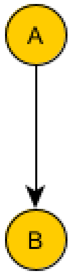


Figure 11 A diagrammatic representation of the prerequisite relationship

The directed edge represents outcome B as being prerequisite to outcome A. To gain deeper insight into the intuitive meaning of a precedence relationship, consider the following example taken from Tatsuoka (2009):

$$Item\ A\colon add\ \frac{1}{2}\ and\ \frac{2}{3}$$

$$Item\ B\colon add\ \frac{1}{3}\ and\ \frac{2}{3}$$

Item A is harder than item B because in addition to the attributes required for item A, it requires mastery of the cognitive process of getting the common denominator of 1/2 and 2/3.

Given two items from a Q-matrix, precedence between the items is discussed in terms of the relationship between their measured attributes. If item i_m measures a set of attributes A_m and i_n measures a set of attributes A_n, where

$$A_m \subseteq A_n$$

then the cognitive processes required to solve i_m are *prerequisites* to those required to solve i_n. This relationship forms an order on the items, attributes, and learning outcomes.

Tatsuoka (2009) observes that by looking into the set theoretical representations of the prerequisite relationships, it is possible to write chains or totally ordered sequences of items. She diagrammatically represents the precedence relations using what she calls an *item tree* (see Figure 12).

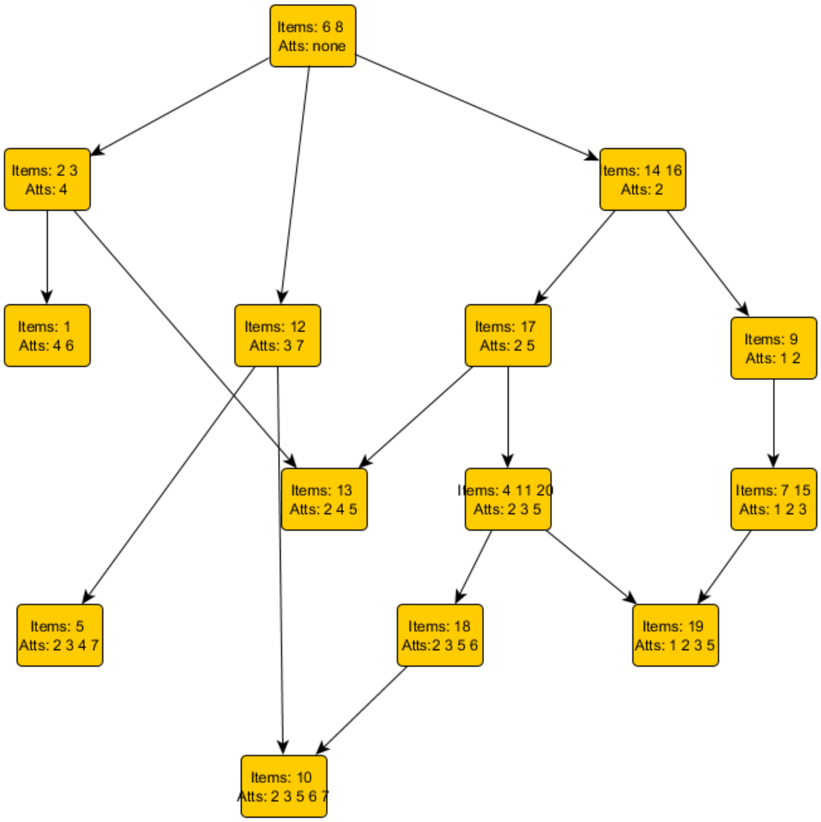


Figure 12. An item tree for fraction subtraction problems (Tatsuoka, 2009, page 130)

To construct the item digraph, let us first root the item tree at a node of interest, and restrict interest to those nodes connected to the root. For example, we root the item tree in Figure 12 at the node containing item 10 measuring attributes {2, 3, 5, 6, 7}. Figure 13 is obtained by rotating the item tree in Figure 12 by 180 degrees so that the most difficult item (item 10) is at the top.

Note that the *item tree* does not satisfy the mathematical definition of a tree as it allows for more than one path between two vertices (Figure 12). Therefore,

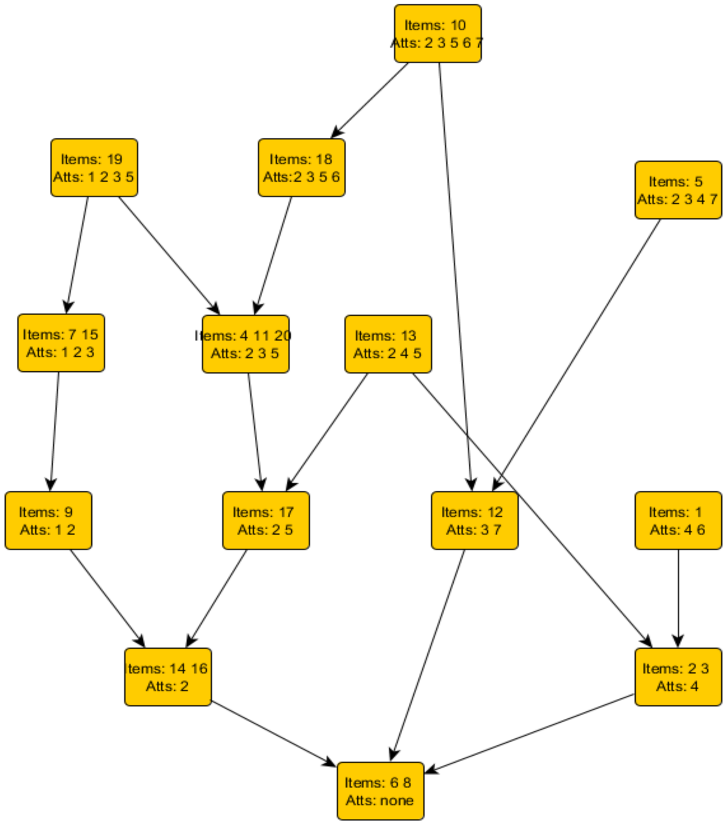


Figure 13. A rooted item digraph depicting the item tree in Figure 12 rotated by 180^0

the rotated item tree in Figure 13 will be referred to as a *rooted item digraph* or *item digraph* for brevity. The learning outcome for which diagnosis is sought is represented as the root such that items closer to the root are more difficult than items further from the root. The *item digraph* is a *directed acyclic graph* that is both simple and rooted where set inclusivity of attributes is represented in as few edges as possible (see Required Background on Graph Theory). It is *simple* as learning outcomes are not discussed as being prerequisite to themselves. Further, it is *acyclic*, as topics learned at the same level are not considered to be prerequisites to one another. Thus, edges are always directed from a lower to a higher level.

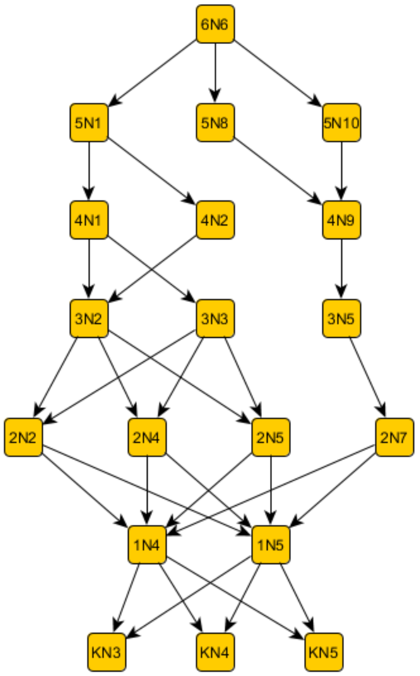


Figure 14. An item digraph derived from The Alberta Mathematics Kindergarten to Grade 12 Scope and Sequence Document (Education, 2017)

Each node of the item digraph represents a learning outcome from a curriculum of studies. The terms *node* and *learning outcome* will often be used here interchangeably. Each node contains both a set of items and the attributes measured by the items. It follows that each node is, implicitly, associated with a unique Q-matrix. Further, each edge represents set inclusivity between the attributes of the nodes.

A well-structured hierarchy of learning outcomes and their associated attributes, allows for the construction of an item digraph. For example, Figure 14 is the item digraph extracted from the *Percents* topic from Alberta Grade 6 Mathematics Specific Outcome 6 (Education, 2007). The arrow from tail to head represents a prerequisite relationship from a more difficult learning outcome in a higher grade to an easier learning outcome in a previous grade.

The levels in the item digraph are representative of hierarchical sets of attributes such that for any nodes v_i and v_j in a path from root to sink we have

$$level(v_i) \leq level\ (v_j)\ iff\ A(v_j)\ \subseteq\ A(v_i)$$

for all $0 \leq i \leq L,\ 0 \leq j \leq L$ where A(v_n) represents the set of attributes belonging to node v_n and L represents the depth of the item digraph. This means that a node at a lower *level* —closer to the root—in a hierarchical path requires additional skills to those nodes at higher levels. Let us refer to this property as *set inclusivity* of the item digraph.

I. B. iv The Theory of Difficulty to Attribute Alignment. To hypothesize a prior difficulty for each item in the item digraph, a theory of difficulty to attribute alignment is required. The theory is stated as follows:

The number of true attributes measured by an item determines the item difficulty.

This theory is central to the GRAPH-CAT test scale and the assignment of mastery to the mastery graph. For brevity the theory will be referred to as *difficulty to attribute alignment. Difficulty to attribute alignment* allows for a prior difficulty estimate for each item characteristic curve (ICC) distribution in a root to sink path. In other words, because the cognitive complexity in a root to sink path decreases monotonically, an underlying continuous IRT test scale can be assigned for both prior difficulty and ability. The accuracy of prior difficulty of items and estimated ability of the learner are based on the accuracy of threshold assignment to each grade. A longitudinal approach to providing a *posterior* measure of item difficulty, based on student responses, is proposed as a direction for future research.

Given a high-quality test such as TIMMS, *difficulty to attribute alignment* is shown to hold (Mullis et al., 2008). Mullis et al. (2008) conducted a study whereby 8 booklets, with 163 items in mathematics, were administered to grade eight students from 42 different countries. An analysis of the correlation between the attributes measured by the item and the item difficulties, indicated that most of the variance in the items (76%) could be explained by the attributes identified by the Q-matrix (Tatsuoka, 2009, p.279). This means that the Q-matrices came close to accurately specifying the set of attributes that were both necessary and sufficient to providing a correct response to the items. Further, this indicates the attributes were of a similar granularity. The test scale based on *difficulty to attribute alignment* assumes that these two properties of the Q-matrix hold. The remainder of this section explores these two assumptions in more detail.

The first assumption is straightforward and states that the specified set of prerequisite attributes measured by an item is sufficient and every attribute in the set is necessary. Such a set of prerequisite attributes ensures that the item digraph is content-complete. The content-complete item digraph is discussed further under Directions for Future Research.

The second assumption is that *attribute granularity* is standardized where *attribute granularity* refers to a predetermined range of acceptable time required to master an attribute among a population of students at the same ability level. For example, suppose we have two attributes *a4* and *a5*. Attribute *a4* states a general skill: *the student will be able to work with simple fractions*. Meanwhile, attribute *a5* states a more specific skill: *the student will be able to add the numerators of two fractions with a common denominator*. It is clear that the ability to "work with simple fractions" is composed of many subtasks, where "add(ing) the numerators of two fractions" might be one of the subtasks. Therefore, to master *a5* would, on average, require significantly more time to master than *a4*. By changing *a5* to a more specific skill we come closer to keeping the attributes at a small and consistent granularity. An example of a more specific skill for *a5* might be: *the student will be able to express a fraction in reduced form*.

This assumption may require substantiation as, according to Rupp, Templin, and Henson (2010, p.53), unitization of attribute granularity does not make sense since cognitive complexity differs depending on ability. Therefore, they argue, we have no frame against which to determine the unit of standardization. The example they use to illustrate the ineffectiveness of unitizing attributes is that a grade eight item will be easier for a grade eight student than for a grade three student.

Contrary to the position taken by Rupp, Templin, and Henson (2010), cognitive complexity of a task must be examined at the ability of the learner as it is determined by the number of attributes required by the task. This can be seen by exploring the expected distribution of attributes and attribute mastery. According to the natural evolution of curriculum development, we expect the number of attributes measured by any learning outcome at a given grade to be approximately the same. We further expect that any student at a given grade has mastered approximately the same number of attributes. In other words, the number of mastered attributes at any grade should correlate highly with the number of measured attributes at any grade. To use a parallel example to that provided by Rupp, Templin, and Henson (2010), a grade eight item will be equally as easy for a grade eight student as a grade three item will be for a grade three student.

By satisfying the two assumptions of difficulty to attribute alignment, cognitive complexity of each learning outcome at a given grade comes closer to being standardized. While theoretically the two assumptions of the test scale are feasible, they are not expected to hold with current curriculum structures. If the two assumptions of the prior test scale do not hold, then the observed number of attributes an item measures is not expected to account for such a large portion of the variance in IRT item difficulty as that observed by Mullis et al. (2008).

I. B. v New Definitions Central to GRAPH-CAT. *Difficulty to attribute alignment* allows for a test scale to be constructed without student responses. To understand the test scale used by GRAPH-CAT a few new definitions must be introduced. To begin, let us examine each path of the item digraph.

A *hierarchical path* can be defined as follows: Given a root to sink path $P = (v0, e0, v1, e1, v2, e2 \ldots vL, eL)$ in an item digraph of depth L, a set of ordered pairs, H, is created where $H = \{ (v_0, b_0), (v_1, b_1), (v_2, b_2) \ldots (v_L, b_L)\}$. Each node (vertex) v_i of H is assigned a value of $b_i \epsilon \mathbb{R}$ only if $b_i < b_j$ whenever $i > j$, for all i and j, where $1 \leq i \leq L$ and $1 \leq j \leq L$ (see Figure 15).

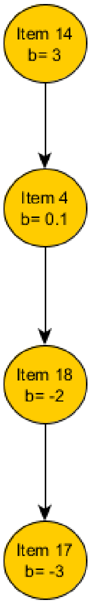


Figure 15. A conceptual diagram of the hierarchical path

Stated conceptually, higher level nodes—those nodes further from the root— are assigned lower values of IRT item difficulty. This allows for the construction of a test scale that is a partitioned IRT scale where both ability and difficulty are strictly increasing with grade. The administration of CAT to a hierarchical path using such a test scale will be referred to as a *path-CAT*.

Let b_i represent the highest difficulty value belonging to the node v_i for grade i in a hierarchical path of an item digraph. Given an ability value of

$\theta \epsilon \mathbb{R}$ on the continuous latent ability scale assigned to a hierarchical path, a *mastery path* is defined by the following *assignment of mastery*:

$$\mathrm{M}(v_i) = \begin{Bmatrix} 1 & iff & b_i < \theta \\ 0 & iff & b_i \geq \theta \end{Bmatrix}$$

for any grade level i where $1 \leq i \leq L$. It follows from *assignment of mastery*, the definition of a *hierarchical path*, and the transitive property of inequality that any node labelled with a 0 is always of a lower level than any node labelled 1 (Figure 16).

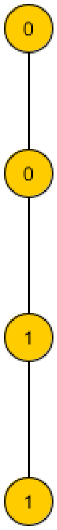


Figure 16. An example of a mastery path

The highest mastered outcome (HMO) represents the mastered node closest to the root in a mastery path. According to the BDF, the learner has mastered all attributes in the HMO in a mastery path. It follows that the node directly above it contains at least one attribute that is not mastered (Figure 17). The HMO in a mastery path will also be referred to as the *mastery node*. The adjacent node closer to the root contains the IRT ability (θ) and it will be referred to as the *ability node*. By the BDF the *mastery node* has a value of 1 and the *ability node* has a value of 0. In Figure 17 below, the ability node is at level 1 from the root and the mastery node is at level 2 from the root. The ability node is estimated using a path-CAT.

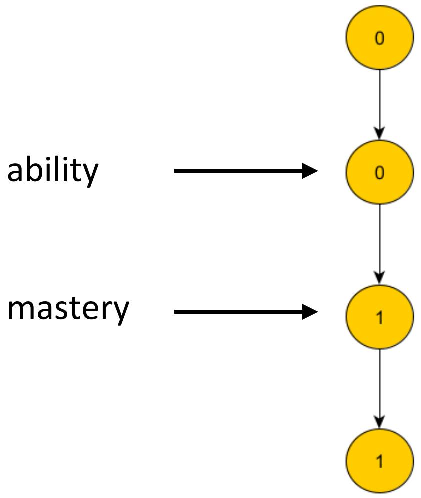


Figure 17. Ability versus mastery in a mastery path

A *mastery graph* is an item digraph where every node has been assigned a mastery value of 0 or 1 (Figure 18). Every student has exactly one true mastery graph for any given learning outcome. The mastery graph is the estimated output from an administration of GRAPH-CAT and it acts as the cognitive model for the student. It is from the mastery graph that the estimated knowledge state is derived.

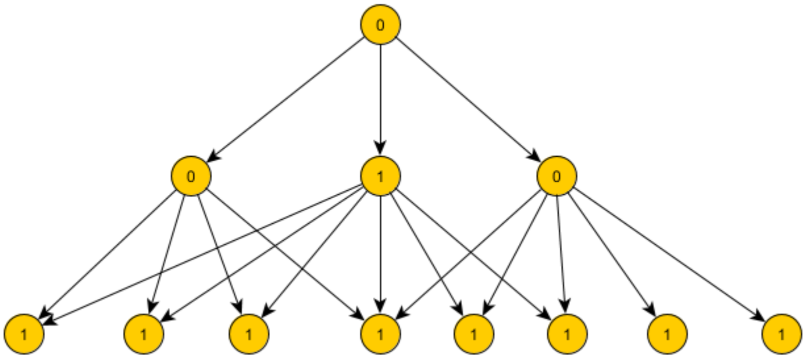


Figure 18. An example of a mastery graph

Using the item digraph for the Specific Outcome for Grade 6 *Percents* (Figure 5), a hierarchical path would resemble the highlighted path in Figure 19. The *b* value in Figure 19 refers to the upper threshold on the continuous IRT scale of a node according to the test scale. An alternative way to view the b value, is the highest item difficulty belonging to a node, if the item bank provided an item for an infinitesimally small granularity of item difficulty.

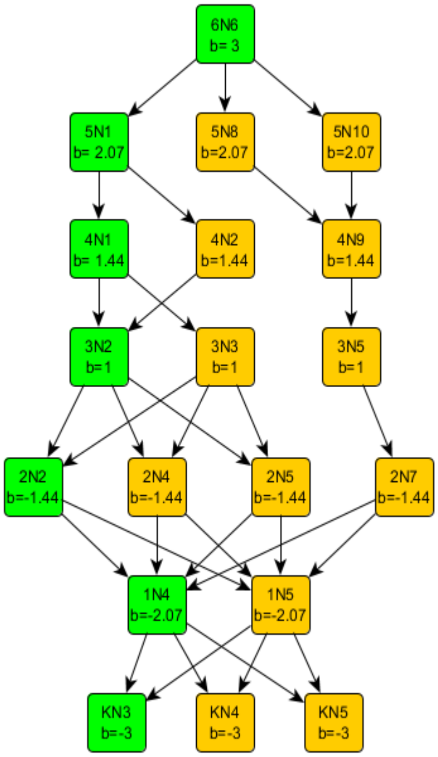


Figure 19 An example of a hierarchical path in the Grade 6: Number-Specific Outcome 6 (Percents) item digraph

Suppose that, in the Grade 6-*Percents* item digraph (Figure 19), a student has an ability on the IRT scale of θ= 0 for the highlighted hierarchical path. Then according to *assignment of mastery*, the hierarchical path in Figure 19 becomes the mastery path in Figure 20. If we suppose a student learning the Grade 6-*Percents* topic has an ability of $\theta = 0$ in every hierarchical path in Figure 19, then their mastery graph can be seen in Figure 21.

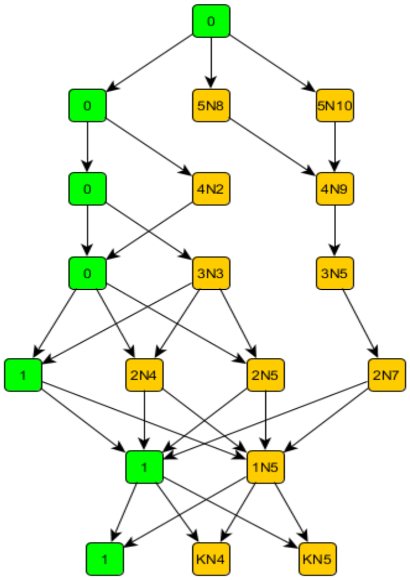


Figure 20 A mastery path associated with the hierarchical path in Figure 19 for a student with an IRT ability of 0

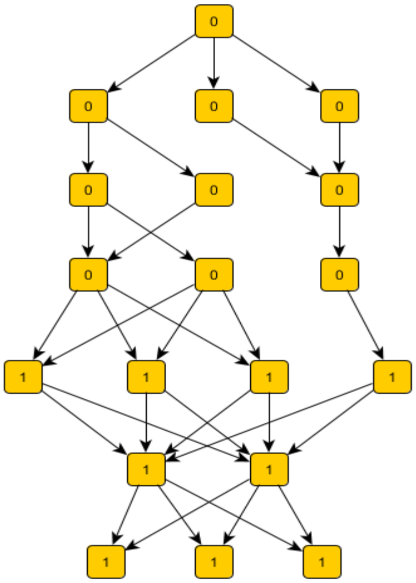


Figure 21. A mastery graph for a student with an IRT ability of 0 in every hierarchical path of the Grade 6-Percents item digraph in Figure 19

The notion of knowledge state is central to all diagnostic tests. According to Tatsuoka (2009), the original purpose of the Boolean-algebra formulated from a Q-matrix is to determine the universal set of knowledge states. Given a Q-matrix with n attributes, she infers that there are 2^n possible attribute mastery patterns (Tatsuoka, 2009, p 79). This value represents all possible assignments of 0 and 1 to each attribute to represent non-mastery and mastery, respectively.

There is an issue with this inference as the number of mastery graphs will generally be less than 2^n. Observe that the mastery graph is an alternative way to visualize the knowledge state (Figure 21). Assignment of mastery will always assign 0 to a lower level node than any node assigned 1. This places a restriction on the assignment of mastery to nodes so that a restriction is also placed on the assignment of mastery to attributes. As a result, the number of possible knowledge states is less than 2^n for any set of attributes with at least one precedence relationship.

I. B. i Partially Ordered Set Theory. Insight can be gained into the unobservable properties of the precedence relationship using *partially ordered set* (POSET) theory (Godin, Missaoui, & Alaoui, 1991). To understand a POSET, a *set* and a *binary relation* must first be defined. Intuitively, a *set* is a collection of objects that satisfy a certain property (Jech, 2013). The binary relation indicates that, for certain pairs in the set, one element precedes another. A *POSET* is a set, A, together with a *binary relation* on A. For all practical purposes, let A be the set of nodes of the item digraph. The binary relation acts on the set of attributes belonging to each node.

The following notation is described to assist with definitions A to E below. The notation $a \in A$ indicates object *a* is a member of set *A*. The subset notation, $A \subseteq B$, indicates that every object belonging to set *A* is also a member of set *B* (Youri, 2008). The cross-product notation, $A \times B$, denotes the set of all ordered pairs (a, *b*) such that *a* ∈ *A* and *b* ∈ *B* (Youri, 2008).

Definition A: Given a set A, a *binary relation* on A is a subset $B_R \subseteq A \times A$ (Levine, 2011)

Definition B: Given a *binary relation,* R, on a set A, R is reflexive iff (X, X) ∈ R for all X ∈ A (Levine, 2011)

Definition C: Given a *binary relation,* R, on a set A, R is transitive iff (X, Y) ∈ R and (Y, Z) ∈ R ⇒ $(\mathrm{X}, \mathrm{Z}) \in R$ (Levine, 2011)

Definition D: Given a *binary relation,* R, on a set A, R is antisymmetric iff $(X, Y) \in R$ and $(Y, X) \in R \Rightarrow X = Y$ (Levine, 2011)

Definition E: A set A with a binary relation $R \subseteq A\ x\ A$ is *partially ordered* if the elements of R are reflexive, transitive, and antisymmetric (*see Def A, B, C, and D*) (Levine, 2011)

The Hasse Diagram can be used to represent a POSET and gain a deeper understanding of the data and the knowledge associated with the data (Figure 22) (Godin, Missaoui, & Alaoui, 1991). According to Weisstein (2019), a Hasse diagram is a graphical representation of a POSET with an implied upward orientation. Each node of the Hasse diagram represents an element of the POSET.

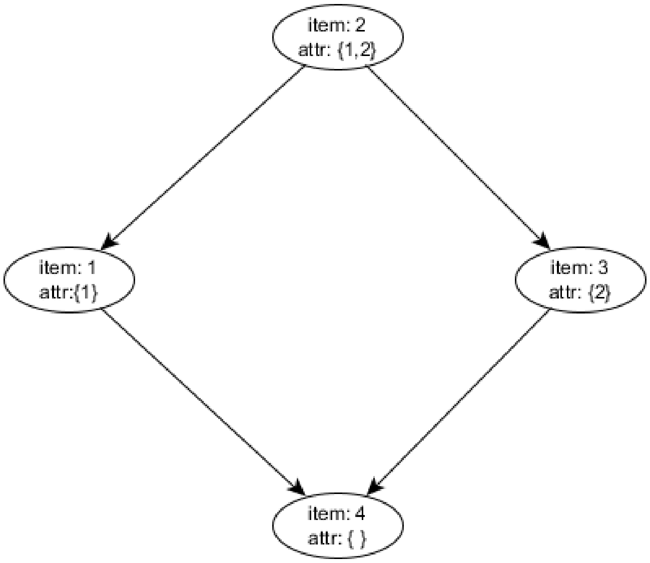


Figure 22. A Hasse diagram representation of the Q matrix in Table 1 (Yamada, 2008)

To understand the two rules of the Hasse diagram provided below, note that element z of a POSET (R, ⊆) is said to *cover* another element x provided that there exists no third element y in the POSET for which x$\subseteq y \subseteq z$ (Weisstein, 2019). With this understanding, a line segment is drawn between nodes according to the following rules:

1. For x and z in the POSET, if x$\subseteq z$ then the node corresponding to x occurs lower in the diagram than the node corresponding to z (Weisstein, 2019)

2. The line segment between the nodes corresponding to any two elements x and y is included in the diagram if and only if x *covers* y or y *covers* x (Weisstein, 2019)

A consequence of this definition is that the Hasse diagram represents the transitive relationship in as few edges as possible, as is required with set inclusion in the item digraph (Aho, Garey, & Ullman, 1972). Viewing the item digraph as a Hasse Digram may allow for further understanding of the precedence relationship and its associated knowledge. A deeper understanding of the structure of the item digraph also helps to introduce a natural test shortening technique, referred to here as Cascading Mastery Estimation (CME).

I. B. ii Cascading Mastery Estimation. CME helps address the problem of content magnitude for diagnostic testing by iteratively reducing the number of learning outcomes that must be tested. The general idea is that if the specified precedence relationships are accepted as valid then estimation of the *ability node* determines assignment of mastery and ability to all connected nodes.

Thus, upon discovering ability in each hierarchical path, test length is shortened by inferring mastery and non-mastery in other hierarchical paths.

Before describing CME in detail, let us put together the processes described so far. To begin, a path is selected and the path-CAT is administered. When the path-CAT ends, the IRT ability value (θ) is mapped to a node in a hierarchical path. The grade level of the ability node is used as the estimated grade ability node of the learner. The hierarchical path is then converted to a mastery path according to assignment of mastery. Finally, subsequent paths are shortened by using CME to infer mastery and non-mastery in other paths.

To validate the CME process, the following two propositions are proved. Definitions F and G are introduced to facilitate the proofs. For an intuitive understanding of CME, a concrete example from Grade 6 *Percents* is provided at the end of this section. The following proofs are not required to understand CME and they may be skipped if desired.

Definition F: Let R be a rooted digraph with root v_0. An *upwards subgraph* is a rooted subgraph of R, R', such that R' has exactly one sink, v_k, and every path P in R' starts at v_0 and ends at v_k(Figure 23) .

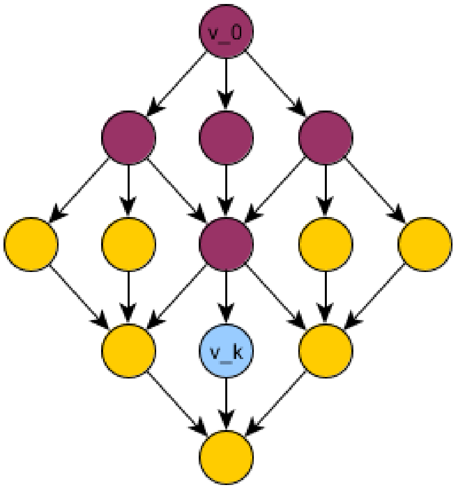


Figure 23. Illustration of an upwards subgraph

Definition G: Let R be a rooted digraph. A *downwards subgraph* is a rooted subgraph of R, R', with root at $v_0{}' \in R$ such that for every path P=$(v_0{}', e_0, v_1, e_1, \ldots\ e_{k-1}, v_k)$ in R', v_k is a sink node of R (Figure 24).

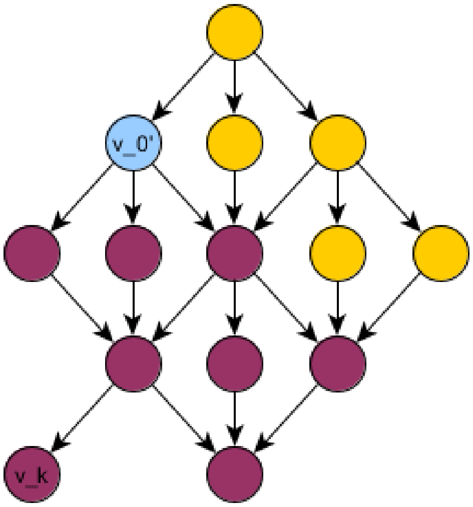


Figure 24. Illustration of a downwards subgraph

Proposition 1. If node z in an item digraph is mastered then for every node, v, in the downwards subgraph (Def G) with root at z we have BDF(A_v)=1 where A_v is the set of attributes belonging to v.

Proof: Let *z* be a mastered node in an item digraph R and let R′ be the *downwards subgraph* with root at *z*. Let P be any path in R' such that for any node v in P , A_v represents the set of attributes belonging to v.

For all $v_i \in P$ we have $level(z) \leq level\ (v_i)$ as z is the root of R′.

Therefore, for any node $v_i \in P$ it follows from set inclusivity of the item digraph that $A_{v_i} \subseteq A_z$. Since all attributes in A_z are mastered it follows that all attributes in A_{v_i} are mastered. As v_i was arbitrariliy selected, it follows from the BDF that $BDF(A_v) = 1$ for all $v \in P$. As P was arbitrarily selected, it follows that for every path P∈ R′ , we have *BDF*(A_v)=1 for every *v* ∈ *P*. Thus, for all nodes, *v,* in the downwards subgraph with root at z we have *BDF*(A_v)=1. ■

Proposition 2. If node z in an item digraph is an ability node then for every node, v, in the upwards subgraph (Def F) with sink at z we have BDF(A_v)=0 where A_v is the set of attributes belonging to v..

Let *z* be an ability node in an item digraph R and let R′ be the *upwards subgraph* with sink at *z*. Let *P* be any path in R′ such that for any node *v* in P, A_v represents the set of attributes belonging to v. For any node $v_i \in P$ it follows from set inclusivity that $A_z \subseteq A_{v_i}$ as $level(v_i) \leq level\ (z)$. Since *z* is an ability node, it follows from the BDF that there exists an attribute in A_z that is not mastered. By set inclusivity it follows that there exists an attribute in A_{v_i} that is not mastered. By the BDF we have $BDF(A_{v_i}) = 0$. As v_i was arbitrarily selected it follows that *BDF*(A_v)= 0 for all $v \in P$. As *P* was arbitrarily selected, it follows that for every P∈ R$'$ we have *BDF*(A_v)=0 for every v∈ P. Thus, for all nodes v in the upwards subgraph with sink at *z* we have *BDF* (A_v) = 0.■

Recall from the Grade 6-*Percents* example provided in Figure 19, that a student with ability $\theta = 0$ in the first path will have the mastery path displayed in Figure 25. It follows by *assignment of mastery* that they have mastered 2N2 while their level of ability is at 3N2 (Figure 25).

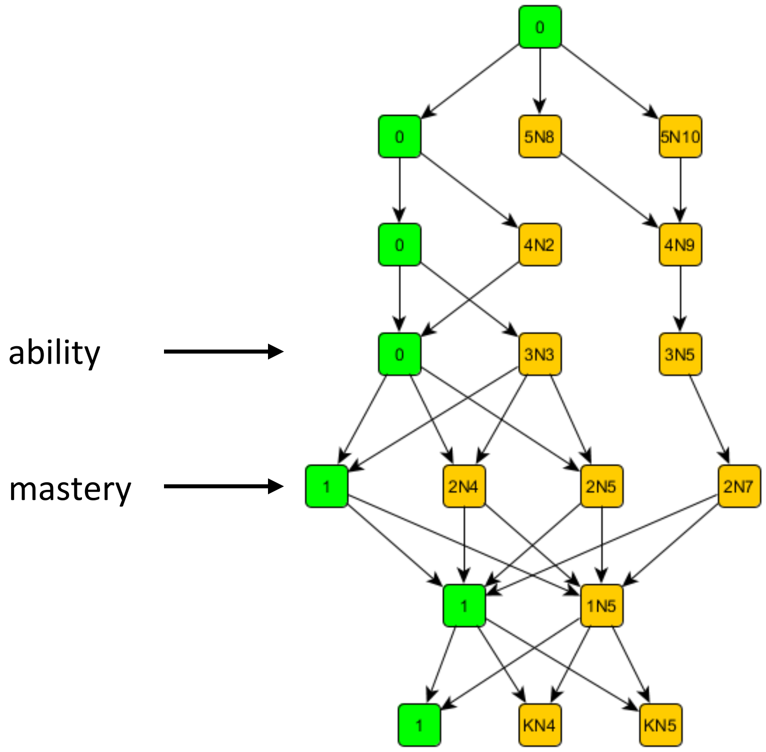


Figure 25. Ability and mastery of a student with IRT ability of 0 in path 1 of the Grade 6-Percents item digraph

As 2N2 is the highest mastered outcome in path 1, it follows from Proposition 1, that the student has also mastered 1N5, KN4, and KN5 as these are nodes in the *downwards subgraph* with root at 2N2 (Figure 26). Similarly, as 3N2 is the ability node, it follows from Proposition 2 that the student has not mastered 4N2 as this is a node in the *upwards subgraph* with sink at 3N2.

Notice that the yellow nodes in Figure 25 are the nodes that remain to be administered. When Figure 25 and Figure 26 are compared, it is clear that CME reduces the number of learning outcomes that must be tested in subsequent paths.

Thus, each single estimation of ability can be viewed as having a cascading effect on the item digraph, where the estimation of ability in one path spreads to nodes in other paths (Figure 26). It also follows that increasing the number of edges between a given set of nodes will allow the test to converge faster.

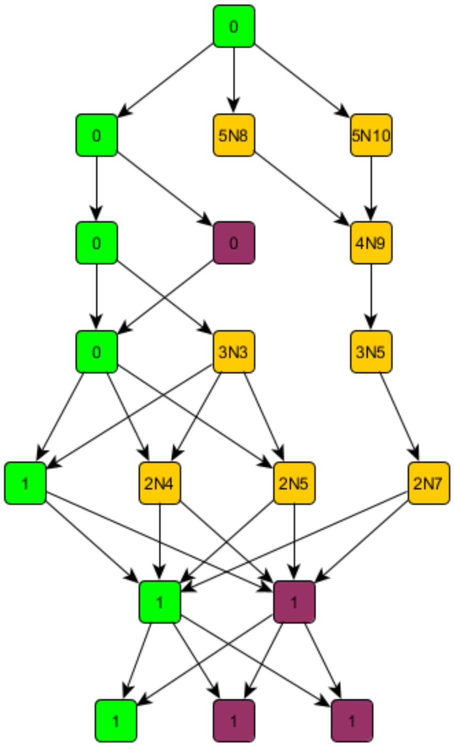


Figure 26. Mastery node labelling according to Cascading Mastery Estimation (CME) as determined by the ability and mastery in Figure 25

I. C Concluding Remarks on CD-CAT with the Item Digraph

The item digraph allows us to use a continuous measure of ability to determine attribute mastery according to the item response theory framework. By using the item digraph to locate the ability node of a learner using a traditional CAT, we are determining a personalized zone of proximal development (ZPD) with an associated standard error of measurement (Shabani et al., 2010). Recall, that the ZPD is a learning space where the content is neither too easy nor too difficult for the student (Shabani et al., 2010). As the ability node is the most remedial non-mastered node, intervening learning resources aligned with the ability nodes target what can be seen in Figure 27 as the ZPD.

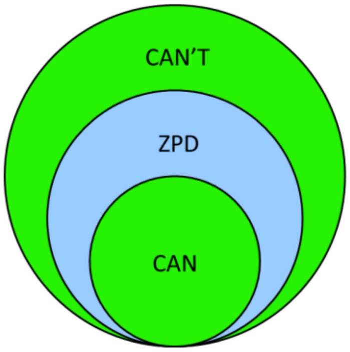


Figure 27. A diagrammatic representation of the Zone of Proximal Development adapted from Kym Buchanan

II. Problem Statement

Previous implementations of CD-CAT tend to replace a single ability measure, such as IRT θ, with a measure of attribute mastery. These algorithms classify examinees into their true knowledge state based on novel statistical measures of response classification (Yamada, 2008). Previous CD-CATs proved to be strong in accuracy and test length. A major shortcoming of these algorithms, however, is in their stochastic item selection criteria (Falmagne, Cosyn, Doignon, & Thiéry, 2006; Tatsuoka, 2009; Yamada, 2008).

III. Research Questions

In this simulation study of GRAPH-CAT, the following questions are posed: (a) what is the efficiency of GRAPH-CAT? (b) what are the item parameters required to ensure GRAPH-CAT is both valid and reliable? (c) what are the test parameters required to ensure GRAPH-CAT is useable in terms of accuracy and test length?

IV. Thesis Statement

Combining the item digraph with traditional CAT based in IRT allows for a more robust measure of attribute mastery that provides an estimation of attribute mastery with an associated standard error of measurement. Representation of the attribute hierarchy as an item digraph fully realizes the connectivity of the attributes to improve test efficiency over previous CD-CAT implementations. The item digraph approach further improves on previous CD-CATs by allowing for item administration that is both sequential and systematic by content and difficulty.

V. Methods

This section reviews how the concepts introduced so far are used in the GRAPH-CAT simulation. The topics covered include a description of the simulation and measurement accuracy values. Notice that each grade in a hierarchical path occurs exactly once and each outcome belonging to a grade in the path is assigned a unique set of attributes. This allows for an estimation of knowledge state based on the mastery graph.

The percent matching is calculated between attribute mastery in the so-called *real knowledge state* and the estimated knowledge state. Yamada (2008) considers the *real knowledge state* to be the attribute mastery derived from real learner responses gathered from a full-length test containing all the items. This study considers real knowledge state to be that generated by simulation. Learner responses are then generated from the knowledge state using Monte Carlo simulation (MCS).

MCS is conducted in the programming language of Python to provide estimated measures on average test length and precision of GRAPH-CAT as measured by classification accuracy (Van Rossum & Drake, 2011). A strength of using MCS over other methods is that real responses are not required. This is important as real student response data used in the Yamada (2008) study is not available. Learner responses are simulated using difficulty and discrimination of the item parameters from the Yamada (2008) item bank. The Yamada (2008) study compares a CD-CAT based in the RSM (RSM-CAT) to a CD-CAT based in POSET (POSET-CAT).

V. A Simulation Variables

Diagnostic testing places stronger interest in remedying academic weaknesses than on estimating ability. For this reason, it is preferable for an academic diagnostic test to underestimate rather than overestimate student ability (McGlohen & Chang, 2008; Tatsuoka, 2009). In other words, more emphasis is placed on estimating non-mastered skills. Under the proposed approach, it is possible to shift the estimated ability node to a higher level that is further from the root. The $delta$ (Δ) value is used to shift the estimated ability (θ) downwards by a factor between -1 and 0 of the standard error of measurement, as estimated by a path-CAT. This increases the frequency of estimated grades per path that are lower than the real grade.

In the simulations, *n* is the number of items per outcome in the item bank. The *maxlen* value is the maximum test length set as the stopping rule for the test whereas the *len* value is the average number of items administered. The *grDIFF* value is the distance from the root at which all attributes are mastered.

V. B Simulating and Estimating the Learner

The ability for each *grDIFF* is simulated based on a random selection of non-mastered attributes from the level above *grDIFF (*$1 \leq \text{grDIFF} \leq 4$) (see Table 4). All attributes at the *grDIFF* level are labelled as mastered. Assignment of mastery to outcomes is then performed using the *BDF*. The following steps are used to simulate knowledge state of the learner:

1. Select a level of learner ability
2. Set all attributes that occur below the level of ability as mastered
3. From the set A of attributes that first occur at the ability level, randomly select 1 to |A| attributes to be non-mastered
4. Use the BDF to assign mastery to the nodes of the item digraph

The steps for administration of a path- CAT and simulating learner responses are based on the Monte Carlo procedure. The steps can be summarized as follows:

1. Select a path in the item digraph
2. Set the estimated IRT ability of the learner to be at the middle of the test scale for the path-CAT
3. Select an item using the maximum likelihood estimation function
4. Given ability (θ) and item difficulty (b), calculate the probability (P) that the learner will answer correctly based on the 2-PL IRT model
5. Generate a random number between 0 and 1
6. If the random number is greater than P, then set the response to this item to 0 (incorrect). Alternatively, if the number is less than or equal to P, then set the response to this item to 1 (correct)
7. Repeat steps 3 to 6 until the stopping condition for the path is met
8. Once the estimated ability node is determined, update the state of the item digraph to maintain the integrity of a mastery graph structure
9. Repeat steps 2 to 8 for every path until the test stopping condition is met

The MLE function is used as the maximum likelihood estimation function of ability for the path CAT. The Brute Force Method is used to discover the maximum of the MLE.

V. C Simulating Population Parameters

The Yamada (2008) simulation uses data from a simulation experiment performed by Tatsuoka (1984) based on a full-length 40 item test administered to 536 junior high students. As the response data is unavailable, examinee responses are simulated to match as closely as possible to ability distributions used by Yamada (2008). The original ability distribution included approximately 38% of examinees that scored either very high or very low on the test. More accurately, 20.71% (111) of examinees scored above or at 36/40= 0.90 and 16.98% (91) scored below or at 5/40= 0.13. To simulate these distributions, test scores are assumed to reflect attribute mastery.

Under this assumption, high ability students are to have mastered six or seven (0.90 ≤) attributes (see Table 4 below). If the student has mastered all seven attributes then they are drawn from *grDIFF*=1. If the student has mastered six attributes then they are drawn from *grDIFF*=2. Those students at *grDIFF*=2 are split into two categories. The first category at *grDIFF*=2 has mastered six attributes (see 2A in Table 4). The second category at *grDIFF*=2 has mastered four or five of the attributes (2B in Table 4).

Similarly, low ability students are to have mastered one or zero attributes (≤0.13). Those students at grDIFF=3 are split into two categories where the first category has mastered one attribute and the second category has mastered two or three attributes. The simulated population sizes for each ability level are shown in Table 4 below. To maintain the proportion of high and low ability students, a total population size of N=538 is simulated.

Table 4 *Population Size of Each Simulated Population (N=538)*

grDIFF	**Mastery Grade**	**Population Size**	**Number of Attributes Mastered**
4	1	46	0
3B	2	46	1
3A	2	167	2 or 3
2B	3	167	4 or 5
2A	3	56	6
1	4	56	7

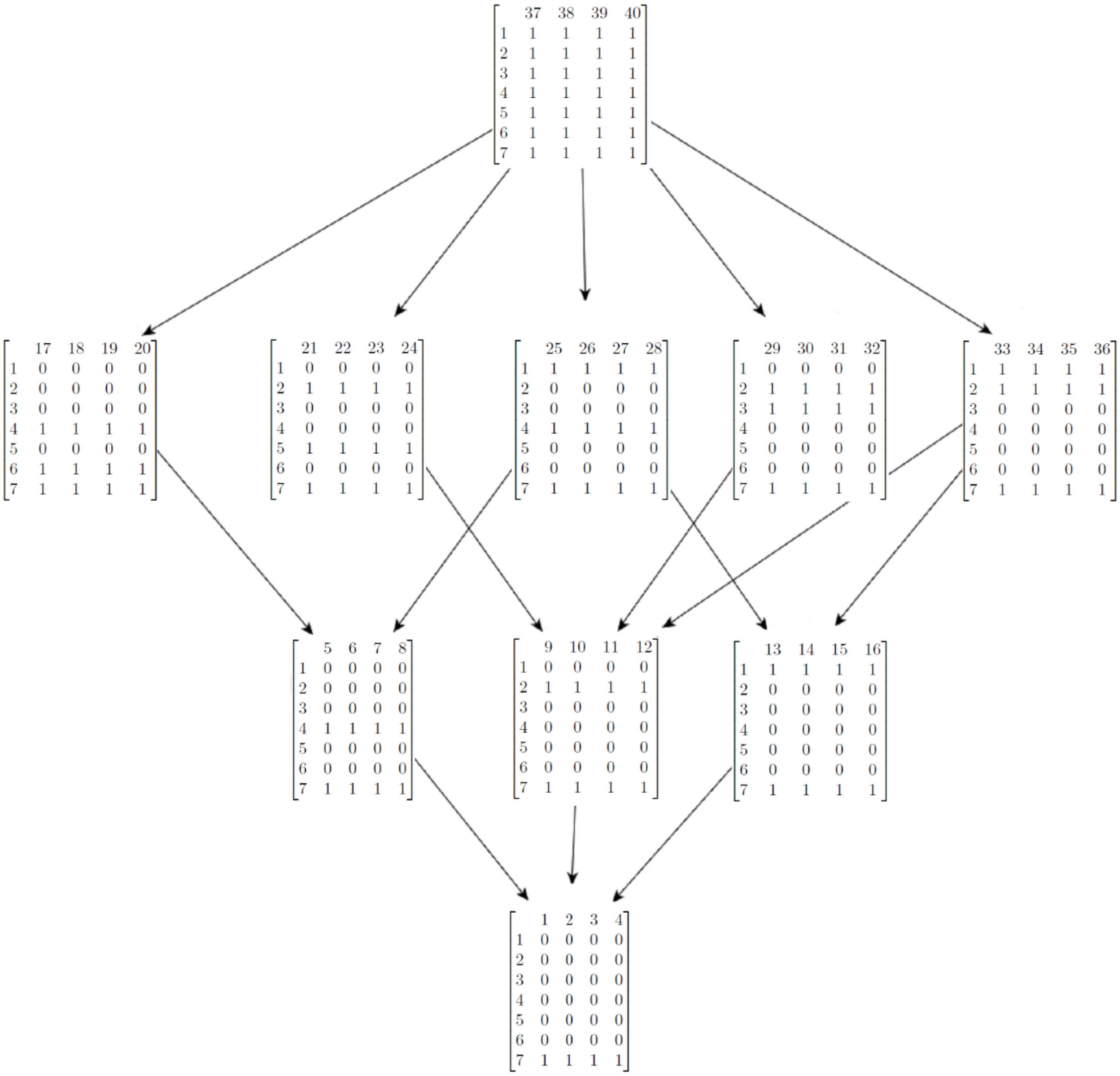
Note: Each population is uniformly distributed with a mean of 1≤ $grDIFF$ ≤ 4.

Similarly, an item bank that closely resembles that used by Yamada (2008) is simulated. Yamada (2008) derived difficulty and discrimination parameters from the test results based on student responses. Every discrimination value of every item is set to be the average of the items in the Yamada (2008) item bank (*a*= 1.78). The IRT difficulty values of the items are based on the test scale derived from difficulty to attribute alignment (Table *5*). The attributes measured by each item in the simulated item bank are the same as the attributes measured by each item in the Yamada (2008) item bank, with the exception of the root node.

Table 5 Parameters of Simulated Item Bank

Item#	IRTb	Attributes	Item#	IRTb	Attributes
1	-12.00	7	**21**	-4.89	2, 5, 7
2	-11.11	7	**22**	-4.00	2, 5, 7
3	-10.22	7	**23**	-3.11	2, 5, 7
4	-9.33	7	**24**	-2.22	2, 5, 7
5	-8.44	4,7	**25**	-4.89	1, 4, 7
6	-7.55	4,7	**26**	-4.00	1, 4, 7
7	-6.67	4,7	**27**	-3.11	1, 4, 7
8	-5.78	4, 7	**28**	-2.22	1, 4, 7
9	-8.44	2, 7	**29**	-4.89	2, 3, 7
10	-7.55	2, 7	**30**	-4.00	2, 3, 7
11	-6.67	2, 7	**31**	-3.11	2, 3, 7
12	-5.78	2, 7	**32**	-2.22	2, 3, 7
13	-8.44	1, 7	**33**	-4.89	1, 2, 7
14	-7.55	1, 7	**34**	-4.00	1, 2, 7
15	-6.67	1, 7	**35**	-3.11	1, 2, 7
16	-5.78	1, 7	**36**	-2.22	1, 2, 7
17	-4.89	4, 6, 7	**37**	0.44	1, 2, 3, 4, 5, 6, 7
18	-4.00	4, 6, 7	**38**	4.88	1, 2, 3, 4, 5, 6, 7
19	-3.11	4, 6, 7	**39**	9.33	1, 2, 3, 4, 5, 6, 7
20	-2.22	4, 6, 7	**40**	13.77	1, 2, 3, 4, 5, 6, 7

The root node has been added to the item digraph as such a unifying outcome was not present in the item bank used by Yamada (2008) (Figure 28). The root node represents the learning outcome being assessed by the test. The item digraph derived from the parameters of the Yamada (2008) item bank has depth 4 and degree 5 (Figure 28).



$$\left[\begin{array}{ccccc} & 37 & 38 & 39 & 40 \\ 1 & 1 & 1 & 1 & 1 \\ 2 & 1 & 1 & 1 & 1 \\ 3 & 1 & 1 & 1 & 1 \\ 4 & 1 & 1 & 1 & 1 \\ 5 & 1 & 1 & 1 & 1 \\ 6 & 1 & 1 & 1 & 1 \\ 7 & 1 & 1 & 1 & 1 \end{array}\right]$$

$$\left[\begin{array}{ccccc} & 17 & 18 & 19 & 20 \\ 1 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 \\ 4 & 1 & 1 & 1 & 1 \\ 5 & 0 & 0 & 0 & 0 \\ 6 & 1 & 1 & 1 & 1 \\ 7 & 1 & 1 & 1 & 1 \end{array}\right] \quad \left[\begin{array}{ccccc} & 21 & 22 & 23 & 24 \\ 1 & 0 & 0 & 0 & 0 \\ 2 & 1 & 1 & 1 & 1 \\ 3 & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 \\ 5 & 1 & 1 & 1 & 1 \\ 6 & 0 & 0 & 0 & 0 \\ 7 & 1 & 1 & 1 & 1 \end{array}\right] \quad \left[\begin{array}{ccccc} & 25 & 26 & 27 & 28 \\ 1 & 1 & 1 & 1 & 1 \\ 2 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 \\ 4 & 1 & 1 & 1 & 1 \\ 5 & 0 & 0 & 0 & 0 \\ 6 & 0 & 0 & 0 & 0 \\ 7 & 1 & 1 & 1 & 1 \end{array}\right] \quad \left[\begin{array}{ccccc} & 29 & 30 & 31 & 32 \\ 1 & 0 & 0 & 0 & 0 \\ 2 & 1 & 1 & 1 & 1 \\ 3 & 1 & 1 & 1 & 1 \\ 4 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 \\ 6 & 0 & 0 & 0 & 0 \\ 7 & 1 & 1 & 1 & 1 \end{array}\right] \quad \left[\begin{array}{ccccc} & 33 & 34 & 35 & 36 \\ 1 & 1 & 1 & 1 & 1 \\ 2 & 1 & 1 & 1 & 1 \\ 3 & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 \\ 6 & 0 & 0 & 0 & 0 \\ 7 & 1 & 1 & 1 & 1 \end{array}\right]$$

$$\left[\begin{array}{ccccc} & 5 & 6 & 7 & 8 \\ 1 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 \\ 4 & 1 & 1 & 1 & 1 \\ 5 & 0 & 0 & 0 & 0 \\ 6 & 0 & 0 & 0 & 0 \\ 7 & 1 & 1 & 1 & 1 \end{array}\right] \quad \left[\begin{array}{ccccc} & 9 & 10 & 11 & 12 \\ 1 & 0 & 0 & 0 & 0 \\ 2 & 1 & 1 & 1 & 1 \\ 3 & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 \\ 6 & 0 & 0 & 0 & 0 \\ 7 & 1 & 1 & 1 & 1 \end{array}\right] \quad \left[\begin{array}{ccccc} & 13 & 14 & 15 & 16 \\ 1 & 1 & 1 & 1 & 1 \\ 2 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 \\ 6 & 0 & 0 & 0 & 0 \\ 7 & 1 & 1 & 1 & 1 \end{array}\right]$$

$$\left[\begin{array}{ccccc} & 1 & 2 & 3 & 4 \\ 1 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 \\ 6 & 0 & 0 & 0 & 0 \\ 7 & 1 & 1 & 1 & 1 \end{array}\right]$$

Figure 28 The item digraph derived from the Yamada (2008) item bank. The items and attributes for each learning outcome are indicated within each node.

Forty items are used in the Yamada (2008) study and the item digraph contains ten specific outcomes. The number of items per outcome (*n*) is set to be equal so that *n*=40/10=4. This produces an item bank having forty items with four items per outcome (Table 5).

V. D Measuring Accuracy

To measure knowledge state classification using GRAPH-CAT, the number of correctly classified attributes in the estimated mastery graph are compared to the mastered attributes in the real mastery graph, using the *Jaccard Index*:

$$|E \cap R| \,/\, |E \cup R|$$

where *E* is the set of estimated mastered attributes and *R* is the set of real mastered attributes (Boyce & Ellison, 2005). The Jaccard Index is known to be a reliable measure of percent matching between two binary sets (Boyce & Ellison, 2005). In this formula, the numerator represents the intersection of the two sets and the denominator represents the union.

Of primary importance in the results are the measures of accuracy and test length as these are the key indicators of test usability (Yamada, 2008). Also provided are the average attribute mastery and the outcome mastery. Several different varieties of these measurements are calculated: *attmatch*, *nonattmach*, *HMOmatch*, *HMOmatch2*, *aborbel*, *ocmatch*, *nonocmatch*, *ocbygr, attbygr,* and *nonattbygr*.

The *attmatch* value provides a measure of the *knowledge state* accuracy and it is equivalent to the *knowledge state proportion match* provided by Yamada (2008). It measures the accuracy of the estimated mastered attributes. At a larger granularity, *ocmatch* measures accuracy of the estimated mastered learning outcomes. This measure of frequency match (fm) is unique to GRAPH-CAT as sets of attributes are estimated through the organization of attributes into learning outcomes.

On the other hand, *nonattmatch* provides a measure of the attributes that the learner has not mastered whereas *nonocmatch* is a measure of learning outcomes the learner has not mastered. Neither of these measures are provided by the Yamada (2008) study, but they can be seen as a complementary form of knowledge state. Rather than creating a model of what the learner knows, a model of what the learner does not know is created. The non-attribute match is, arguably, more useful than knowledge state, as it directly addresses the goal of determining what the learner does not know.

Notice that proportion matching of outcome mastery and attribute mastery are not expected to be the same. As an example, consider a student of the Grade 3 learning outcome in Figure 29 who has the outcome mastery of {GRADE2A, GRADE1A, GRADE1B, GRADE2B, GRADE1C, GRADE1D}.

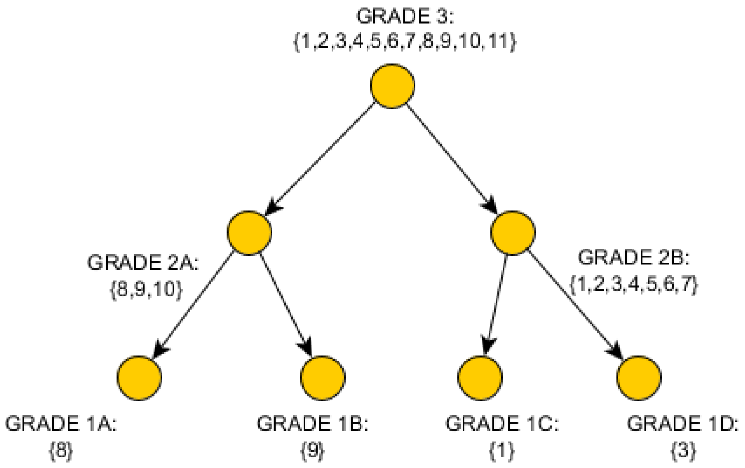


Figure 29. An item digraph with imbalanced attribute assignment

If GRAPH-CAT underestimates ability in two hierarchical paths by one grade, then the estimated outcome mastery might be {GRADE2A, GRADE1A, GRADE1B, GRADE1C, GRADE1D}. In this example, the outcome match of (*ocmatch* =5/6= 83%) and the attribute match of (*attmatch* =5/10= 50%) are not the same. As this demonstrates, the non-standardized item digraph may lead to estimations with high outcome matches and low attribute matches or vice-versa. For this reason, a distinction is made between outcome match and attribute match.

The highest mastered outcome frequency match measures grade level mastery accuracy per path (*HMOmatch*). In the results presented, the estimated *HMOmatch* value is considered to be accurate if the grade of the estimated highest mastered outcome (HMO) in a path is at most one grade below the real HMO of the learner. Similarly, *HMOmatch2* is considered to be accurate if the grade of the estimated HMO in a path is at most two grades below the real HMO of the learner. In terms of usability *HMOmatch*, *HMOmatch2*, and *aborbel* provide a measure of GRAPH-CAT navigability. Navigability indicates the ease with which a learner would be able to find appropriate intervening learning resources given exposure to the full database of learning resources and an initial recommendation by the test results.

The by-grade measures provide a measurement of accuracy on average in the universe of an individual grade. The outcome by-grade measure (*ocbygr)* indicates the accuracy of estimated outcome mastery on average in the universe of each grade. Similarly, *attbygr* provides a measure of attribute mastery estimation in the universe of each grade (Figure 30). See Table 6 for a summary of all measures.

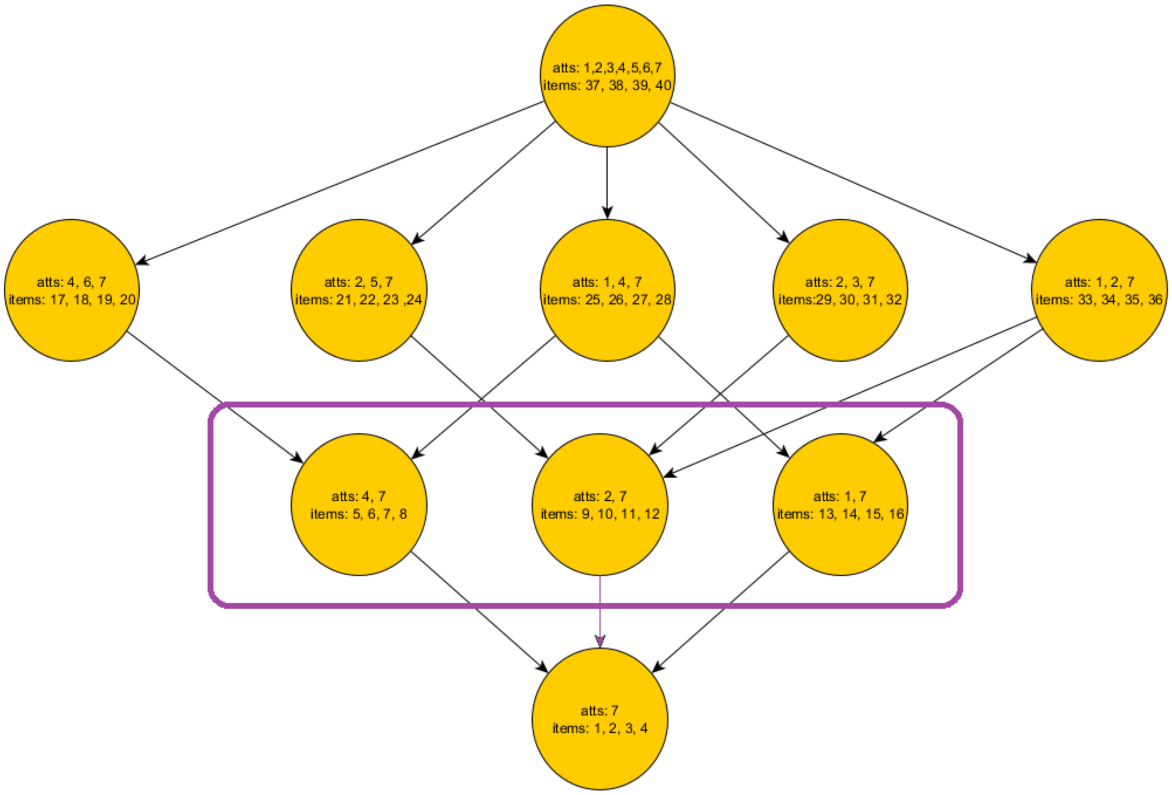


Figure 30. Diagrammatic representation of attbygr and ocbygr

Table 6 *A Description of the Frequency Match Categories*

attmatch	the fm between the mastered attributes belonging to each specific outcome
nonattmatch	the fm between the nonmastered attributes belonging to each specific outcome
aborbel	the fm between the estimated highest grade per path and the real grade when the difference is 1, 0, or -1.
hmomatch	the frequency match (fm) between the highest mastered grade per hierarchical path when the difference is either 0 or -1.
hmomatch2	the fm between the highest mastered grade per hierarchical path when the difference is either 0, -1, or -2.
ocmatch	the overall match between estimated mastered outcomes and real mastered outcomes
nonocmatch	the overall match between estimated non-mastered outcomes and real non-mastered outcomes
ocbygr	the average mastered outcome match per grade
attbygr	the average mastered attribute match per grade

Note: These values are used to provide a measure of GRAPH-CAT efficiency

V. E Simulation Conditions

The Maximum Likelihood Estimation function is used with the stopping rule for each path-CAT set to a minimum standard error (SE) of 0.20. The path shortening effects of CME require that semi-complete paths end before the SE error is reached. This will predictably lead to a test that ends with a higher SE.

Two test stopping rules are tested in these simulations. The first simulation ends the test after every outcome has been assigned mastery. The second simulation ends the test once a maximum of ten items has been administered.

VI. Results

VI. A Simulation 1: Test Stopping Rule of Complete Outcome Mastery Assignment

Using a test stopping rule of complete outcome mastery-assignment, the GRAPH-CAT test achieves a weighted average attribute match of 92% ($\overline{attmatch} = 0.92$), across the four grade level abilities in twenty items with SE= 0.33 (see Table 7 below). The estimated theta ability (θ) has not been shifted downwards (Δ=0.00 of the SE). The matching attributes on average per grade is 95% ($\overline{attbygr} = 0.95$). The $\overline{attbygr}$ measure indicates a high percent matching between intervening learning resources and learner ability.

Table 7 The Attribute Frequency Matching Values for the Yamada (2008) Item Bank Using a Test Stopping Rule of Complete Outcome Mastery Assignment

grDIFF	len	attmatch	nonattmatch	attbygr	nonattbygr	SE
1	20	99%	96%	100%	99%	0.35
2A	20	97%	92%	99%	97%	0.36
2B	22	94%	89%	95%	96%	0.35
3A	21	84%	89%	90%	93%	0.34
3B	20	89%	95%	94%	97%	0.33
4	11	100%	100%	100%	100%	0.23
weighted AVE	**20**	**92%**	**92%**	**95%**	**96%**	**0.33**

Note: The path-CAT runs by MLE using a stopping condition of *SE*=0.20 and the test achieves average SE=0.33. The number of items in the item bank is *n*=4 per learning outcome, $\Delta = 0.00$.

The average outcome match (*ocmatch*) is 91% indicating the accuracy of overall outcome match. In the case where learning resources are delivered to the student in cumulative batches across all grades of outcome deficiency, *ocmatch* indicates the accuracy of learning resource alignment to student ability (Table 8). For 92% of learner-paths (*HMOmatch*=0.92*)*, the estimated grade level in the path is less than the real grade per path by at most one grade (Table 8). This is a measure of the closeness of average estimated grade level ability to real ability, when the test estimation is below the real ability. In 94% of the learner-paths, the estimated grade is less than the real average grade per path by at most two grade levels (*HMOmatch2*=0.94). The HMO values indicate that the path-CAT is able to closely estimate the grade level mastery of the learner using the MLE function.

Table 8 The MLE Outcome Frequency Matching Values for the Simulated Yamada (2008) Item Bank Using a Test Stopping Rule of Complete Outcome Mastery Assignment

grDIFF	**len**	**ocmatch**	**nonocmatch**	**HMOmatch**	**HMOmatch2**
1	20	99%	96%	99%	100%
2A	20	97%	91%	97%	98%
2B	22	95%	88%	95%	96%
3A	21	82%	89%	85%	88%
3B	20	89%	96%	89%	89%
4	11	100%	100%	100%	100%
weighted AVE	**20**	**91%**	**91%**	**92%**	**94%**

Note: The path-CAT runs by MLE using a stopping condition of SE=0.20 and the test achieves SE=0.33. The number of items in the item bank is n=4 per learning outcome, Δ= 0.00.

The average *aborbel* value is 98% meaning that on average per path, the estimated grade level is almost always one grade above or one grade below the real grade. The *aborbel* value is high for all ability levels, ranging between 95%-100% (Table 9). The *aborbel* value is the best measure of resource navigability.

The average *diff* value provides a measure of the per path difference between real and estimated grade level (Table 9). The estimated grade level ability per path is, on average, very close to the real grade level ability (*diff*= 0.00). On average, 6% of paths are estimated to be at a grade ability higher than the real ability grade (*above* = 0.06).

Table 9 The Average Difference Between Estimation Grade and Real Grade for the Simulated Yamada (2008) Item Bank Using a Test Stopping Rule of Complete Outcome Mastery Assignment

grDIFF	len	ocbygr	nmocbygr	aborbel	diff	above
1	20	100%	98%	99%	-0.02	0.00
2A	20	99%	96%	99%	-0.03	0.02
2B	22	97%	96%	99%	0.00	0.04
3A	21	89%	93%	95%	0.00	0.12
3B	20	94%	97%	99%	0.10	0.11
4	11	100%	100%	100%	0.00	0.00
weighted AVE	**20**	**95%**	**96%**	**98%**	**0.00**	**0.06**

Note: The path-CAT runs by MLE using a stopping condition of SE=0.20. The number of items in the item bank is n=4 per learning outcome, Δ= 0.00.

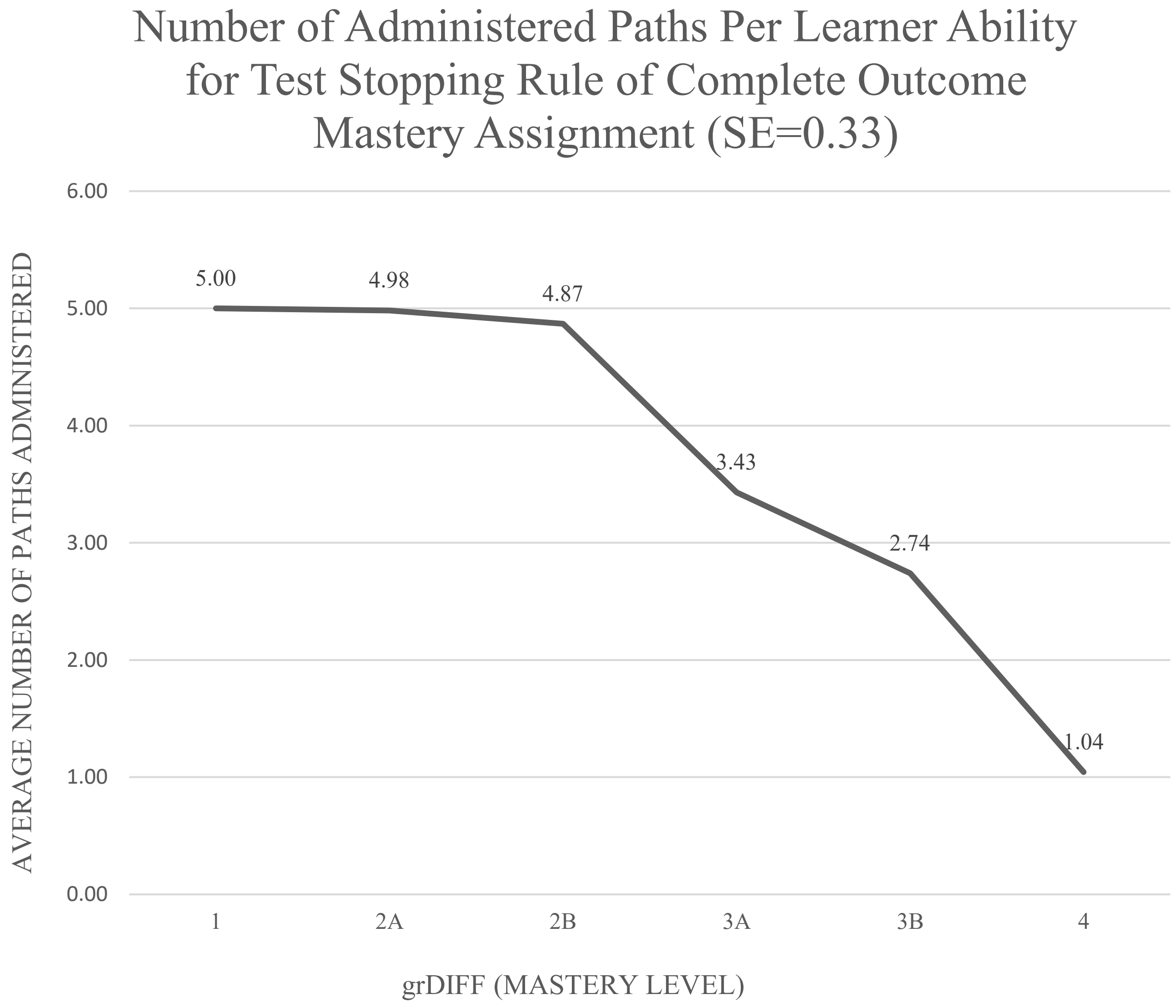


Figure 31 Number of paths administered for a test stopping rule of complete outcome mastery assignment. Note that grDIFF indicates mastery level where grDIFF=1 is high ability and grDIFF=4 is low ability (Table 4).

VI. B Simulation 2: Test Stopping Rule of Maximum Ten Items

When test length is limited to ten items the GRAPH-CAT achieves a weighted average attribute match of 83% in an average of ten items with SE= 0.38 (Table 10). The estimated theta ability (θ) is not shifted downwards (Δ=0.00). The matching attributes on average per grade remain high at 89% (*attbygr*).

Table 10 The Attribute Frequency Matching Values for the Yamada (2008) Item Bank Using a path-CAT Stopping Rule of Maximum Ten Items

grDIFF	len	attmatch	nonattmatch	attbygr	nonattbygr	SE
1	10	77%	36%	88%	80%	0.41
2A	10	83%	64%	89%	89%	0.39
2B	10	89%	85%	90%	94%	0.36
3A	10	73%	84%	85%	91%	0.41
3B	10	86%	95%	92%	96%	0.42
4	9	98%	99%	99%	99%	0.26
weighted AVE	**10**	**83%**	**80%**	**89%**	**92%**	**0.38**

Note: The path-CAT runs by MLE using a stopping condition of *SE*=0.20 and the test achieves average SE=0.38. The number of items in the item bank is *n*=4 per learning outcome, Δ= 0.00.

The average outcome match (*ocmatch*) is 83% indicating the accuracy of learning resources delivered to the student in cumulative batches (Table 11). The average *HMOmatch* is 92% and *HMOmatch2* is 95%. The *HMO* values indicate that the per path estimation of ability is highly accurate to within two grades below the real ability. As the path-CAT is ended with lower accuracy by limiting the test length to ten (*SE*=0.38), the estimated outcome matches are understandably lower than those with a stopping rule of full outcome mastery estimation (*SE*=0.33).

Table 11 The MLE Outcome Frequency Matching Values for the Simulated Yamada (2008) Item Bank Using a Test Stopping Condition of Maximum Ten Items

grDIFF	**len**	**ocmatch**	**nonocmatch**	**HMOmatch**	**HMOmatch2**
1	10	82%	36%	100%	100%
2A	10	87%	64%	99%	99%
2B	10	90%	83%	96%	98%
3A	10	70%	84%	85%	90%
3B	10	86%	96%	88%	88%
4	9	98%	100%	96%	96%
weighted AVE	**10**	**83%**	**79%**	**92%**	**95%**

Note: The path-CAT runs by MLE using a stopping condition of SE=0.20 and the test achieves SE=0.38. The number of items in the item bank is n=4 per learning outcome, $\Delta = 0.00$.

The average difference between estimated grade and real grade remains high across all categories of estimated grade (*ocbygr*=0.91, *nmocbygr*= 0.91, *aborbel*= 0.97) (Table 12). These values can be used as a measure of effectiveness of formative feedback when intervening learning resources are provided in batches by grade. The outcome by grade match (*ocbygr* =0.91) and non-mastered outcome by grade match (*nmocbygr*= 0.91) indicate that intervening resources are well-aligned with learner ability, notwithstanding a lower overall attribute match than that achieved when the test converges using complete outcome mastery assignment.

Table 12 The Average Difference Between Estimation Grade and Real Grade for the Simulated Yamada (2008) Item Bank Using a Test Stopping Condition of Maximum Ten Items

grDIFF	**len**	**ocbygr**	**nmocbygr**	**aborbel**	**diff**	**above**
1	10	93%	80%	100%	-0.26	0.00
2A	10	94%	89%	100%	-0.16	0.01
2B	10	95%	93%	98%	-0.12	0.02
3A	10	82%	90%	92%	-0.19	0.10
3B	10	92%	96%	100%	0.03	0.12
4	9	99%	99%	100%	0.04	0.04
weighted AVE	**10**	**91%**	**91%**	**97%**	**-0.13**	**0.05**

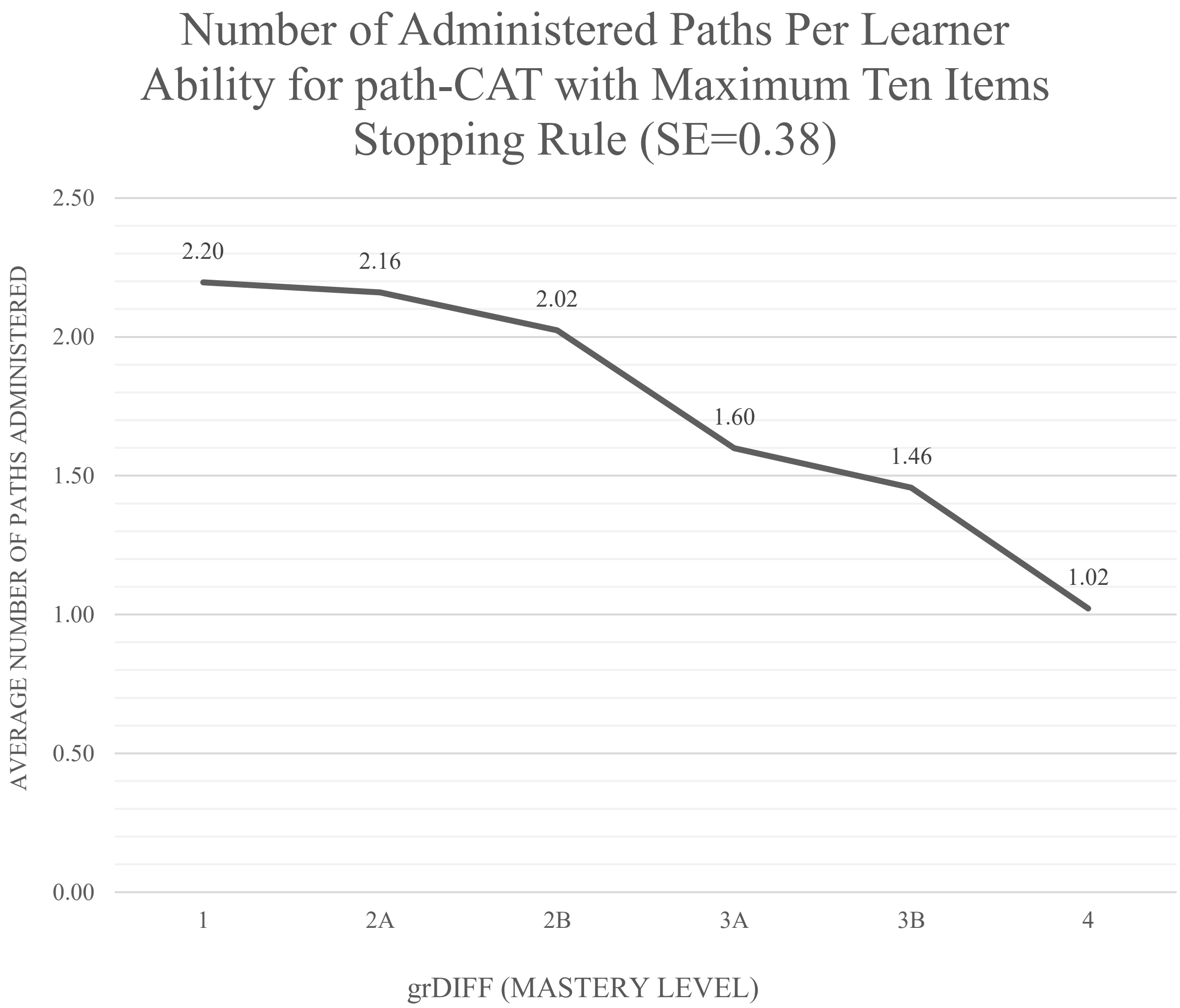


Figure 32 Number of paths administered for a test stopping rule of maximum ten items. Note that grDIFF indicates mastery level where grDIFF=1 is high ability and grDIFF=4 is low ability (Table 4)

VII. Discussion

The purpose of this simulation study is to test the efficiency of an implementation of GRAPH-CAT; a new approach to estimating attribute mastery. A simulation is conducted using the programing language of Python to create item bank parameters similar to those used by a previous study comparing RSM-CAT and POSET-CAT (Yamada, 2008). This simulation uses a maximum likelihood estimation (MLE) and the Brute Force Method to estimate ability in each path-CAT.

VII. A Highlight of Major Findings

Using a path-CAT stopping rule of SE=0.20 and a test stopping rule of complete outcome mastery assignment, GRAPH-CAT achieves a proportion knowledge state matching (PKSM) of 92% accuracy in an average of twenty items with SE=0.33 (see *attmatch* in Table 7). When the test is stopped in ten items GRAPH-CAT achieves PKSM of 83% accuracy in an average of ten items with SE=0.38 (see *attmatch* in Table 10). Using a similar data set, POSET-CAT based in partially ordered set theory, achieves a PKSM of 87.50% in ten items with

SE=0.20 (Yamada, 2008). In twenty items, RSM-CAT achieves 82.46% PKSM. This comparison indicates that using the most straightforward approach to CAT, GRAPH-CAT will likely perform somewhere in the range of efficiency between POSET-CAT and RSM-CAT.

Under both test stopping conditions, the average estimated grade closely matches the real grade (see *aborbel* in Tables 9 and 12). This indicates the ease with which the learner will be able to locate their ideal learning resources if all curriculum learning resources are made available to them. Navigating through learning resources is best visualized as the learner moving up or down by at most one level in a hierarchical path to find their Zone of Proximal Development (Shabani, 2010).

Both simulations indicate that the number of paths administered decreases as level of deficiency increases (see Figure 31 and 32). This is due to the quick reduction of testable outcomes, according to CME, once a low difficulty outcome is marked as non-mastered. On average, approximately twice as many paths are administered using a stopping rule of outcome mastery assignment when compared to a stopping rule of ten items. This is likely due to the requirement of at least ten items to achieve SE=0.20 in a path-CAT.

The efficiency measures are affected by numerous parameter settings, introduced by the multiple pathways. A few of these parameter settings are explored below. They include standard error of measurement (SE), maximum test length, number of items in the bank, and the test scale.

VII. A. i Standard Error of Measurement (SE). The SE is the measure of inaccuracy of an IRT θ value as measured by a path-CAT. Alongside SE, the difference between the previous item SE and the current item SE—*seDIFF*—is another value that can be used to stop a CAT. The *seDIFF* values tend to converge must faster than the SE values. As with any other CAT, a stopping condition with higher accuracy (SE=0.20) in a path-CAT results in a longer test while lower accuracy (SE=0.50) allows for a shorter test. This relationship holds true for both the accuracy of estimated theta and the accuracy of estimated attribute mastery.

VII. A. ii Maximum Test Length. A limit can also be set on the maximum number of items administered. If using either test length or SE to end a path-CAT, a prevalent problem is that the two values must be balanced to ensure that multiple paths are administered. If a balance is not found, then the mastery graph will usually be estimated based on ability from the first path alone. This is a result of administering the maximum allowable test items in the first path-CAT to achieve the desired SE. This effect can be seen by comparing the number of items and paths administered in simulation 1 and simulation 2 (Figure 31 and Figure 32). Notice that many outcomes have likely not been tested when the maximum test length is limited to ten.

VII. A. iii Number of Items in the Bank. Increasing assignment of items to each node decreases granularity between item difficulties. This increases the chance of the confidence interval containing the threshold between two grades. This is a problem when toggling between estimated θ ability and estimated grade level. An item bank with two to four items per outcome appears to achieve desirable results.

VII. A. iv Test Scale. In summative assessment the IRT scale tends to be in the range of around -3 to 3 (Thompson, 2009). The GRAPH-CAT test scale is an extended IRT scale that accommodates four grade levels rather than one. This places the lowest difficulty item at b=-12.00 and the highest difficulty item at b=13.77 (Table 5). These values are assigned according to difficulty to attribute alignment. In other words, difficulty is assigned according to the number of attributes the item is supposed to measure. It is found that as the range of the scale increases from (-3, 3) to (-16, 16), the accuracy of PKSM increases.

VII. B Limitations of the Study

A major shortcoming of this study is that the data distributions for student responses is simulated. Simulated data is used as original student data from previous studies on CD-CAT is unavailable. Student response data, while not readily available, would provide a more direct comparison of efficiency.

Another shortcoming is that the item data is not the same as that used by Yamada (2008). The simulated item bank maintains the attributes measured by each item as used by Yamada (2008), but the difficulties are generated according to *difficulty to attribute alignment*. This impedes a more direct comparison of GRAPH-CAT to POSET-CAT and RSM-CAT.

As a final note on limitations, recall that the MAP and EAP approaches to CAT accommodate for responses that are all correct or all incorrect. The MLE function is used in this study as it is the most straightforward approach to estimating ability. It is likely that other approaches to CAT, such as the MAP or EAP, will further improve efficiency results, but these simulations are left as directions for future research. The following section introduces a framework for GRAPH-CAT and this lends itself to other possible directions for future research.

VII. C Directions for Future Research

As a generative test, GRAPH-CAT faces issues related to item over-exposure and, as an embedded assessment, it must satisfy the ethical requirements of learning analytics. While this simulation has shown the potential of GRAPH-CAT to be useable in terms of efficiency, it remains to demonstrate its practical use. This includes how test quality measures such as validity and reliability are to be provided. Further, by using a graph to model the learner, new measurements relevant to the field of measurement theory must be provided.

DIFFR is a framework for GRAPH-CAT that attempts to outline both the theoretical and practical aspects of these topics. The DIFFR name is selected to indicate the encouragement of differences between students while emphasizing an accommodation for those differences. A framework is necessary to bridge the gap between obtaining results of the diagnostic test and providing effective formative feedback. According to Bejar (2002), a framework for any generative test must address implementational issues associated with equity, security, and interface design.

VII. C. i DIFFR: Equity. Equity concerns are addressed by ensuring content-quality, learning goal alignment, motivation, feedback, and adaptation (Leacock and Nesbit, 2007). The item digraph supports these domains of equity, by revealing hidden dimensions of a learning construct and adapting the item digraph according to student responses. The goal of providing equity is further developed by creating the *ideal item digraph.*

The *ideal item digraph* is an item digraph that is measurably shown to be content-complete, fair, and free of bias. For example, a content-complete item digraph in mathematics includes attributes belonging to reading comprehension. By including reading comprehension as a measured construct for mathematics, bias due to reading deficiency is reduced. To ensure the item digraph is content-complete each attribute must be shown to be necessary and the complete set of attributes shown to be sufficient.

Ensuring prerequisite structures are content-complete may be insurmountable without assistance from computers. One possible approach is to use a covariance matrix to indicate the cooccurrence of attribute mastery. A high frequency of mastery-cooccurrence may indicate membership to a common item digraph. This is a variant of an approach already taken by others towards modelling the latent attribute space (Rupp, Templin, & Henson, 2008). Adapting the existing approach to the item digraph is left as a direction for future research.

Another possible measure introduced by the item digraph is *rate of attribute mastery*. Recall from the two assumptions of difficulty to attribute alignment that we require the attributes to be of a similar, if not the same, granularity. According to Leighton and Gierl (2011) the factors required to determine a valid attribute grain size are poorly understood. One interpretation of rate of attribute mastery, is the expected time required to master an attribute, as observed from student responses in a given population. Deviation from this expectation may indicate an attribute granularity that is either too large or too small.

Similarly, *rate of outcome mastery* should remain constant across populations. A difference in rate of improvement towards a learning objective according to external variables, such as gender or race, may indicate the occurrence of bias in the inclusion of certain attributes. An example of such bias would be a hierarchy that has included the scientific method as a prerequisite to scientific thinking. Including such a precedence in the hierarchy may conflict with indigenous *ways of knowing* (Shepherd, 2016). This may cause students from indigenous cultures to become hesitant or even non-cooperative in the learning experience. If this should occur, a significant difference in rate of outcome mastery is expected to be observed between students from indigenous cultures and students from non-indigenous cultures.

Another method of detecting bias is by validating the precedence relationships between attributes. A monitoring measure (M) is introduced to indicate possible fault in the organization of attributes. Given a bank of *well-constructed items* and an item digraph with validated precedence relationships, it will be observed that higher difficulty items belong to learning outcomes closer to the root. Note that, in addition to satisfying traditional IRT measures of quality, a *well-constructed item* further provides a precise measure of the indicated attributes.

To define M formally, suppose we have an item digraph with a bank of well-constructed items. Let b_i be the difficulty of any item from node v_i in any hierarchical path from the item digraph. If we let

$$\mathrm{M} = \frac{(b_i - b_j)}{(level(v_j) - level(v_i))}$$

where $i \neq j$, then it will be observed that $\mathrm{M} \geq 0$. On the other hand, a negative M ($M < 0$) indicates a questionable precedence relationship between attributes.

The introduction of these measurements alludes to a constant tailoring that must take place to craft the item digraph into its ideal state. The ideal item digraph provides a valid and reliable measure of attribute mastery and therefore, in addition to formative assessment, it can be used for summative assessment. Given an ideal item digraph, mastery of all items in the root node indicates mastery of all attributes in the item digraph. Thus, summative and formative assessment are merged into two different mastery-states of the same item digraph. This theoretical union of summative and formative assessment is in keeping with what has been observed by Redecker and Johannessen (2013) as a natural consequence of embedded assessment.

The feasibility of achieving the ideal item digraph requires a system that is able to securely collect and store longitudinal student data. A major challenge behind implementing such a system is in securing student data. The data collection and storage procedures must not only meet industry standards, but should also satisfy the requirements of parents, students, teachers, and other stakeholders.

VII. C. ii DIFFR: Security. Security must be maintained for both the learner and the test. Learner-security ensures that student data is kept secure and school authorities are held accountable for the maintenance of security. Test-security, on the other hand, ensures that the test remains valid by reducing item exposure.

A highly sensitive and pressing issue surrounding data security in educational multimedia is the protection of student data. There are two approaches to securely providing access to the GRAPH-CAT test interface. Both approaches can be described as *privacy-by-architecture* meaning privacy is built into the software or hardware rather than enforced through policy (Spiekermann & Cranor, 2009). The first approach is to establish an on-line accessibility to the test. In this case, the student remains anonymous by being provided with a user ID. The ID is supplied to the teacher and given to the student. The second approach makes the software available to teachers and students on a closed network provided by the school administration.

According to Le Metayer (2008), the presence of digital technology in schools requires that privacy issues, such as secure accessibility, be addressed from both a technological and a legal standpoint. Particular care must be given to the security of digital records as data collection by third party interests has become normalized. Pardo and Siemens (2014) recommend including such features as limiting data that students can provide, limiting access to authorized individuals, and allowing learners to access their own data. In some cases, a data privacy officer may be desired, and perhaps necessary, to ensure data security. Spiekerman and Cranor (2009) recommended that these issues be addressed in the early stages of framework development.

In addition to securing item responses, the longitudinal use of a testing system also requires the maintenance and security of an item bank. Test security requires that item exposure is kept to a minimum (McGlohen & Chang, 2008). Yet, according to McGlohen and Chang (2008), all items in the bank must be active so that the number of learner responses to each item is greater than zero. McGlohen and Chang (2008) note that a desirable item exposure is an administration to under 20% of examinees. One possible approach of meeting this high demand for items, is to integrate automatic item generation (AIG) into the generative testing process.

AIG is the use of computers to generate items based on an *item model*, also known as a *blueprint* (Gierl & Lai, 2013). AIG has become essential to generative testing, as the switch towards embedded computerized learning demands greater measures be taken to reduce item exposure (Redekken & Johannessen, 2013). Further, AIG allows for a much greater efficiency in the item development process.

The item development process involves modelling items and using the models to generate new items. Parts of the item model string, called *elements*, are alterable so that by replacing these elements with a selection from a set of *values* we create new items (Gierl & Lai, 2013). Meanwhile, other substrings of the item remain constant.

The prototypes for item models are often taken from pre-existing items (Gierl & Lai, 2013). A common practice is to adapt an existing item into an item model that fits the attribute hierarchy (Leighton & Gierl, 2011). This type of *retrofitting* is not recommended as it leads to item-attribute misfit. A better approach, according to Gierl, Alves, and Majeau (2010), is to design new item models that are specifically designed to measure the intended attributes.

An issue that arises in the AIG process is the expectation that generated items and the template item have the same difficulty (Bejar, 2002). In 1977 Merwin found that the difficulty of generated items was not the same as the item template used to create the item model (Bejar, 2002). Bejar (2002) concludes that a theoretical analysis is necessary for generative testing so that greater agency is gained over parameters of the generated items. In the context of generative diagnostic testing, this theoretical analysis comes in the form of creating cognitive models.

The cognitive model specifies those elements that effect the difficulty level of the generated items (Gierl & Lai, 2013). According to Gierl and Lai (2013), a cognitive model enhances control over the psychometric properties of the generated items. As the items belong to the item digraph, the properties of the cognitive model must align with the attributes included in the Q-matrix of each learning outcome.

Expected costs for the development of a high-quality CD-CAT are associated with the development of *base pools* of item models. To form a quality base pool requires at least the same amount of time from subject matter experts as that to form an item bank. Yet, by switching from the item to the item model as the base unit of a test, the efficiency of designing high quality items can be greatly enhanced.

VII. C. iii DIFFR: Interface Design. To further explore issues related to security and equity, a user interface must be made available to both teachers and students. Interface design has long been recognized as a prominent issue in educational software design. A comprehensive example of domains of interest when considering the interface design are outlined in the Learning

Objective Review Instrument (LORI) developed by Leacock and Nesbit (2007). These domains include motivation, presentation design, interaction usability, accessibility, and reusability. The LORI instrument is readily available, and helps to ensure inclusivity in the learning experience.

A high-quality interface design is essential for improving both student motivation and student cognitive learning outcomes. A next step for future research involves the provision of such a user interface to students to measure the effectiveness of DIFFR in a classroom setting. According to All, Plovie, Castellar, and Van Looy, J. (2017), a pre-post test design is an effective approach for measuring improvement in cognitive learning outcomes as it allows for the control of prior ability. An important direction for future research includes measures of improvement in student cognitive learning outcomes and improvement in student motivation. User-satisfaction of both students and teachers should also be measured using methods such as direct observation, interviews, and surveys.

VIII. Conclusion

The graph theory approach to attribute mastery offers a paradigm-shift from traditional forms of test measurement. The paradigm-shift from classical test theory is revealed by the error attribution. As with the IRT framework, the error in measurement is absorbed by inaccuracy of the test design and associated test parameters (Lord, 1980). The conceptual departure from traditional IRT is that multi-dimensionality is allowed by including secondary dimensions as visible constructs in the item digraph. These dimensions are each measured separately by sub-tests associated with a single path-CAT.

Diagnostic testing with the item digraph relies on the prerequisite relationship for the administration of items. The item digraph continually improves through validation and adaptability of the prerequisite relationships. Under the assumption that the item digraph is complete, fair, and free of bias, a continuous progress of the learner will be observed. Continuous and predictable progress of all learners is the hallmark of an ideal item digraph. Similarly, delayed or unpredictable progress is an indication that the item digraph is not ideal.

A non-ideal item digraph could, in part, be due to a low-quality item bank. Creating a high-quality item bank requires that items align well with the attributes they are intended to measure. Accuracy in prior difficulty of the items, as assigned by the test scale, requires that the items are well-constructed.

The predictability of item difficulty offered by *difficulty to attribute alignment* overcomes previous issues with implementing MLE for CD-CAT. A major strength of introducing the MLE, MAP, or EAP function is that a standard error of measurement can be associated with the estimated knowledge state. This provides the user with a level of confidence in estimated ability.

The most salient feature of the item digraph approach, is that it overcomes the stochastic item administration depended upon by previous CD-CAT implementations. Impenetrable, complicated algorithms and unwieldy test usability have inhibited widespread use of computerized formative assessment in the classroom (Lim, 2015; Yamada, 2008). The item digraph approach to CD-CAT may bring contributions that can address these issues, by offering a user-friendly algorithm based in the framework of statistical theory.

Bibliography

Ackerman, T. A. (1989). An Alternative Methodology for Creating Parallel Test Forms Using the IRT Information Function.

Aho, A. V., Garey, M. R., & Ullman, J. D. (1972). The Transitive Reduction of a Directed Graph. *SIAM Journal on Computing*, *1*(2), 131-137.

All, A., Plovie, B., Castellar, E. P. N., & Van Looy, J. (2017). Pre-Test Influences on The Effectiveness of Digital-Game Based Learning: A case study of a fire safety game. *Computers & Education*.

Baker, F. B., & Kim, S. H. (2004). *Item Response Theory: Parameter Estimation Techniques*. CRC Press.

Bejar, I. I. (2002). Generative testing: From Conception to Implementation. *Item Generation for Test Development*, 199-217.

Boyce, R. L., & Ellison, P. C. (2001). Choosing the Best Similarity Index When Performing Fuzzy Set Ordination On Binary Data. *Journal of Vegetation Science*, *12*(5), 711-720.

Čisar, S. M., Radosav, D., Markoski, B., Pinter, R., & Čisar, P. (2010). Computer Adaptive Testing of Student Knowledge. *Acta Polytechnica Hungarica*, *7*(4), 139-152.

Diestel, R. (2017). Graph Theory, volume 173 of *Graduate Texts in Mathematics*.

Education, A. (2007) The Alberta K–9 Mathematics Program of Studies with Achievement Indicators. *Mathematics Kindergarten to Grade*, *9*.

Education, A. (2017) Alberta Kindergarten to Grade 12 Scope and Sequence. Alberta Education

Embretson, S. E., & Reise, S. P. (2013). *Item Response Theory*. Psychology Press.

Falmagne, J. C., Cosyn, E., Doignon, J. P., & Thiéry, N. (2006). The Assessment of Knowledge, In Theory And In Practice. In *Formal concept analysis* (pp. 61-79). Springer, Berlin, Heidelberg.

Gass, S. I., & Assad, A. A. (2005). Model World: Tales from The Time Line—The Definition of OR and the Origins of Monte Carlo Simulation. *Interfaces*, *35*(5), 429-435.

Gierl, M. J., Alves, C., & Majeau, R. T. (2010). Using Principled Test Design to Develop and Evaluate A Diagnostic Mathematics Assessment in Grades 3

And 6. In *Annual Meeting of the American Educational Research Association Denver, CO, USA*.

Gierl, M. I., & Lai, H. (2013). Using Weak and Strong Theory to Create Item Models for Automatic Item Generation. *Automatic Item Generation*, 26-39.

Geyer, C.J. (2003). *Theory of Maximum Likelihood Estimation*. University of Minnesota. Retrieved from https://www.stat.umn.edu/geyer/5931/mle/mle.pdf

Godin, R., Missaoui, R., & Alaoui, H. (1991, November). Learning Algorithms Using a Galois Lattice Structure. In *Tools for Artificial Intelligence, 1991. TAI'91., Third International Conference on* (pp. 22-29). IEEE.

Harary, F. (1955). The Number of Linear, Directed, Rooted, And Connected Graphs. *Transactions of the American Mathematical Society*, *78*(2), 445-463.

Hsu, C. L., Wang, W. C., & Chen, S. Y. (2013). Variable-Length Computerized Adaptive Testing Based on Cognitive Diagnosis Models. *Applied Psychological Measurement*, *37*(7), 563-582.

Jech, T. (2013). *Set Theory*. Springer Science & Business Media.

Kantrowitz, T. M., Dawson, C. R., & Fetzer, M. S. (2011). Computer Adaptive Testing (CAT): A Faster, Smarter, And More Secure Approach to Pre-Employment Testing. *Journal of Business and Psychology*, *26*(2), 227.

Kroese, D. P., Brereton, T., Taimre, T., & Botev, Z. I. (2014). Why the Monte Carlo Method Is So Important Today. *Wiley Interdisciplinary Reviews: Computational Statistics*, *6*(6), 386-392.

Lai, H., & Gierl, M. (2013). Generating items Under the Assessment Engineering Framework. In M. Gierl (Ed.) Automatic Item Generation (pp. 77-101). Routledge.

Leacock, T. L., & Nesbit, J. C. (2007). A Framework for Evaluating the Quality of Multimedia Learning Resources. *Educational Technology & Society*, 10 (2), 44-59.

Le Métayer, D. (2008, October). A Formal Privacy Management Framework. In *International Workshop on Formal Aspects in Security and Trust* (pp. 162-176). Springer, Berlin, Heidelberg.

Leighton, J. P., & Gierl, M. J. (2011). *The Learning Sciences in Educational Assessment: The Role of Cognitive Models*. Cambridge University Press.

Levine, C. (2011). Partially Ordered Set Theory. *18.312 Algebraic Combinatorics*. Massachusetts Institute of Technology.

Lim, Y. (2015). *Cognitive Diagnostic Model Comparison* (Doctoral Dissertation). Georgia Tech Thesis and Dissertations. Retrieved from https://smartech.gatech.edu/handle/1853/53513

Linacre, J. M. (2000). Computer-Adaptive Testing: A Methodology Whose Time Has Come. *Chae, S.-Kang, U.–Jeon, E.–Linacre, JM (Eds.): Development of Computerised Middle School Achievement Tests, MESA Research Memorandum*, *69*.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Arlbaum Associates, Inc..

Margoliash, J. (2010). Matrix-Tree Theorem for Directed Graphs.

McGlohen, M., & Chang, H. H. (2008). Combining Computer Adaptive Testing Technology with Cognitively Diagnostic Assessment. *Behavior Research Methods*, *40*(3), 808-821.

Mullis, I.V.S, Martin, M.O., Gonzales, E.J., Gregory. K.D., Garde. R.A., O'Connor, K.M., Chrostowski, S. J., & Smith, T. A., (2000). *TIMSS 1999 International Mathematics Report.* Chestnut Hill, MA: International Study Center, Boston College.

Pardo, A., & Siemens, G. (2014). Ethical and Privacy Principles for Learning Analytics. *British Journal of Educational Technology*, *45*(3), 438-450.

Redecker, C., & Johannessen, Ø. (2013). Changing assessment—Towards a new assessment paradigm using ICT. *European Journal of Education*, *48*(1), 79-96.

Roussos, L. A., DiBello, L. V., Stout, W., Hartz, S. M., Henson, R. A., & Templin, J. L. (2007). The Fusion Model Skills Diagnosis System. *Cognitive Diagnostic Assessment for Education: Theory and Applications*, 275-318.

Rupp, A., Templin, J., & Henson, R. A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. Guilford Press.

Shabani, K., Mohamad K., and Saman E. (2010). Vygotsky's Zone of Proximal Development: Instructional Implications and Teachers' Professional Development. *English language teaching* 3.4, 237-248.

Shaffer, C. A. (1997). *A Practical Introduction to Data Structures and Algorithm Analysis*. Upper Saddle River, NJ: Prentice Hall.

Shepherd, T. H. (2016). The Convergence of Modern Scientific and Traditional Earth Ways of Knowing, with Implications for Global Education.

Spiekermann, S., & Cranor, L. F. (2009). Engineering Privacy. *IEEE Transactions on software engineering*, *35*(1), 67-82.

Tatsuoka, K. K. (2009). *Cognitive Assessment: An Introduction to The Rule Space Method*. Routledge.

Thompson, N. A. (2009). Ability Estimation with Item Response Theory. *Assessment Systems Corporation*. Retrieved from http://www.assess.com/docs/Thompson_(2009)__Ability_estimation_with_IRT.pdf

Thompson, N. A., & Weiss, D. J. (2011). A Framework for The Development of Computerized Adaptive Tests. *Practical Assessment, Research & Evaluation*, *16*.

Wang, C., Chang, H. H., & Huebner, A. (2011). Restrictive Stochastic Item Selection Methods in Cognitive Diagnostic Computerized Adaptive Testing. *Journal of Educational Measurement*, *48*(3), 255-273.

Weisstein, E. W. (2019). Cover Relation. *MathWorld--A Wolfram Web Resource*. Retrieved from http://mathworld.wolfram.com/CoverRelation.html

Weisstein, E. W. (2019). Hasse Diagram. *MathWorld--A Wolfram Web Resource.* Retrieved from http://mathworld.wolfram.com/HasseDiagram.html

William, D. (2011). *Embedded Formative Assessment*. Solution Tree Press.

van der Linden, W. J., & Pashley, P. J. (2009). Item Selection and Ability Estimation in Adaptive Testing. In *Elements of Adaptive Testing* (Pp. 3-30). Springer, New York, NY.

Van Rossum, G., & Drake, F.L. (2011). The Python Language Reference Manual. Network Theory Ltd..

Yamada, T. (2008). *Comparison of Cognitively Diagnostic Adaptive Testing Algorithms*. Columbia University.

Yan, D., & Magis, D. (2016). Computerized Adaptive and Multistage Testing with R.

Youri, Z. (2008). Basic Set Theory. *Semantics I*. Boston University.