THOMAS, MARIE HUFFMASTER, Ph.D. Modeling Differential Pacing Trajectories in High Stakes Computer Adaptive Testing Using Hierarchical Linear Modeling and Structural Equation Modeling. (2006)
Directed by Dr. Richard M. Luecht. 94pp.

This study compares two statistical methods for modeling changes in response latency (timing) patterns on a high-stakes adaptive test: (1) hierarchical linear modeling (HLM2) and (2) growth modeling using structural equation modeling (SEM). The testing context involves the NCLEX-RN®, a variable-length, computerized adaptive test used to license registered nurses in the United States and its territories. Item-level response-time data from 4,415 first-time takers of the NCLEX-RN® examination are used to create and evaluate the different "examinee pacing trajectory" models. The examinee pacing trajectory models are separately fit to examinees at three proficiency levels: (1) clear failers who take an abbreviated test of only 75 items; (2) indeterminate examinees who take a variable-length test of up to 265 items; and (3) clear passers who also take an abbreviated test of only 75 items.

The HLM- and SEM-based approaches provided comparable but not identical results. The estimated intercept terms for the pacing trajectory models were different for each of the three proficiency groups, indicating a possible association between ability and pacing skill. However, the intercepts did not differ dramatically across the two modeling methods because those estimates are essentially based on empirical means. The pacing trajectory slopes varied both in scale and relative magnitude across proficiency groups and by analysis method. The HLM slope estimates reflect the average scale variances of the empirical response data across blocks of items. In contrast, SEM-based growth

modeling employs arbitrary constraints imposed to statistically identify the model. These scaling differences made it difficult to directly compare the HLM-based and SEM-based slopes. Ultimately, however, it was concluded that either method (HLM or SEM) is able to model pacing trajectories in a meaningful way.

MODELING DIFFERENTIAL PACING TRAJECTORIES IN HIGH STAKES COMPUTER ADAPTIVE TESTING USING HIERARCHICAL LINEAR MODELING AND STRUCTURAL EQUATION MODELING

By

Marie Huffmaster Thomas

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2006

Approved by

_____
Committee Chair

APPROVAL PAGE

This dissertation has been approved by the following committee of the Faculty of Graduate School at The University of North Carolina at Greensboro.

Committee Chair _____

Committee Members _____

_____

_____

_____

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGEMENTS

This project would not have been possible without the support of many people. I want to thank my adviser, Richard M. Luecht, PhD, for his encouragement and guidance. Also thanks to my committee members, Terry Ackerman, PhD, Daniel L. Bibeau, PhD, Robert A. Henson, PhD, and Rick Morgan, PhD who offered guidance and support. Thanks to Tom O'Neill, PhD, Associate Director, NCLEX® Examinations, and Casey Marks, PhD Associate Executive Director — Business Operations, National Council State Boards of Nursing (NCSBN), for their invaluable assistance.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Testing is a ubiquitous part of everyday life but nonetheless a source of anxiety for most people. Historically, testing was a very regimented, structured, and orderly process. Every test taker took the same test under the same conditions. All examinees were given the same amount of time and took a test of the same difficulty. This was an important basis for the comparability of test scores. Final scores could be compared directly to determine the relative level of performance (Plake, 1999).

Computerized adaptive testing (CAT) engenders a different environment for test taking. Based on a particular probabilistic item response theory (IRT) model, adaptive testing allows the administration of very different tests measuring the same latent construct or trait (Plake, 1999). In its typical implementation, CAT requires a large pool of items that have been developed and calibrated to determine the item characteristics (estimated parameters) for a particular IRT model. For example, the common three-parameter (3PL) IRT model mathematically expresses the probability of a correct answer to item i—that is, u_i=1 (and where u_i=0 indicates an incorrect response) as

$$prob\left(u_i = 1 \middle| \theta; a_i, b_i, c_i\right) \equiv P_i = c_i + \frac{1 - c_i}{1 + \exp\left[-a_i\left(\theta - b_i\right)\right]} \qquad (1)$$

where:

θ denotes the examinee's proficiency score with respect to the measured construct or latent trait;

b_i represents the item difficulty;

a_i, indicates the item's capability to differentiate between high and low proficiency examinees—often called the item discrimination or sensitivity; and

c_i is the assumed probability that the low proficiency examinee will be able to answer the item correctly due to chance guessing.

The items in the CAT database or item bank are pre-calibrated using experimental field-test or pilot data to determine statistical properties (a_i, b_i, and c_i in Equation 1-1). The CAT item selection algorithm then proceeds as follows. First, each examinee is given several initial items from the item bank—usually selected randomly from items within the mid-range of difficulty. Based on the examinee's initial or provisional proficiency score, as estimated using his or her scored performance on those initial items, the CAT algorithm selects the next item for administration to maximize the accuracy of the current provisional proficiency estimate. In the simplest IRT models, this "accuracy maximization" is accomplished by matching the item's difficulty to the examinee's provisional proficiency estimate. Accordingly, the next item selected by the CAT algorithm will be easier if the examinee answers most of the initial items incorrectly and more difficult if the examinee gets most of the initial set of items correct. In general, an adaptive test actually maximizes the IRT test information function

$$I + I_{n+1} = \sum_{i=1}^{n} \frac{D^2 a_i^2 Q_i \left(P_i - c_i\right)^2}{P_i \left(1 - c_i\right)^2} + \frac{D^2 a_{n+1}^2 Q_{n+1} \left(P_{n+1} - c_{n+1}\right)^2}{P_{n+1} \left(1 - c_{n+1}\right)^2} \qquad (2)$$

by incrementally selecting items that provide the greatest contribution to the test information function for $n + 1$ items—i.e., choosing the item with the largest value of the rightmost term in Equation 2. This is a type of statistical optimization. The adaptive testing process continues until one of the stopping criteria is met. Common stopping rules are: (a) reaching a fixed number of items; (b) attaining a specified degree of accuracy with respect to the provisional proficiency score estimates; or (c) when a certain degree of decision accuracy is attained (in the context of mastery or with respect to a pass/fail standard).

Because the adaptive mechanism selects maximally "challenging" items for each test taker, there may be pacing and other consequences that are experienced by some examinees. The purpose of this study is to understand the nature of the interactions between items and examinees over the course of a computer adaptive test in terms of pacing.

It is helpful to distinguish between the related concepts of "pacing" and "speededness" before proceeding. Speededness is a characteristic of the test that is related to time constraints. A speeded test is often contrasted with a power test. A speeded test requires examines to work as quickly as possible. Items are typically written to be at a uniformly low difficulty level and time constraints are so severe that nobody can finish all of the items. Many early performance tests were speeded by design, with a typical example being a typing test. Power tests are incrementally ordered by difficulty and provide ample time for everybody to at least attempt every item. It often becomes

clear how much a test taker knows by how far he or she can progress in the power test by providing consistently correct answers.

Bridgeman, McBride and Monaghan (2004) in their work with the Graduate Record Examination (GRE) define speededness in testing as the effect that time limits have on test takers' scores. If examination time limits prevent examinees from considering and answering each question, the examination is "speeded". The examinee score would be lower than if the examinee had been given an unlimited amount of time to complete the test. Henderson (2004) in his work with the Law School Aptitude Test (LSAT) concluded that measures of speed and level of ability (ability to accurately perform difficult tasks) have very low correlations. In the psychometric field, the general consensus is that the results of an assessment or aptitude test can be confounded if the speededness component is large and speed is construct irrelevant for the construct being measured by the test. A speed test is one where an examinee is given the minimum amount of time that is adequate for a typical student to complete the exam while feeling pressure to complete it. For a power exam the examinee is given as much time as needed. Hornke (2000) in her work with German military services computer adaptive test data found low correlations between test scores under timed and un-timed conditions. Bridgeman and Cline (2004) also found similar results when manipulating the time per item allowed on the Graduate Record Examination and Scholastic Aptitude Test. Examinees under restrictive time limits use different test-taking strategies than when they take the same exam under unlimited time conditions. Thus, speededness is a test validity issue (Stout & Heck, 1997). The validity of any inferences we make about an

examinee's proficiency is often based solely on his or her test score. If examination factors such as speededness or secondary examinee proficiency factors such as pacing skill directly or indirectly affect performance on the test, the resulting score will not properly represent the examinee's true proficiency level. Speededness and pacing therefore both represent potential threats to test score validity (Wise & Kong, 2005).

As noted earlier, CAT has further changed testing by altering the difficulty of the test for each test taker. Therefore, examinees do not receive the same items. The definition of test speededness has evolved to measure the extent that time limits affect examinee test performance (Hadadi & Luecht, 1998). In contrast, when testing was paper and pencil; speededness was commonly assessed by computing the percentage of examinees failing to reach a percentage of items on the exam (Schnipke & Scrams, 1997). With the advent of computer-based testing, it is easy to store actual response times on each question and even track test taking strategies such as item review and how often answers are changed.

Examinations with a time limit evoke two possible strategies for examinees who do not normally respond fast enough to complete the examination: some examinees do not change their pace, running out of time, while other examinees resort to rapidly guessing at items as time nears the end. Speededness in the CAT environment is linked to *rapid guessing behaviors* (Schnipke & Scrams, 1997). Rapid guessing behavior is triggered as a test taking strategy when the examinee has an external cue that he or she is running out of time or in the case of the NCLEX-RN® examinee that the test is continuing beyond the 75 initial questions. When an examinee is displaying rapid

guessing behavior, the examinee works rapidly through the items, skimming for keywords, answering based on partial information or guessing the answer. Contrast this with *solution behaviors* that demonstrate the examinee read and answered each question carefully. Solution behavior is demonstrated when the examinee reads each item carefully and fully considers the solution. Response times resulting from solution behavior depend on item length, item difficulty and person specific characteristics. Accuracy depends on examinee ability and item difficulty and other item characteristics (Schnipke & Scrams, 1997).

Hadadi and Luecht (1998) in their work with the National Board of Medical Examiners argue that speededness is not an isolated test characteristic; they posit that in order to understand speededness an understanding of examinee pacing strategies is needed. Speededness is a characteristic of the exam while pacing is an examinee characteristic. Pacing strategies may lead to item omission or rapid guessing. Self-governing functions such as confidence, persistence, test-wiseness, test anxiety, cognitive or personal characteristics may explain examinee test taking behaviors when taking a test they find speeded (Martinez, 1999). The examinee's ability is a mix of his mental ability, the time devoted to the item, and the persistence given to finding the right solution. A correct answer is given if the examinee has sufficient ability and stays on task, working on the problem until the solution is found. An incorrect answer is more likely if the examinee shifts away from the problem and ceases to work on the task despite his or her sufficient ability (Hornke, 2000).

Conflicting research relating response time to ability does exist. Smith (2000) in working with Graduate Management Admission Test (GMAT) data states that it is possible those items, which are similar with respect to difficulty and discrimination, differ with respect to cognitive demand or the number of necessary procedures to correctly answer an item. It is possible for two examinees of equal ability to be administered items that are equal in terms of item characteristics of difficulty, discrimination and propensity for guessing correctly, but that one examinee was administered a set of items that was more cognitively demanding. The increased cognitive demand requires more mental energy and time. Test designed to be equivalent in item parameters can vary significantly in testing time. Hornke (2000) reported that the wrong responses required more time than right responses in an adaptive testing environment, indicating that total response time is independent of proficiency. Correct solution response times may represent an effort based on response time of a successful mental process, while wrong solution response times may represent time of lesser effort plus loss of interest leading to failure. Some examinees give up mental effort on an item and guess, thus indicating that they may employ different cognitive processes. Bridgeman and Cline (2000) found that because more difficult items, which by their nature tend to take longer, are administered disproportionately to higher ability students, there is a positive correlation between test time and ability of the student. Thus for CAT the usual relationship between ability and speededness is totally reversed so that the examination is more speeded for the higher ability student.

New strategies are needed to identify not only speededness but also pacing strategies and changes in response patterns that could provide valuable information to test designers to allow for an even distribution of questions with a similar time demand and cognitive load.

An example of how speededness research can inform test design practices comes from the NCLEX-RN®, the subject test for the present study. Under the original test design, Bontempo and Julian (1997) found that at least five percent of the examinee population was significantly affected by the time limit on the NCLEX-RN® examination. An implication of Bontempo and Julians' research was that for this five percent of the examinee population, the NCLEX-RN® may have been somewhat speeded. As a result, the original time limit for the NCLEX-RN® examination was increased from four hours and forty-five minutes to five hours and forty-five minutes.

However, NCLEX-RN® examinees are still required to take different length tests, depending on how they perform. As of 2005, the NCLEX-RN® can be from 75 to 265 items long. The number of items administered must be completed within the total time limit of six hours for the NCLEX-RN®. The time limit includes the tutorial, sample items, all breaks and the examination. The initial test length is determined by the examinee responses to the first 75 test items. Examinees are then gated as "clear passers", "clear failers", or are required to continue testing. Some examinees actually make it all the way to 265 items or five hours and forty-five minutes of testing, at which time a final pass/fail decision is made. It is important to note that the total testing time limit does not change for this latter group that continues testing beyond 75 items.

In this study, a sample from the examinee population of 1999-2000 was analyzed to evaluate the interaction between items and examinees over the course of the examination for these three groups: clear passers, clear failers, and the extended testing group. Using Hierarchical Linear Modeling (HLM) and Structural Equation Modeling (SEM) this study explores the performance and timing trends for items and examinees over the course of the examination. This study specifically attempts to identify examinees with inadequate pacing and performance trajectories in a high stakes computer adaptive test.

Ultimately, this study focuses on a primary research question. That is, what do pacing trajectories look like in a computer adaptive test for groups at various proficiency or mastery levels and how can statistical models help understand any apparent pacing trajectories?

The following hypotheses will be tested.

H_1 : There is no difference in pacing trajectory among Clear pass at 75 items (clear passers), Indeterminate at 75 items (indeterminate at 75), Clear fail at 75 items (clear failers) and Indeterminate at 76 to 265 items (indeterminate at 76 to 265 items) proficiency groups.

H_2 : The Clear pass at 75 items proficiency group establishes a pacing trajectory above the pass cut score and maintains that trajectory for 75 items.

H_3 : The Indeterminate at 75 items proficiency group fails to establish a consistent pacing trajectory above or below the pass cut score and is unable to maintain a consistent trajectory for 265 items.

H_4 : The Indeterminate at 76 to 265 items proficiency group fails to establish a consistent pacing trajectory above or below the pass cut score and is unable to maintain a consistent trajectory for 265 items.

H_5 :The Clear fail at 75 items proficiency group establishes a pacing trajectory below the pass cut score and maintains that trajectory for 75 items.

The application of statistical modeling to identify pacing/speededness problems is not new. Several researchers have attempted to use statistical models to understand the extent of speededness effects and examinee pacing behaviors. Bontempo (2000) used the NCLEX-RN® data and a probabilistic model to identify speededness and adjust time limits; Yamamoto (1995) used a latent-class HYBRID model and data from the Test of English as a Foreign Language (TOEFL) to identify speededness. He went on to propose adjustments to the IRT parameters that would minimize the impact of speededness, Bergstrom, Gershon, and Lunz (1994), using a hierarchical linear model and data from a national real estate examination, attempted to identify test items with detectable characteristics that tend to make the items too time consuming. Their intent was to exclude those items from the item bank or to adjust the item selection algorithm to account for the differential item speededness.

Halkitis, Jones and Pradhan (1996) developing a regression model using data from a real estate appraisers national licensing examination to identify test characteristics and properties of test items that could be adjusted so accurate testing time required by examinees could be identified. Similarly, Schinpke and Scrams (1997) used a mixed effects IRT model and data from a non adaptive version of the GRE to measure response

time and detect rapid guessing on individual items. In one of the most comprehensive modeling efforts to date, van der Linden and van Krimpen-Stoop (2003) (also see van der Linden, 2005) developed a lognormal model for time that is capable of estimating a pacing parameter for each examinee and a timing parameter for each item. Van der Linden (2005) has demonstrated some interesting results for simulated data, based upon the Law School Admissions Test, as well as for the Uniform Certified Public Accountants Examination.

The present study involves an item-by-item analysis of the interaction between items and examinees over the course of the NCLEX-RN® examination. Although other research has dealt with pacing and speededness issues, no one has specifically examined the conditional response and pacing patterns of examinee groups on an item-by-item basis over the course of the examination (i.e., based on examinees' apparent levels of proficiency).

In addition to the primary research question and hypotheses stated earlier, the study examines whether any of the variance in response time may be due to specific item types or other surface item characteristics. This study further explores identifiable patterns of examinee behavior that could explain how different types of examinees progress through the test. Ultimately, understanding subject performance domains may reveal aspects of examinee performance that are employed, such as self-regulatory abilities and meta-cognitive abilities that are crucial for high levels of performance (Martinez, 1999).

CHAPTER II

LITERATURE REVIEW

This chapter presents a summary of the relevant research literature on speededness and pacing in high-stakes assessment settings. The chapter begins with some attempts to operationally define speededness and pacing and moves on to cover findings from research studies that highlight the impact of test speededness and that explore pacing as an examinee attribute. Some of the advantages of computer-based testing are discussed insofar as making it possible to collect item-level timing data on every test taker. Finally, various past attempts at and proposals for modeling speededness and pacing are summarized.

Definitions of Speededness

Previous researchers have searched for a definition of speededness. Anastasi (1976) defines a pure speed test as one in which individual differences depend entirely on speed of performance. These tests are constructed from items of low difficulty with a time limit so short that no one can finish all the items. Each person's score reflects only the speed with which he worked. Power tests have a time limit long enough to permit everyone to attempt all items. The difficulty is steeply graded and includes items too difficult for anyone to solve so no one gets a perfect score. Both tests are designed to prevent the achievement of perfect scores. The essential question is: to what extent

is the individual difference in test scores attributable to speed? Speededness in testing in the most simplistic view is the effect of time limits on the candidate's scores. An exam is speeded to the extent that a candidate has a lower score taking the test within the time limit than they would if they had had unlimited time. Bontempo and Julian (1997) proposed that speededness be defined as how much faster examinees work when an examination has a particular time limit versus how fast the examinees would work on an examination with an infinite amount of administration time. Oshima (1994) states by definition a test is a power test when a candidate's total incorrect score is equal to the number of items for which the person is not pressured by time and simply responds incorrectly. A test is speeded when the candidate's total incorrect score is equal to the number of items that the candidate does not attempt. Bejar (1985) states that a test is speeded when a portion of the test-taking population does not have sufficient time to attempt every item. Stafford (1971) supports a speededness quotient (SQ) based on the percentage of un-attempted items. A SQ of zero would result if all errors on the test resulted from inaccurate responses. Such a test is a power test: all errors are due to lack of ability or knowledge. If all items are correct but not all items were attempted, a SQ of one would result. That would indicate that an examination is a speeded test and errors are due to slowness rather than lack of ability or knowledge. Bugbee and Bernt (1992) propose examinee feedback as a means of evaluating exam speededness. Postulating that two students might still be working when the time limit is reached but the tasks they are engaged in may be different; they propose an analysis of examinees' perceptions of time pressure during the examination to indicate speededness. Donlon (1973) states that if the

examinee completes all items on the exam, the allotted time are used as the estimated time needed; if not, then an estimate is calculated based on the number of items completed.

There is not much support for the validity of speeded tests in academic and certification/licensure contexts. Mislevy (1996) suggested that speeded tests ignore the learners' active search for meaning, constructing and representing the subject matter. He argues that levels of proficiency or achievement might be better defined and measured not in terms of numbers of facts recalled or recognized, but by determining the test-takers' levels of understanding of key concepts and principles. Linn (2000) states that speededness; the rate at which questions are presented and the kinds of responses that are required is an equity issue. He suggests that changing the rate of responding required by individual changes the pattern of responses of ethnic, gender, and cultural groups. Is this because the ethnic groups/gender groups use different pacing strategies or that because less able students are generally slower on paper-and-pencil tests and that gender and ethnic group distributions differ on the ability that speededness impacts the groups differently? Linn (2000) asks, what is the mechanism? And why does rate of question response influence the performance of ethnic, gender, and cultural groups? A major mechanism governing conditions where the examinee expects failure yet hopes for success is heightened anxiety. Heightened anxiety typically causes examinees to devote some of their reasoning capacity to the anxiety and less capacity to selecting from their repertoire. Do learners select more conservative or safer strategies under vulnerable conditions, or do they make poor choices? Linn's (2000) spatial reasoning studies

indicate that heightened anxiety might cause the examinee to check to be sure they are following procedures and waste time in checking and rechecking their behaviors. In high-stakes testing these measures waste valuable testing time. Safe practices increase the time required to solve individual problems and stand in the way of achieving a higher score.

Despite the denial of speededness by most testing organizations, Oshima (1994) argues that speededness remains a significant problem on many large-scale tests. However, his arguments are based upon research involving paper and pencils tests administered with precise time limits, where speededness was ascertained by calculating the percentage of candidates who did not reach the items toward the end of the test. As noted below (also see Luecht & Hadadi, 1998), computer-based testing (CBT) has greatly enhanced our ability to detect speededness and investigated related examinee pacing issues.

Definitions of Pacing Skills

Pacing skill may be defined as time-management proficiency exhibited by the examinees. That is, whereas speededness tends to be viewed as a characteristic of a test and associated test administration time limits, pacing skill is more clearly a test-taker attribute. It should be noted that speededness factors and pacing skill can, and often do, interact.

Empirical pacing research has been difficult to conduct until the advent of computer-base testing (CBT). With CBT, item-level response times can be accurately

collected for each examinee. That is, each examinee-by-item transaction can be timed, and the resulting timing variables stored as part of the examination records.

The role of pacing skill in measuring a prescribed proficiency is largely a validity issue. The more pacing skill enters into the mixture of knowledge and skills being assessed (intentionally or by accident), the greater is the onus on the testing organization to justify its inclusion.

Mislevy (1996) explains that a learner's state of competence at any given point is a complex constellation of facts and concepts, and the networks that interconnect them; of automated procedures and conscious heuristics, and their relationships to knowledge patterns that signal relevance; of perspectives and strategies, and the management capabilities with which the learner focuses his efforts. Tests need to present tasks that learners are likely to carry out in observably different ways: not only correctly as opposed to incorrectly, at what speed, with what products, looking for patterns of similarity, dissimilarity or independence across tasks that probe knowledge structures and problem-solving strategies. Poor or faulty pacing strategies may lead to item omissions or to rapid guessing. Self-regulatory functions such as confidence, persistence or test-wiseness, test anxiety, cognitive or personal characteristics may explain examinee performance (Martinez, 1999). The examinee's ability is a mixture of his mental or cognitive ability, the time devoted to the item, and the persistence devoted to finding the correct solution. A correct answer is ideally given if the examinee has sufficient ability and stays on task, working on the problem until the solution is found. An incorrect answer is more likely if the examinee shifts away from the problem and ceases to work on the task despite

whether he or she has sufficient ability or proficiency to correctly respond to the test item (Hornke, 2000).

Factors Affecting Pacing Skill in High-Stakes Examinations

The research literature covering test speededness and examinee pacing skill on high-stakes, large-scale tests *per se* is rather sparse. Different views of pacing skill and its consequences are presented in the literature, with most of the discussion focusing on characteristics of the test (e.g., item formats) that might interact with performance. There is little to no specific discussion of pacing skills. Rather, the subject of pacing skills is often relegated to the general category of "nuisance" characteristics within an examinee population. Nonetheless, the literature does at least indirectly suggest that pacing skill should be a matter of great concern.

In most testing settings—at least those in which speed is not a crucial component in the measured proficiency—examinees need adequate time to complete an examination without pressure. If pressures such as time limits are imposed, inaccurate responding or omissions may result, reflecting factors or examinee-test interactions that include more than the examinee's lack of knowledge or proficiency (Bugbee & Bernt, 1992).

Anastasi (1976) points out that speededness can cause undesirable consequences increasing examinee anxiety, encouraging guessing behaviors, and offering an obvious advantage to those skilled in pacing and applying related test-taking strategies. Under her perspective on speededness, "pacing" is viewed more or less as a consequence of the test

taking process, affected by off-proficiency-construct behaviors or psychological traits such as anxiety.

Martinez (1999) approached the issue from more of a cognitive-science perspective that relates general cognitive demands of various item types to speededness/pacing skill. Martinez recognized that multiple-choice (MC) items can be written to require simple cognitive processes such as recognizing a correct answer or recalling isolated facts, or they can be constructed to activate complex cognitive states and complicated performances, including understanding, prediction, evaluation, and problem-solving. MC items that merely require recall and recognition skills are far less cognitively demanding than items that mandate the retrieval of complex, associative information (i.e., declarative or procedural knowledge) from long-term memory. The relationships between cognitive demands and speededness/pacing skill are obvious. Test items that require very complex search strategies to access and organize memories stored from multiple learning episodes may be more time-consuming than those that only require low-level associative searches (e.g., retrieval of a common vocabulary word). However, MC items can also change the cognitive demands by interacting with examinee characteristics. A typical example is a MC item that penalizes high-ability examinees by including distracters that are partially correct under only certain conditions. The high-ability examinees "think too much" about the distracters.

Messick (1995) argues that multiple-choice items are limited in that cognition must lead to convergence on a single response selected from a small set of options. The ability to generate novel applications of scientific principles or to posit original

interpretations of solutions does not fit within this constraint. The essence of such competencies is lost when forced into this mold. Cognitive performances such as problem solving involve divergent production and complex performance. If adequate assessment of a construct requires cognitions of this sort, the construct may not be adequately measured in the multiple-choice format (Messick, 1995).

Examinee characteristics such as age can also affect pacing skill. Henderson (2004) states older examinees tend to do worse on analytical reasoning related to differences in test-taking speed. Psychometric literature cautions that the variable of age must be controlled in correlation studies because test-taking speed changes for the worse after early adulthood.

Bridgeman (2004) found that various assessments of mental speed have been proposed as measures of intelligence and evaluating semantic processing speed (response time). Although some indicators of mental speed may be related to existing measures of fluid intelligence, crystallized intelligence is unrelated to performance on a broad variety of speed measures composed of cognitive tasks (Roberts & Stankov, 1999).

Bridgeman (2004) also pointed out an indirect consequence of speededness and pacing on performance in an adaptive testing setting. Because of the dependence between item-selection and examinee performance, aberrant responses on examination scores induced by pacing or speededness artifacts can alter the subsequent items selected as well as the estimated proficiency of the examinee. Speededness effects early on tend to have greater impact than effects present later in an adaptive test because, near the end

of the adaptive test, the provisional ability estimate is less likely to change as a function of specific item responses.

Other researchers have taken a more direct line of research in looking at pacing and associated speededness effects. Embretson and Prenovost (2000) examined information processing efficiency by analyzing response times. Cognitive processes examined in this study were attention, mood, motivation and personality. Embretson and Prenovost demonstrated that attention has a significant impact on response time. They went on to suggest that attention involves task-relevant and task irrelevant thoughts. Relevant cognitive intrusions are extraneous thoughts that relate to the testing situation. Task relevant intrusions appear related to tension and worry. Greater attention relates to better task performance. Embretson and Prenovost went on to discuss the cognitive links between attention and arousal. Arousal is a complex cognitive system. An energetic arousal is a subjective experience of high energy or alertness where deactivation of this system is experienced as low energy or feeling tired. A tense arousal system leads to an experience of tension where as deactivation relates to a sense of calm. Tense arousal activation inhibits performance while energetic arousal facilitates performance. Personality traits may be related to cognitive performance. Longer response times were associated with task relevant intrusions. Response time was postulated to be the mental analogue of cognitive processes. Long response times indicate less effective processing and are associated with solving items. Response time represents a sum of processing speed, the number of processes attempted and other activities the examinee is engaged in while solving complex tasks. Long response time represents processing difficulties and

are positively associated with attentional failures. Longer response times are also associated with successfully solving matrix completion problems requiring correct discovery of several relationships. Response time, measured as persistence on task, is highly related the completion of the multiple relationships required to correctly complete the matrix. Embretson and Prevenost concluded that response times provide distinct information and may provide new information about individual differences. van der Linden (2005) has very recently expanded on this research area by developing a statistical modeling framework for response times.

Impact of Speededness on Test Performance in High-Stakes Examinations

Jones (1999) points out that it is important to understand the purpose of high-stakes assessments as well as the population of interest. Examples of high-stakes examination purposes include high-school graduation, college admissions, professional certification, and professional licensure. High-stakes assessments are usually given to gauge a student's proficiency or achievement relative to an expected standard or norm. Some high-stakes examinations offer unlimited retakes, other severely restrict or may even preclude retakes.

The testing environment for most professionally developed, high-stakes assessments is usually standardized so that most of the examinees (except for special accommodations granted by the testing agency) take the test under the same conditions. That is, the testing environment is made as equivalent as possible and every examinee is

provided the same amount of time to complete the assessment. Most high-stakes assessments have time limits ranging from two hours to 2 days (14 to 16 hours).

These fixed time limits have a pragmatic basis. The cost of testing "seat time" is usually based upon facility and related test administration costs. It is simply not feasible to allow every examinee unlimited time to complete his or her assessment.

Where there are time limits, test speededness and pacing become issues. In many high-stakes testing situations such as licensure and certification examinations, there is a conscientious and concerted effort to remove any extraneous testing factors that might influence the examinee (Bontempo & Julian, 1997).

As noted above, most high-stakes examinations allow a reasonably large block of time for examinees to complete a fixed or minimum/maximum number of items. For example, the NCLEX-RN® (which is the subject of this study) allows examinees to have almost six hours to complete between 75 and 265 multiple-choice items. Many examinees only take 75 items, at which point, a clear pass/fail decision is made. "Indeterminate" examinees proceed with testing until an accurate pass/fail decision can be made or until they complete 265 items, whichever comes first. The test is designed so that every examinee has, on average, more than one minute to answer each item and therefore may only be speeded for the subset of test takers who take all 265 items. The 2005 NCLEX-RN® Examination Candidate Bulletin states that the examinee needs to maintain a steady pace spending one minute on each item. That recommended pacing rate allows the examinee to easily complete the examination in the allotted time if all 265 items need to be administered in order to make a pass/fail decision.

The current time limits for the NCLEX-RN® only came about following some speededness research by Bontempo and Julian (1997). They found that examinees who take the minimum number of items (75 items) seldom run out of time. Only examinees who perform close to the passing standard, and who must take a longer examination, are at risk of running out of time. Bontempo and Julian further predicted that if the early-finishing examinees had been required to take the maximum number of items, and if their pacing levels did not change, over one-third of those examinees would have also run out of time. To come to this conclusion, Bontempo and Julian (1997) used a large sample of over 86-thousand NCLEX-RN® examinees. They reported a mean response time rate of 61.8 seconds per item. The average examinee could therefore be expected to finish 265 items in 4 hours and 33 minutes. Under the previous NCLEX-RN® time limits of 4 hours 45 minutes, Bontempo and Julian predicted that 80 percent of examinees would answer 229 items and 99 percent of examinees would answer 149 items. Therefore, given the opportunity, only 63 percent of the examinees would finish a 265 item examination in 4 hours and 45 minutes.

Bontempo and Julian (1997; also see Julian and Bontempo, 1996) also found that the examinees tended to increase their response rate after the initial 75 items. Note that this finding resulted from the examinees designated as neither clear passers nor failers. Eight-eight percent of the examinees increased their response rate by 12.5 seconds per item and 91 percent increased their response rate on the last 50 items by 18.6 seconds per item. The researchers suggested that the initially slower examinees increased their speed while faster examinees did not increase their speed. Luecht and Hadadi (1998) had

similar findings for a field trial of the United States Medical Licensing Examination. Bontempo and Julian reported that 66 percent of the examinees increased their speed after the initial 75 items, while 77 percent increased their speed for the last 50 items. Ultimately, Bontempo and Julian (1997) estimated that only about five percent of examinees were affected by the time limit. This was similar to the three percent number that Julian and Bontempo (1996) reported in an earlier study. Nonetheless, these studies jointly suggested that the NCLEX-RN®, as a 265-item fixed-length computer-adaptive test (CAT) was slightly speeded for a small percentage of examinees under the four-hour-and-forty-five-minute time limit. As a result of this research, the time limits for the NCLEX-RN® were increased to 5 hours and 45 minutes in 2005.

However, there are still pacing issues to consider for the NCLEX-RN®. Because response times are longer for items near an individual's ability level when taking an adaptive test (i.e., the "challenge" factor is approximately equal for all examinees on all items), an adaptive test may indirectly impose an element of test speededness by design. Overton and Harms (1997) actually found a reduction in test length did not lead to a decrease in testing time.

Collection of Response-Time Data on Computer-Based Tests

Computer-base test (CBT) administration allows researchers to collect timing data on every examinee-by-item transaction. These times can be averaged across all items and called "item-response latencies", or across examinees, in which case they become a measure of pacing skill. Timing data can also be analyzed conditionally—that is, for

examinees at different proficiency levels or across the course of the examination—which provides interesting options for modeling speededness and examinee pacing.

Pacing and Speededness for CAT

Computer-adaptive testing (CAT) applies another dimension to the speededness issue by modifying the difficulty of a test as a function of an examinee's apparent (provisional) ability or proficiency level. The adaptive item selection mechanism in CAT actually equalizes the "challenge" (i.e., the probability of a correct response) for each examinee, but may also add to the cognitive load—especially for higher ability examinees who are accustomed to confidently and correctly answering a majority of items. Assessments should include efforts to reduce the mental effort and fatigue associated with performance, and reduction of performance deterioration related to stress. Assessment of appropriate action control and self-regulation, i.e,. coping strategies, is important as well. CAT puts a new twist on the assumption of non-speededness by changing the difficulty of the test items that are presented. Item difficulty increases and decreases according to the estimated proficiency level of the examinee and the issue of response time becomes time used with the examinee facing progressively more difficult items with less and less time to complete the items. The issue is no longer one of omitted or not reached areas because omitting items is no longer an option; therefore, even when the intent is not meant to be speeded, the result is a speeded examination. An attempt at incorporating speededness constraints into a CAT assembly algorithm has been proposed by van der Linden, Scrams, and Schnipke (1999).

That does not imply that time limits should be eliminated in CAT. Examinee seat time at a commercial testing center—the site of choice for many professional certification and licensure testing programs that have migrated to CBT—is extremely expensive and time limits are necessary. The question, however, is, "Are the established time limits contributing to differential test speededness that affects test scores, and, by extension, score validity?"

Bridgeman (2000) states that time limits may contribute to the validity of an assessment but raise equity issues if the limit is imposed for administrative convenience rather than an essential part of what the test in measuring. Bridgeman and Cline (2004) concluded that higher ability students tend to take substantially longer than lower ability student because higher ability students are administered more difficult items. Thus, it may be that the expected relationship between cognitive ability and pacing skill is reversed for a CAT; that is, the test becomes more speeded for higher ability students. Test taking strategies are also confounded by the fact that most CAT implementations prevent the test taker from reviewing previous answers, as well as from omitting answers. Pacing skill, as a examinee-controlled attribute, is therefore complicated by the forced answering and the changing item difficulty implicit in most CAT implementations. Bridgeman (2004), in his work with the Graduate Record Examination-Analytical, found that an examinee who worked at the mean rate for the first 20 items would take 71 minutes to complete the test or 11 minutes longer than the 60 minutes allowed. For higher ability students generally working faster but taking harder and more time consuming items the entire exam would take 77 minutes. Earlier, Bridgeman (2000)

explored the possibility that higher ability examinees might have lower scores if they have a greater number of items that take longer to complete. They found that examinees that took the test items that should have taken longer got higher test scores. There was no evidence that examinees taking long tests were disadvantaged in terms of total scores, at least when using IRT scoring to take difficulty into account. However, Bridgeman (2000) recommended that some index of estimated solution time be incorporated in the item selection algorithm so that no individual gets more than a fair share of time-consuming items. This recommendation was also proposed by van der Linden, Scrams, and Schipke (1999) and by van der Linden (2005).

Bridgeman, Cline and Hessinger (2003), again using GRE data, found that speededness is a potential threat to the validity of the test because examinations should reflect the intellectual power of the examinees work, not the rate at which examinees work. They concluded that the variation among examinees in the rate of response to test items constitutes an irrelevant source of difficulty in test performance. Nonetheless, the capacity to work rapidly or process information efficiently may be relevant to academic ability. Speededness may be an irrelevant or relevant indicator of academic ability, to the extent that scores are dependent on time is of interest to potential score users. Random guessing substantially drops the ability estimate because the scoring algorithm assumes the missed items reflect lower ability rather than random guessing. It would be useful to know if any test taking strategies or test wiseness might help students use their time more effectively.

Plake (1999) states that with CAT it is possible for one test-taker to be administered the most challenging questions from the item pool while another could be administered easier items. The scoring model for the situation determines a person's score not only by the number of questions answered correctly but accounting for the level of difficulty of the test items answered. The stronger candidate answering more difficult items correctly than her weaker friend will receive a higher overall score even if they get the same number of items correct. In computerized adaptive testing the probability of a correct response is a function of the ability level of the examinees (latent proficiency), the item difficulty, item discrimination, and the probability that an examinee with very low ability will be able to answer the item correctly due to chance alone (guessing). In computer adaptive tests all critical dimensions of testing in terms of item content coverage, item selection, and balance of differing levels of cognitive load must be anticipated and the item selection algorithm must take all of these into account, as well as item difficulty, discrimination and guessing, as the items are sequentially selected. There is a possibility that the examinees will be administered examinations with items that differ in substance, not just difficulty, discrimination, and the propensity to be guessed correctly. Because the examination is developed dynamically, the item selection algorithm makes its selections based on pre-calibrated item characteristics; items that are equal on these calibrations are considered equivalent in the algorithm selection. Item selection algorithms do not consider length, cognitive complexity, or number of steps. The possibility exists for pairs of items to be equivalent in calibrated features but differ dramatically on other features. One examinee could be administered an examination that

is loaded with more cognitively demanding, time consuming questions than another equally able examinee. Test dynamics can unfairly affect the examinee that will use more cognitive energy and potentially be more time challenged than an equally able examinee that got equally calibrated but differentially less cognitively challenging questions.

Issues of fairness arise when high stakes decisions are based on the results of such differences by psychometrically equivalent tests. High stakes tests should be able to assure that equally able examinees receive tests equal in cognitive demand. Smith (2000) used Graduate Record Examination (GRE) and Graduate Management Admission Test (GMAT) data to show that it may be possible for two examinees of equal ability to be administered items that are equal in terms of item difficulty, item discrimination and propensity for guessing correctly, but one examinee will be administered an examination that was more cognitively demanding. The difficulty level of problem solving and critical reasoning items increase the number of steps required to answer a problem correctly. Smith (2000) went on to demonstrate a quadratic relationship between response time and difficulty and critical reasoning. The response time increases monotonically with the cognitive demand. Item difficulty and examinee ability may be highly correlated on an adaptive test and evidence suggests the computer adaptive tests do not control for needed response time. The assumption here is that more difficult items require greater cognitive loads, which leads to an increased processing time.

Modeling Speededness and Pacing

The literature provides differing perspectives about the nature of the interaction between test speededness factors and examinee pacing skill. Some of these differences are purely theoretical, others stem from empirical research.

Following a strictly statistical efficiency rationale, Mislevy and Bock (1989) suggested that adaptive testing reduces the time needed for testing. That is, a CAT is theoretically more time efficient because equally reliable tests can be constructed with fewer items, requiring less time. Therefore, speededness is essentially ignorable. Presenting a contrasting perspective, Overton and Harms (1997) suggested that adaptive testing required more time than the paper-and-pencil testing (PPT). They found that items chosen under the adaptive testing conditions appeared to be more difficult for the examinees because more items were at the appropriate levels for the examinees. Their findings recommended that the test should be shortened (i.e., fewer items administered to the examinees) in order to retain the same amount of total-testing time. They further found no evidence that the adaptive testing enabled a more accurate selection based on equivalent total-testing times. Overton and Harms analysis suggested that average response time per item depends on nature of the relationship between the item difficulty and the examinee's apparent proficiency. They concluded that items requiring extensive problem-solving and analytical-reasoning steps tend to slow down the examinee. This issue of depth-of-processing/item complexity, pacing, and test speededness would definitely appear to confound the increased efficiency/time reduction assumed by Mislevy and Bock (1989). Scrams and Schnipke (1997) proposed using response times

in standardized tests to compare speed and accuracy as different components of proficiency. Although speed and accuracy have long been used in employment-based performance assessments such as typing tests and in athletic competitions, the idea of considering pacing as a legitimate and valid measure of proficiency has not been widely accepted in educational and psychological testing.

Psychometricians and psychologists tend to look at human performance as involving both speed and accuracy aspects, but assessment of these attributes have decidedly different goals. Cognitive psychologists are commonly interested in the mechanisms of cognitive processing. Information-processing speed can be used as a surrogate for depth of processing, efficiency or generalization of search strategies, etc. In contrast, psychometricians tend to have a [somewhat dated] view of psychology that focuses on proficiency as demonstrated knowledge or skill in a prescribed content domain such as "mathematics" or "reading comprehension". Many psychometricians are essentially unconcerned about any data other than the scored item responses they consider the measure(s) of proficiency. Scrams and Schnipke (1997) suggested taking advantage of both approaches—that is, using response accuracy and response speed to provide separate measures of performance. Computer-based tests, as noted earlier, allow us to record cumulative response times for each item. This provides a new measured variable for each examinee and for each item. Considerations changing the relationships between the response-time and performance-related variables may have important implications for research involving different ability groups, different ethnic groups, etc.

Earlier psychological research did consider modeling pacing skills and modeling speededness but was hampered by the lack of accurate ways to collect timing data. Donlon (1980) looked at four methods of evaluating test speededness: (1) the Gulliksen approach; (2) the Cronbach and Warrington approach; (3) the Stafford approach and (4) the Swineford approach. However, because of the lack of specific item-level response time data, all four approaches assume a constant rate of responding throughout the exam. That assumption is probably false. In fact, work by Luecht and Hadadi (1998) specifically demonstrates that constant timing rates of response are not evident and vary by proficiency level of the examinees (perhaps due to differential pacing skills) and other factors.

Yamamoto and Everson (1997) proposed a HYBRID model to estimate the proportion of examinees that switch to a guessing strategy as each item is sequentially administered. Their model assumes that, if an examinee switches to a guessing strategy halfway through an examination, the probability of making correct responses on the subsequent items no longer adheres to the IRT model. Changes in the conditional probabilities of a correct response under a particular IRT model are therefore most noticeable for the more proficient examinees. The HYBRID model incorporates the notion of speededness of an examination by directly modeling changes in conditional probabilities while analyzing omitted responses. This model-based approach effectively looks at the impact of potential speededness on estimated item parameters and is very useful for examinations in which item difficulty systematically increases. This model characterizes examinee strategy use, detects extraneous strategy influences in estimated

model parameters and incorporates partial knowledge of latent classes. The model is limited by an inability to handle random responses in the first part of the test or to deal with random guessing versus the induced rapid guessing that occurs when examinees run out of time. Like the Schnipke and Scrams' (1997) mixture model, the HYBRID model has not yet been used in any operational testing settings.

Bergstrom, Gershon and Lunz (1994) proposed using hierarchical linear modeling to compute variance components or facets for different item and test factors. Using increasingly complex models, Bergstrom, et al, computed variance component estimates for within-person and between-person facet models. They found that while relative item difficulty, item correct, item sequence, item length, item content and position of the correct answer did predict detectable amounts of variance in response times, these effects did not vary across persons. In general, they found that more difficult items took longer to answer. They also found that examinees spend more time on items they got incorrect than on items they got correct. Increased total item-text length and increasing item difficulty resulted in proportionately increases in response times as well. Interestingly, examinees took longer to respond if the correct answer is A, B, or C than D, possibly due to the requirement to backtrack and re-evaluate the distracters relative to one another. Some other findings were: (a) that test anxiety significantly predicted variance between examinees—hence anxious examinees take longer on their tests; (b) items answered earlier in the test took longer response times—perhaps due to a caution or warm up period—and examinees increased their rates of responding as they progressed through the test; and (c) the final estimate of ability was not a predictor of response time.

Halkitis, Jones, and Pradhan (1996) found that accurate projections of testing time required by examinees was important to insure that unintended effects due to response speed did not compromise score interpretations in computer administered tests. Their study found that the determination of testing times for credentialing examinations has a significant impact on test validity, testing efficiency and resource allocation. They developed a regression model based on item difficulty, item discrimination and word count that explained 50 percent of the variance in response time.

Item response theory (IRT) modeling strategies have been proposed to deal with pacing as a separate skill. Thissen (1983) was one of the first to formally develop an IRT model for response times. Unfortunately, the parameters of Thissen's model are difficult to estimate and the model has not seen operational use. Scrams and Schnipke (1997) used a mixture model of response times and item-level performance scores to estimate proficiency as having two component traits: (1) response accuracy and (2) pacing/response speed. They investigated two different types of speed-accuracy relationships: (a) a within-examinee relationship modeled as a speed-accuracy tradeoff and (b) an across-examinee relationship that was modeled as a covariance between the two traits. Examinees are assumed to require some minimal amount of time to respond above a chance guessing level. Scrams and Schnipke (1997) used a minimal encoding/intuition response time of 10 seconds in their model. After this minimal processing time—modeled as the lower asymptote of the response time function—the performance given infinite amount of time moves toward the upper asymptote. Once the examinee reaches the upper asymptote in terms of his or her performance (i.e., given his

or her estimated accuracy or proficiency trait), further processing time has little effect on performance. The concepts of minimal processing time, a monotonic increase in accuracy as a function of processing time and upper asymptote appear to be psychologically useful constructs for considering response time in a testing context.

Scrams and Schnipke (1997) provided a psychological rationale for their approach to modeling speed and accuracy relationships. Examinees have direct control over their choice of test taking strategies and their response time but only indirect control over accuracy. Strategies selected affect the speed-accuracy relationship and increased processing time generally results in better performance. Strategies can affect the minimal processing time, the solution time and accuracy. Examinees choose their response speed based on their selected strategy. Measures of response time alone do not provide enough information to predict the examinee's performance. Speededness is linked to *rapid guessing behaviors*. *Rapid guessing behaviors* are indications that the examinee is working rapidly through the items—that is, changing his or her pacing strategy—skimming for keywords, answering based on partial information or guessing the answer. In contrast, *solution behaviors* demonstrate that the examinee reads and answers each question carefully (Schnipke & Scrams, 1997). Schnipke and Scrams (1977) developed an estimate or an index of what they termed *rapid guessing behavior*. Strict time limits can cause examinees to focus on processing speed to the detriment of response accuracy. As time begins to expire, examinees may begin to respond with minimal processing.

Luecht and Hadadi (1998) suggested that rapid guessing phenomena may be spontaneously activated when a test taker realizes that he or she is running out of time. In

that sense, rapid guessing is not the same as "random guessing" in that it does involve some strategic attempt at correctly answering the remaining items. However, the response patterns that emerge in the data are often too noisy to provide usable measurement information about the examinees' proficiency levels. Identifying rapid guessing responses is necessary to eliminate those response times that are statistically higher than expectation, using a likelihood ratio. Luecht and Hadadi (1998) attempted to expand on Schnipke and Scrams work by providing empirical criteria for determining the point where rapid guessing might be activated, as well as some of the first attempts to model differential trajectories in response times and performance occurring over the course of an examination. It should be noted that Schnipke and Scram's (1997) model has not been operationally used on any examinations to date.

Van der Linden and van Krimpen-Stoop (2003) developed a log-normal model for response times to check times for aberrance in examinee behavior on computerized adaptive tests. The detection rates of the Bayesian checks outperformed those for the classical checks but with more false positives. In all conditions, initial estimate of theta (θ) was set equal to zero. Unfortunately, the results from the empirical examples were more discouraging than anticipated. This method is not recommended for checks at the level of the individual response.

Van der Linden (2005) has extended this log-linear model for response times and helped to distinguish between ability and speed as separate, intentional proficiency factors, versus speededness as a nuisance factor. Van der Linden's response-time model provides ways of estimating specific item and person parameters that respectively

characterize item-level speededness for a population and examinee-level pacing proficiency. The estimated item timing parameters can subsequently be constrained using automated test assembly algorithms or heuristics (van der Linden, 2005) to ideally minimize the influence of speededness when it is considered to be a nuisance factor.

Bontempo and Julian (1997) attempted to model speededness in a variable-length computer-adaptive test (CAT). They focused on the NCLEX-RN® and suggested three methods of analyzing speededness: (1) computing an examinee response rate; (2) investigating changes in response rate within an individual's examination; and (3) exploring changes in examinee response rate across repeated examinations. Bontempo (2000) went on to develop a probabilistic model assessing speededness. Bontempo's rationale was that, given the pass/fail outcomes from tests like the NCLEX-RN® examination, examinees whose proficiency scores place them near to the cut score must have their test score estimated with greater precision than examinees whose proficiency is well above or well below the cut score. (That is, it takes less measurement precision to classify clear passers and clear failers than to classify individuals in the region of the pass/fail cut score.) At the same time, administering these "indeterminate" examinees more test items to increase the decision accuracy of the test may put them at risk for exceeding the time limits. In the NCLEX-RN® context, the problem does not extend to clear passers and clear failers. In fact, each year, less than 5 examinees fail to complete the minimum 75 items within the required time limit. In any case, examinees that need the maximum number of items may change their pacing strategies in an effort to complete the test in the remaining time. Under Bontempo's (2000) probabilistic model,

response times were calculated as the natural logarithm of the item-level response times, and a speed rating scale score for each item was developed by averaging the log-times over examinees. Then, measures of item duration were developed and item difficulty was calculated. Bontempo's model encountered model fit and convergence difficulties in the presence of noisy data (e.g., if examinees guessed or took an extraordinary long time to answer particular questions). Ultimately, Bontempo found that there was no change in the ability estimates as a result of apparent changes in pacing (speed).

Rationale for the Present Study

Most of the research to date clearly recognizes the seriousness of speededness and, for models like Thissen's (1983) model, Scrams and Schnipke's (1997) mixture model, or van der Linden's (2005) response-time model, may actually attempt to compute a "speed" parameter that suggests how skilled the examinees are at pacing. However, what is missing from this research is the recognition that *pacing may change over the course of an examination*. Scrams and Schnipke's work on rapid guessing, as well as Luecht and Hadadi's (1998) work attempted to detect the activation point for the rapid guessing. Yamamoto and Everson's (1997) work with the HYBRID model likewise tried to more-or-less isolate the ability-producing response space from the response space that involved what amounted to rapid guessing.

However, rapid guessing is somewhat of an extreme response state that is only activated when a pacing strategy fails. It may be that pacing involves more gradual changes over the course of an examination. It is therefore useful to consider pacing as a

"trajectory" that may be linearly or nonlinearly decreasing over the course of the examination.

The present research study attempts to conditionally understand the interaction between test items and examinees over the course of the examination, using two potentially useful modeling methods: hierarchical linear modeling and latent growth modeling using structural equation modeling (SEM). These modeling methods, which are described more completely in Chapter 3, have not been previously applied in the context of response-time research and definitely not for a high-stakes assessment that incorporates an adaptive item selection algorithm. This study helps to shed some light on a basic question. That is, how do pacing trajectories look in a computer adaptive test for examinees at different levels of proficiency? Both hierarchical linear modeling and SEM-based growth modeling are appropriate techniques for exploring the nature of "trajectories". In the present study, those "trajectories" refer to the changing interactions in response times and performance between items and examinees over the course of the NCLEX-RN® examination.

CHAPTER III

METHODOLOGY

This study compares several methods for statistically modeling response-time trajectories on the NCLEX-RN®, a high-stakes computer-adaptive test (CAT) administered to registered nurse licensure candidates in the U.S. and its territories. This chapter summarizes relevant characteristics of the NCLEX-RN® and the examinee population. The chapter also details the trajectory-modeling methods being compared, software employed for the analyses, and details about the actual comparative analyses undertaken.

Instrumentation

The NCLEX-RN® examination is designed to test knowledge, skills and abilities (KSAs) that are considered to be essential to the safe and effective practice of nursing at the entry level in the United States. The NCLEX-RN® examination results are a key component used by state and other jurisdictional boards of nursing to make decisions about licensure as a registered nurse (RN).

The NCLEX-RN® examination is constructed and administered by the National Council of State Boards of Nursing (NCSBN). The computerized adaptive forms of the examination were implemented in April of 1994, replacing the previous paper-and-pencil examinations. The NCLEX-RN® is a variable-length computer-adaptive test (CAT).

Every examinee takes a minimum of 75 adaptively administered items, at which point one of three decisions occur: (1) the examinee has clearly passed the examination with a prescribed level of statistical certainty; (2) the examinee has clearly failed the examination with a predetermined level of statistical certainty; or (3) the examinee's performance is "indeterminate" and the examination must continue in an attempt to resolve the indeterminacy about the examinee's pass/fail status. The maximum number of items that any examinee can take is 265—or after 5 hours and forty-five minutes of testing time has elapsed, whichever comes first. However, many examinees classified as "indeterminate" after 75 items can be classified as passers or failers before taking all 265 items.

The NCLEX-RN® items are administered using a computer-adaptive testing (CAT) algorithm. Under this type of item-selection algorithm, a small set of "random" items is initially administered to the examinee from a pre-calibrated item bank. Pre-calibration implies that the statistical properties of each item have been determined through item pre-testing relative to the proficiency of the target population of RN candidates. The test taker's proficiency score for the initial test of items is estimated and an item is chosen having a difficulty targeted to the provisional proficiency score so that the examinee has approximate 50:50 odds of correctly answering the item. Other content balancing and exposure control criteria are also incorporated into the CAT. The examinee's provisional proficiency score is updated after he or she answers that next item and a new item is adaptively selected to be as close as possible in difficulty to that new provisional proficiency score. The process continues until the examinee's score can be

determined with 95 percent statistical certainty to be above or below the established pass/fail cut score for the NCLEX-RN® examination.

The majority of the NCLEX-RN® items are multiple-choice (MC) items with a strict format of grammar, phrasing and language usage. All items are scored right/wrong; there are no partial-credit items. The NCLEX-RN® items are written by subject-matter experts and are extensively reviewed by item writing editors, sensitivity review panels and committees of nursing experts. Items must maintain strict psychometric criteria to remain in the item banks, including model-fit and differential-item-functioning criteria.

As noted above, content constraints are incorporated into the CAT item-selection algorithm to maintain a balance of items. The content blueprint requires proportional distribution of the items across several broad areas of nursing knowledge related to the health needs of clients as well as assessing the candidates understanding of integrated processes fundamental to nursing practice. The content categories include management of care, safety and infection control, health promotion and maintenance issues, and four sub-areas under the heading of "Physiological Integrity". Each NCLEX-RN® examination also includes 15 unscored, experimental pretest items. The pretest items are evaluated for possible use in future item banks.

The use of the CAT algorithm implies that NCLEX-RN® is differentially difficult for the examinees, depending on their demonstrated proficiencies. More able candidates receive a more difficult test than less able candidates. Item difficulty is factored into the scoring so that all examinees are scored on a common metric. In addition, the variable-length testing implies that examinees may need to employ different pacing strategies for

the first 75 items, versus for the remaining items, should they be required to continue testing after the initial pass/fail gating decision has been made.

For purposes of this dissertation, the examinees are therefore classified into three broad categories that reflect both the test length conditions of the NCLEX-RN® examination, as well as the examinees' apparent proficiency level: (1) clear passers at 75 items; (2) clear failers at 75 items; and (3)"indeterminate" examinees at 75 items.

Sampling

Over 120,000 examinees take the NCLEX-RN® examination every year. For this study, a random sample of 4,415 first-time takers was selected from the examinee population of records for the 1999-2000 test administration year. The National Council of State Boards of Nursing (NCSBN) provided the data.

The examinee records did not identify specific candidates. The data identified the candidates' gender, ethnicity, and pass/fail status. In addition, for each examinee, data included the number of items completed, the response(s), and the cumulative response time on the item, the sequence number, and an indication of whether the item is scored. This study focuses primarily on the timing information provided in the response records.

Data Preparation

Using recommendations made by Hadadi and Luecht (1998), the examinee data were organized to allow analysis in two ways: by item and by examinee. The item-level timing requires aggregation of the response timing data by item identifier. The

examinee-level data were aggregated using mean response times in blocks of 15 items. The blocking acts as a smoothing function for the response times. The aggregated block-level response times provides a straight forward way to model the trajectories within and across examinees. Therefore, examinees taking only 75 items had five blocks of items, examinees who continued testing had six or more blocks of data.

Research Questions and Comparative Methods of Analyzing Response-Time Trajectories

The fundamental question addressed by this study is, What do pacing trajectories look like in a computer adaptive test for each of three groups: (1) clear passers at 75 items; (2) clear failers at 75 items; and (3) "indeterminate" examinees who took between 76 and 265 items? A follow-up set of questions focus on the comparison of two statistical modeling methodologies: hierarchical linear modeling (HLM) and growth modeling using structural equation modeling (SEM) to help understand the pacing trajectories. The follow-up questions are: (a) Which method provides better fit, and (b) Which method provides the more interpretable findings?

Tangential questions include: Does an examinee's pacing strategy vary at the beginning, in the middle or near the end of the test? Would a pattern develop if response times were examined in blocks of items? What is the optimum number of item blocks? How many examinees are in danger of running out of time in the new six-hour testing paradigm and do they change their pacing in order to attempt to finish the test?

What pacing strategies can be identified that could provide information for computerized testing algorithms?

As noted above, this research study compares two trajectory-modeling methods. The hierarchical linear (or multilevel) modeling framework (e.g., Bryk & Raudenbush, 1992) is a useful conceptual framework for analyzing trajectories and provides a flexible set of analytical tools. One set of applications focuses on between-group differences in the trajectories. This study used proficiency levels and pass/fail status as a means of classifying the examinees in four groups. That is, each examinee can have an individual response-time trajectory.

The HLM computer program (Bryk & Raudenbush, 1992) was initially designed for analysis of the statistical modeling of two- and three-level data structures, respectively. A two-level model consists of two sub-models at level 1 and level 2. For example, if the research problem consists of data on students nested within groups, the level-1 model would represent the relationships among the student-level variables and the level-2 model would capture the influence of group-level factors. Here, a hierarchical linear model (HLM) using a two level model will allow a model of pacing trajectory to be identified for each examinee and by using the proficiency space we are now able to compare "mean" trajectories for each proficiency group.

Bryk and Raudenbush (1992) state that HLM using a two level model can represent individual change phenomena. In Level 1, each examinee's development is represented by an individual pacing trajectory that depends on a unique set of parameters. The individual growth trajectory parameter then becomes the outcome variable in the

Level 2 model that depends on group membership. The observations for each individual are viewed as nested within each person. Treatment of multiple observations as nested allows the researcher to proceed without difficulty when the number and spacing of time points vary.

An example of how Bryk & Raudenbush (1992) represent the repeated observations model (Level 1) and person level model (Level 2) of a two level hierarchical model is shown in Figure 3-1.

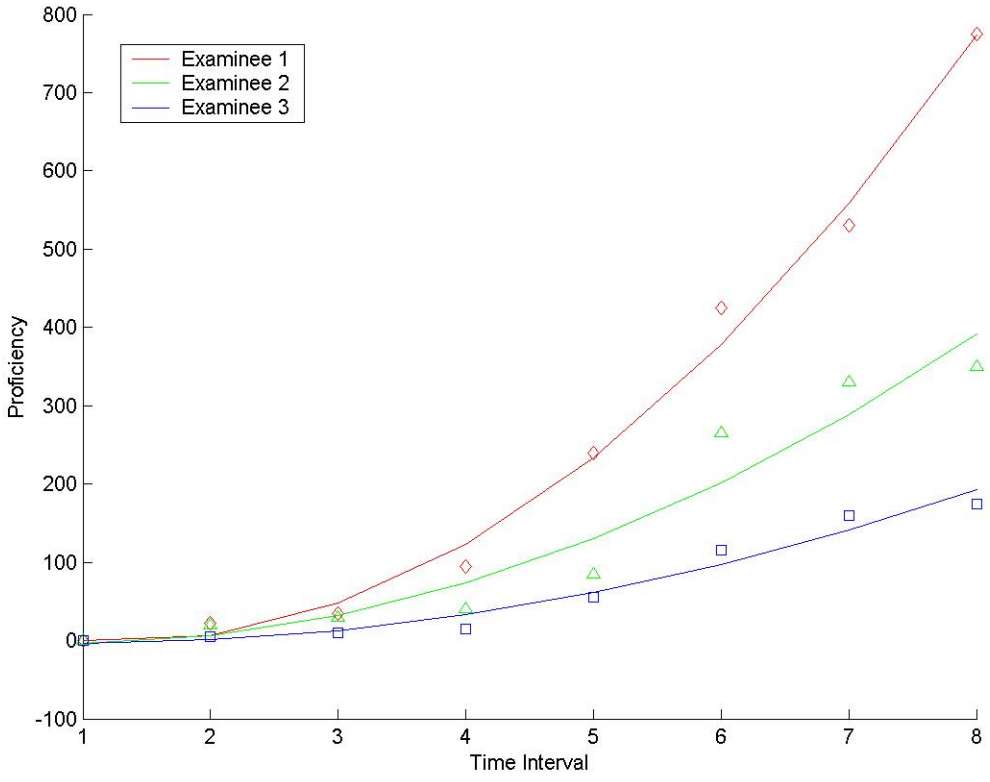

Figure 3-1. Example of individual growth trajectory using HLM

Growth modeling and change modeling can also be carried out using structural equation modeling (SEM). The SEM framework focuses on unobserved latent variables. Latent change analysis models the change in the latent variable means over time or conditions. Latent growth modeling models change in the rate of change (i.e., the developmental growth in the latent variables over time or conditions). A latent intercept and a latent slope are considered the factor variables (of change). The first factor is the initial latent status or intercept. The second factor is the slope, which is allowed to change over time points. If the growth is linear, that trajectory will be obvious as a constant rate of change. If the growth (or decline in the case of response time data in a testing context) is nonlinear, the SEM model provides estimates of the rate of change via the slopes. Statistical tests can be used to evaluate whether the change in rates are different in the population.

Duncan, Duncan, Strycker, Li & Alpert (1999) provide an example of a latent growth curve model with two factors: factor 1 (intercept) and factor 2 (slope). The intercept (factor 1) is a constant for any given examinee across time with a fixed value for factor loading of one of the repeated measures. The intercept for a given examinee has the same meaning as the intercept of a straight line on a two-dimensional coordinate system. The intercept factor presents information in the sample about the mean (M_i) and variance (D_i) of the collection of intercepts that characterize each examinee's growth curve. The slope (factor 2) represents the slope of an individual examinee's trajectory. It is the slope of the straight line determined by the repeated measure. The slope factor has a mean (M_s) and variance (D_s) across the whole sample estimated from the data. The two

factors are allowed to co-vary and the error variance terms are assumed to be zero. Raykov & Marcoulides (2000) depicted this growth model as shown in Figure 3-2.

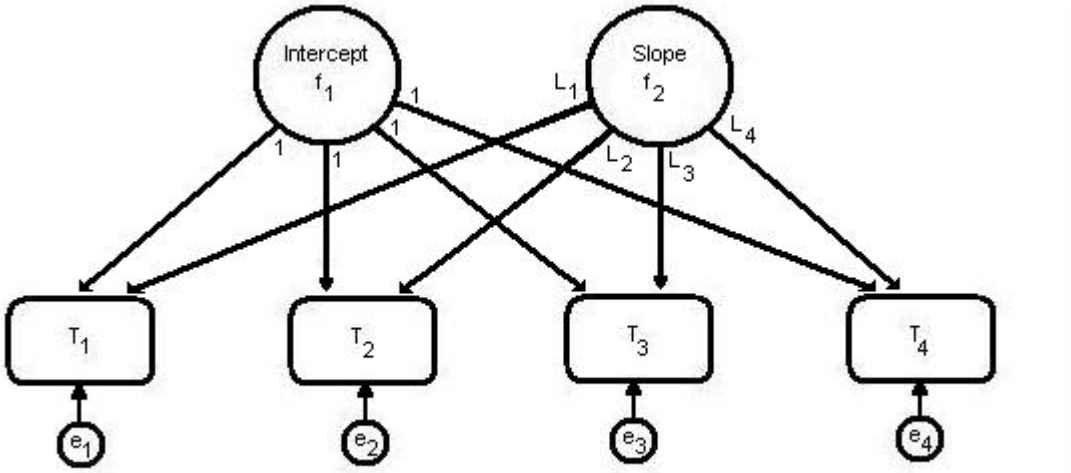

Figure 3-2. Path diagram for a latent growth model.

The SEM-based growth modeling approach requires certain strict identifiability conditions to allow the parameters to be uniquely estimated. It is possible that unstable solutions can result if these conditions are not met (see Stoolmiller, 1995).

Both the HLM and SEM-based growth models attempt to measure the pacing trajectory on the NCLEX-RN® for 15-item blocks. Therefore, every examinee has a minimum of five data points to identify their individual pacing trajectory. The models will identify the intercept and the slope of the individual's trajectory. Each model will be carried out conditional on proficiency level (ultimate pass/fail status) and relative test-length (75 items versus 76 or more items).

Software Resources and Data Analysis

In this research, hierarchical linear modeling (HLM) and multilevel growth structural equation modeling (SEM) were performed with commercially available software. The HLM analyses were carried out using HLM 5.0 for Windows (Raudenbush, Bryk, Cheong and Congdon, 2003). The SEM analyses were conducted using LISREL 8.7 (Jőreskog & Sorbom, 2004). Data manipulation, descriptive analyses, and graphical results were carried out using a combination of other statistical and spreadsheet computing software.

CHAPTER IV

ANALYSIS AND RESULTS

This study compares two methods of modeling response-time trajectories in a high-stakes computer-adaptive test (CAT): hierarchical linear modeling (HLM) and latent growth modeling using structural equation modeling (SEM). The comparative aspects of the study were carried out using a data sample of over 4100 examinees who took the NCLEX-RN® (National Council of State Boards of Nursing) during the 1999-2000 time period. This chapter summarizes relevant characteristics of the sample and primary statistical results from the HLM and SEM analyses.

Descriptive Statistics

A random sample of 4145 examinees who took the NCLEX-RN® examination during the 1999-2000 time period is used for this study. As explained below, that sample was reduced slightly during the data cleaning phase of the analysis.

The sample is 86 percent female and 13 percent male, with one percent not responding to the gender question. Ethnicity is broken down as follows: 49 percent Caucasian, 11 percent African American, 18 percent Hispanic, 3 percent Asian Indian, and 5 percent Pacific Islander (14 percent "other" or non-respondents to the ethnicity question). The percentage of passing examinees in the sample is 54.6 percent.

Data Cleaning and Examinee Groupings

The examination data for an individual examinee consists of an item identifier, the examinee's raw and scored response(s) to the item and the response time. The latter timing data are recorded in cumulative milliseconds spent on an item.

The data set was reviewed for outliers. Seventeen examinees having response times of zero or exceeding 500 seconds were eliminated from further analysis. The final sample of data used for modeling therefore consisted of adaptive-test records for 4118 examinees.

For purposes of the analysis, the examinees are classified into one of three groups: (1) clear passers at 75 items (Pass-75); (2) clear failers at 75 items (Fail-75); and (3) "indeterminate" examinees who took 76 to 265 items (Indeterminate). It should be noted that the "indeterminate" examinees are eventually classified as passers or failers, but, for purposes of this study, are treated as members of the same proficiency group. Sample sizes for these three groups, respectively, were: $n_1 = 1006$, $n_2 = 792$ and $n_3 = 2320$. This study specifically looks at response trajectories for these three groups, where group membership represents a type of conditioning on proficiency level.

Proficiency level is somewhat confounded with the additional potential for speededness due to increased test length for individuals falling within the indeterminate group. The nature of the confounding between proficiency level and test length conditions on the NCLEX-RN® should be somewhat obvious, but readers might benefit from some additional explanation. The NCLEX bulletin (NCSBN, 2006) states NCLEX-RN® examinations can vary from 75 to 265 items long. The length of the examination is

determined by the examinee's responses to the items; that is, a CAT item-select algorithm is used to construct a unique test for each examinee, tailored to his or her apparent proficiency level. Once the minimum of 75 items has been met, a fundamental decision is made by the test delivery software to either stop testing—if a pass/fail decision can be made with 95 percent certainty as to the examinee's likely proficiency level—or to continue testing. Examinees who are classified as clear passers or clear failers at 75 items are allowed to leave the testing center. Examinees who continue to test have any remaining time within their allotted six hour testing session to complete up to 190 additional items (265 items in total). However, few examinees actually complete 190 more items because the decision rules continue to be applied after each subsequent item and the test halts when a pass/fail decision can be made with 95 percent statistical certainty.

Therefore, there are actually four examinee proficiency levels (lowest to highest): (i) clear failers; (ii) failers near the cut score; (iii) passers near the cut score; and (iv) clear passers. The confounding occurs because the examinees in the second and third proficiency-level groups get a longer test, from 76 to 265 items. For purposes of this study, examinees in the middle two groups are collapsed into a categories called "indeterminate at 75 items" (abbreviated as "Indeterminate")

Item Blocking and Baseline Trajectories

Regardless of proficiency level, the data points for investigating the response-time trajectories were computed by averaging the individual item responses within consecutive

blocks of 15 items. This block size was determined to be adequately large to provide an empirical method smoothing the response times across items of possibly varied difficulty and also yielded sufficient data points for modeling the trajectories. Blocks are denoted as "BLOCK0", "BLOCK1", etc., up to a possibility of "BLOCK17" for the examinees who actually took all 265 items[1]. "BLOCK0" is the initial, baseline block of items denoting the start of every subject's CAT.

Figure 4-1 presents a graph of the raw averages for the item blocks for the clear failers at 75 items, the clear passers at 75 items, and for the "indeterminate" group that continued beyond 75 items.

[1] Note that the average for "BLOCK17" is only computed using up to 10 items. The results for this block may therefore be less stable than for other blocks. Furthermore, some examinees may have fewer than 15 items computed in their final block average if the adaptive algorithm ended their test at a point other than at an item corresponding to a multiple of 15.

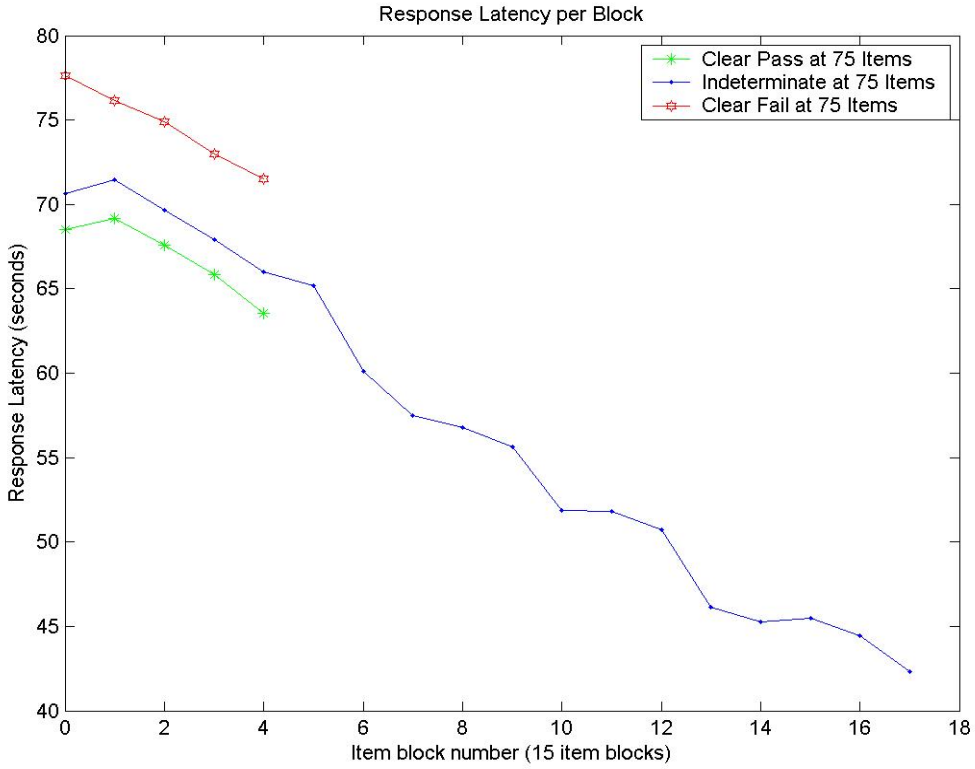

Figure 4-1. Raw response-time trajectories for three examinee proficiency-level groups (N=4045).

These raw data averages in Figure 4-1 demonstrate rather distinct differences in the pacing trajectories for each group. Examinees who are classified as clear passers at 75 items (Pass-75) have the shortest response times on average, Examinees classified as clear failers at 75 items (Fail-75) have the longest response latencies. Examinees classified as "indeterminate" at 75 items (Indeterminate) are between those pass/fail rates of response, being somewhat closer to the clear-passer group. It should also be obvious that all three groups decrease in their response times. The Indeterminate group has

somewhat a steeper apparent slope as the examination progresses. Table 4.1 has the actual values of the means and the associated standard deviations by proficiency group and broken down for each of the relevant item blocks.

Table 4.1 Descriptive Statistics for Response Latency Raw Data

Block	Clear Pass 75 Mean	Standard Deviation	Clear Fail at 75 Mean	Standard Deviation	Indeterminate at 75 Mean	Standard Deviation
0	68.49	24.85	77.63	29.02	71.18	24.96
1	69.20	23.21	76.16	27.17	71.88	24.32
2	67.61	21.80	74.89	25.62	70.28	23.59
3	65.96	22.09	72.99	25.09	68.42	21.59
4	63.55	20.84	71.53	23.45	66.43	21.07
5					65.58	19.53
6					60.43	17.42
7					57.72	16.02
8					56.98	15.41
9					55.74	14.69
10					51.90	13.01
11					51.90	12.38
12					50.80	11.94
13					46.10	11.07
14					44.52	10.07
15					42.30	12.103
16					45.23	10.91
17					45.50	10.56

The decreasing response times are not mirrored by relative performance. Figure 4-2 shows a graph of the examinees' percent-correct scores, computed within the 15-item blocks, for these same three groups.

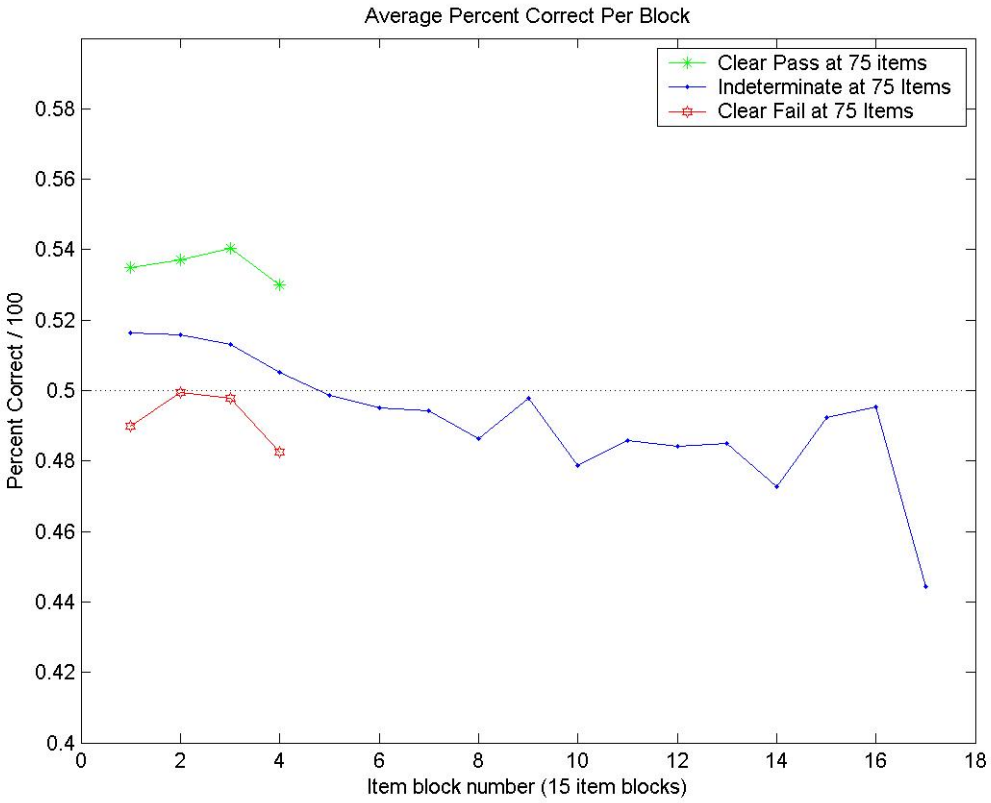

Figure 4-2. Average percent-correct scores by item block examinees in three proficiency-level/test length groups

Figures 4-1, 4-2 and Table 4.1 jointly present important baseline data for the model-based comparisons to follow. These same block-level NCLEX-RN® data were used to estimate the parameters in a hierarchical linear model (HLM) and in a latent growth model using structural equation modeling (SEM).

Trajectory Modeling Results for SEM-Based Growth Modeling

The 15-item block averages of the item-level response times were analyzed using a latent growth model developed under a structural equation model (SEM) framework. A separate growth model was developed for each of the three proficiency groups: (1) clear pass at 75 items (Pass-75); (2) clear fail at 75 items (Fail-75); and (3) indeterminate at 75 items (Indeterminate). An SEM-based growth model includes two latent factors that actually reflect the "growth" in some phenomena of interest. One factor represents the intercept or baseline for the growth model. The second factor represents the slope or rate of change. The slopes are allowed to change over time sequences—here, item blocks—denoting parametric changes in the response rates. The SEM-based latent growth modeling approach further incorporates time-specific measurement error into the model.

Using a typical implementation of SEM-based growth modeling, the change in the intercept is fixed at a constant value (usually 1.0). The change in response time is modeled via rate changes in the slopes. If the rate is constant, the growth is assumed to be linear in time. Nonlinear growth can be modeled by changing values of the slope coefficients (i.e., multipliers of the slopes). Estimation error variances are provided for both the intercept and slope coefficients. This allows hypotheses to be tested—especially concerning changes in the rate of change of the slopes.

An important feature of the SEM-based latent growth model is that model fit can be assessed. By incorporating latent means into the growth model, fit to the observed-variable means (item blocks) can be carried out. This provides an important way to empirically evaluate the estimated parametric growth model. The fit of the model to the

data, however, reflects the degree of statistical correspondence between the observed, sample-based covariance matrix and the model-based covariance model. Two statistics typically used to evaluate fit are a goodness-of-fit chi-square (χ^2) statistic, the adjusted goodness-of-fit (AGFI) statistic (Jöreskog and Sörbom, 1989), and Steigler's (1990) root mean squared error of approximation (RMSEA). A non-significant χ^2 statistic, AGFI values of 0.9 or higher, and a low RMSEA value (typically at 0.08 or smaller) support acceptable model fit (Raykov and Marcoulides, 2000). Of these statistics, the χ^2 is least interpretable because of a strong dependence on sample size and tendency to over-reject good-fitting models.

In this analysis, SEM models were developed individually for each of the three groups of interest: (1) Pass-75, (2) Fail-75; and (3) Indeterminate. The three group-specific SEM models all demonstrated negative slopes with the response latencies becoming shorter over the course of the examination.

Test takers classified as clear passers at 75 items have a 12.8 percent rate of response change in the slopes from BLOCK0 to BLOCK4. The SEM model equation for expected response times is:

$$E(X_1)=E(\tau_i)+\lambda_i E(\xi_j)+E(\delta_i) = \tau_i + \lambda_{ij}\,\kappa_j \qquad (3)$$

where κ_j represents the mean of ξ_j , j=1,..n observations and $E(\delta_i)=0$.

The estimated parameters, stated in model-based equation form are:

BLOCK0: $y_{t_0} = \lambda_1 \kappa_1 = (1.0)\ (69.06) = 69.06$

BLOCK1: $y_{t_1} = \lambda_1 \kappa_1 + \lambda_{2,2} \kappa_2 = (1.0)\ (69.06) + (-0.03)\ (-5.36) = 69.22$

BLOCK2: $y_{t_2} = \lambda_1 \kappa_1 + \lambda_{3,2} \kappa_2 = (1.0)\ (69.06) + (0.35)\ (-5.36) = 67.18$

BLOCK3: $y_{t_{27}} = \lambda_1 \kappa_1 + \lambda_{4,2} \kappa_2 = (1.0)\ (69.06) + (0.60)\ (-5.36) = 65.84$

BLOCK4: $y_{t_4} = \lambda_1 \kappa_1 + \lambda_{5,2} \kappa_2 = (1.0)\ (69.06) + (1.00)\ (-5.36) = 63.70.$

The parameter values and their standard errors are reported in Table 4.1. All of the rate-change parameter estimates are significant from zero at α=0.05, except for BLOCK1.

Table 4.2. Parameter estimates and standard errors of estimation for the latent growth model for clear passers at 75 items (Pass-75)

Parameter Description	Parameter	Estimate	Std. Error
Intercept	κ_1	69.06	0.74
Slope	κ_2	-5.36	0.52
Change in Slope for BLOCK1	$\lambda_{2,2}$	-0.03	0.06
Change in Slope for BLOCK2	$\lambda_{3,2}$	0.35	0.05
Change in Slope for BLOCK3	$\lambda_{4,2}$	0.60	0.05
Change in Slope for BLOCK4	$\lambda_{5,2}$	1.00	0.00

The model-fit statistics indicate an excellent fit of the model to the data for the Pass-75 group (χ^2=15.0, df=7, p<0.05; AGFI=0.99; RMSEA=0.04). Figure 4-3 shows a graphic of the model-based equations to aid in interpreting the growth curve (or decline in this instance of test-taker response times). Each line is modeled as a linear function. However, the end-points of the lines chart the response-time trajectory and may imply a non-linear rate of change—which is only slightly evident in Figure 4-3.

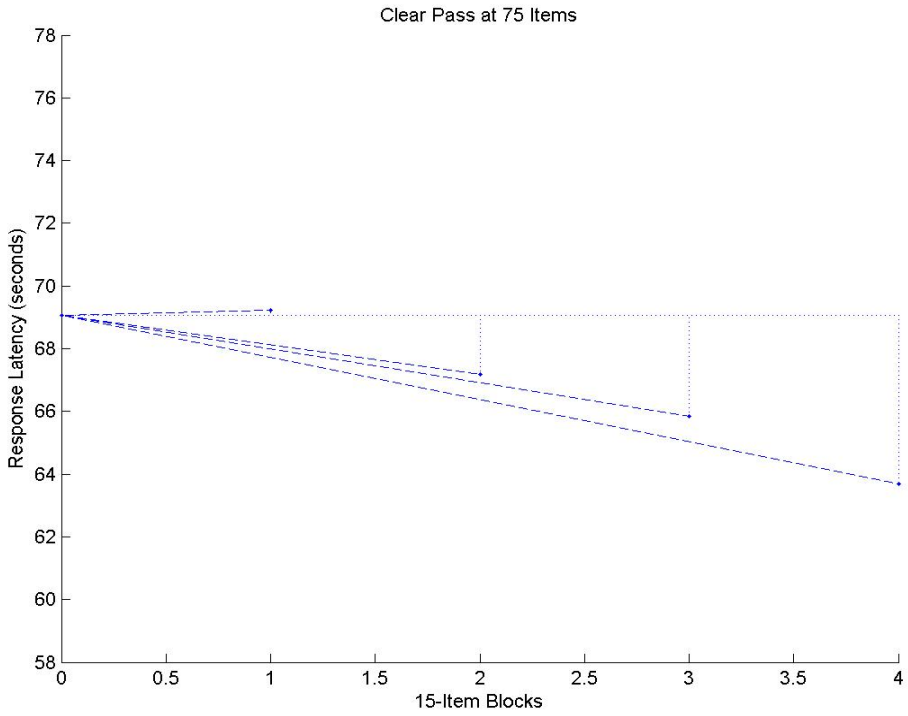

Figure 4-3. Model-based parameter estimates, model-fit statistics, and graphic of the model for the Pass-75 proficiency group.

The initial intercept is 69.06 and the initial slope is -5.36 (denoting a decline in response time). There is little change in BLOCK1. The slope, however, begins changing at a rate of almost 33 percent in BLOCK2 to BLOCK3 to BLOCK4. The coefficients for the rate of change (λ) indicate the rate of change is not constant. The rate of change can be tested by comparing the λ estimates (changes in slopes) for adjacent blocks using a non-central, single-degree-of-freedom *t*-test, where the test statistic is:

$$t = \frac{\hat{\lambda}_1 - \hat{\lambda}_2}{\hat{\sigma}_{pooled}}. \qquad (4)$$

The *t* statistic for the change from BLOCK1 to BLOCK2 is t=37.04 $_{(df=1005)}$, which is significant at α=0.05. For the change from BLOCK2 to BLOCK3, t = 26.8 $_{(df=1005)}$, which is also significant with 95 percent confidence. Refer to the graphic in Figure 4-3 to interpret these rate changes in the response times for adjacent blocks.

Examinees classified into the Fail-75 proficiency group have a 7.7 percent decrease in response latencies from BLOCK0 to BLOCK4. The SEM model equations for expected response times is:

$$E(X_1)=E(\tau_i)+\lambda_i E(\xi_j)+E(\delta_i) = \tau_i + \lambda_{ij}\,\kappa_j \qquad (5)$$

where κ_j represents the mean of ξ_j , j=1,..n observations and $E(\delta_i)$ =0.

The estimated parameters, stated in the model based equation form are:

BLOCK0 $y_{t_0} = \lambda_1 \kappa_1 = (1.0)\ (77.49) = 77.49$

BLOCK1 $y_{t_1} = \lambda_1 \kappa_1 + \lambda_{2,2} \kappa_2 = (1.0)\ (77.49) + (0.21)\ (-5.97) = 76.23$

BLOCK2 $y_{t_2} = \lambda_1 \kappa_1 + \lambda_{3,2} \kappa_2 = (1.0)\ (77.49) + (0.44)\ (-5.97) = 74.86$

BLOCK3 $y_{t_{27}} = \lambda_1 \kappa_1 + \lambda_{4,2} \kappa_2 = (1.0)\ (77.49) + (0.77)\ (-5.97) = 72.89$

BLOCK4 $y_{t_4} = \lambda_1 \kappa_1 + \lambda_{5,2} \kappa_2 = (1.0)\ (77.49) + (1.0)\ (-5.97) = 71.52$

The parameter values and their standard errors are reported in Table 4.2. All of the rate-change parameter estimates are significant from zero at α=0.05.

Table 4.3. Parameter estimates and standard errors of estimation for the latent growth model for Fail-75 proficiency group.

Parameter Description	Parameter	Estimate	Std. Error
Intercept	κ_1	77.49	1.01
Slope	κ_2	-5.97	0.69
Change in Slope for BLOCK1	$\lambda_{2,2}$	0.21	0.06
Change in Slope for BLOCK2	$\lambda_{3,2}$	0.44	0.05
Change in Slope for BLOCK3	$\lambda_{4,2}$	0.77	0.05
Change in Slope for BLOCK4	$\lambda_{5,2}$	1.00	0.00

The model fit statistics indicate an acceptable fit of the model to the data for the clear-fail-at-75-items (χ^2=10.81, df =7, p<0.15; AGFI=0.99; RMSEA=0.062). Figure 4-4 shows a graphic of the model-based equations interpreting the decline of test-taker response times. Each line is modeled as a linear function. As noted earlier, the end-points of the lines chart the response-time trajectory and may imply a non-linear rate of change. In Figure 4-4, however, the trajectory does not appear to (visually) depart from a linear rate change.

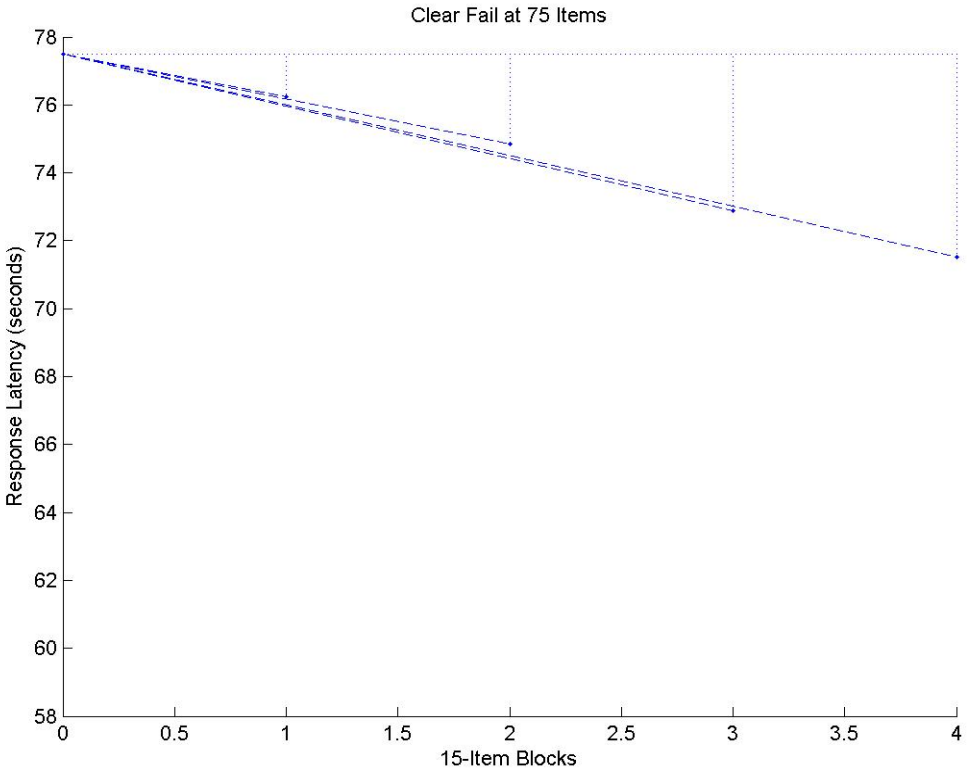

Figure 4-4. Model-based parameter estimates, model-fit statistics, and graphic of the model for the Fail-75 group.

The model fit is acceptably good, although slightly worse for the Fail-75 than the Pass-75 group. This Fail-75-items group has an intercept of 77.49 and a slope of -5.97. Those values indicate a considerably longer start-up pacing rate when compared to the Pass-75 group (see Figure 4-3). The coefficients for the rate of change (λ) indicate the rate of change is not constant. The *t* statistic for BLOCK1 to BLOCK2 is $t = 24.9_{(df=791)}$; for BLOCK2 to BLOCK3, $t = 39.4_{(df=791)}$. The graphic illustrates this change in response times.

Test takers for the Indeterminate group have a 6.96 percent decrease in response latencies from BLOCK0 to BLOCK4. Note that the model is only estimated for BLOCK0 to BLOCK4 to provide comparable results with respect to the other two groups, who only took 75 items. Results for a subset of the Indeterminate group who took all 265 items are reported further on. The SEM model equation for the expected response times is:

$$E(X_1)=E(\tau_i)+\lambda_i E(\xi_j)+E(\delta_i) = \tau_i + \lambda_{ij}\,\kappa_j \qquad (6)$$

where κ_j represents the mean of ξ_j , j=1,..n observations and $E(\delta_i)$ =0.

The estimated parameters, stated in model-based equation form are:

BLOCK0 $y_{t_0} = \lambda_1 \kappa_1 = (1.0)\ (71.39) = 71.39$

BLOCK1 $y_{t_1} = \lambda_1 \kappa_1 + \lambda_{2,2} \kappa_2 = (1.0)\ (71.39) + (-0.13)\ (-4.67) = 70.78$

BLOCK2 $y_{t_2} = \lambda_1 \kappa_1 + \lambda_{3,2} \kappa_2 = (1.0)\ (71.39) + (0.30)\ (-4.67) = 69.98$

BLOCK3 $y_{t_{27}} = \lambda_1 \kappa_1 + \lambda_{4,2} \kappa_2 = (1.0)\ (71.39) + (0.70)\ (-4.67) = 68.12$

BLOCK4 $y_{t_4} = \lambda_1 \kappa_1 + \lambda_{5,2} \kappa_2 = (1.0)\ (71.39) + (1.0)\ (-4.67) = 66.72$

The parameter values and their standard errors are reported in Table 4.3. All of the rate of change parameter estimates are significant from zero at α=0.05.

Table 4.4. Parameter estimates and standard errors of estimation for the latent growth model for "indeterminate-at-75-items" group (Indeterminate)

Parameter Description	Parameter	Estimate	Std. Error
Intercept	κ_1	71.39	0.50
Slope	κ_2	-4.67	0.33
Change in Slope for BLOCK1	$\lambda_{2,2}$	-0.13	0.05
Change in Slope for BLOCK2	$\lambda_{3,2}$	0.30	0.03
Change in Slope for BLOCK3	$\lambda_{4,2}$	0.70	0.03
Change in Slope for BLOCK4	$\lambda_{5,2}$	1.00	

The model-fit statistics indicate an acceptable fit of the model to the data for the Indeterminate group (χ^2=67.73, df =7, p<0.05; AGFI=0.98; RMSEA=0.062). Figure 4-5 shows a graphic of the model-based equations interpreting the growth curve of test-taker response times.

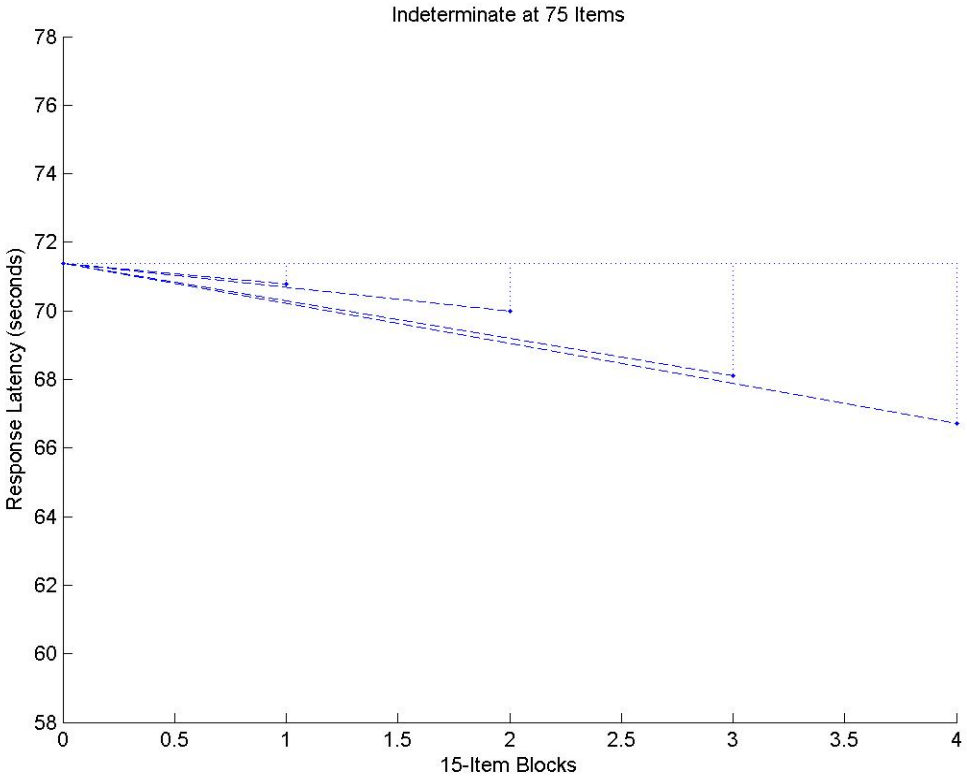

Figure 4-5. Model-based parameter estimates, model-fit statistics, and graphic of the model for the Indeterminate group.

The growth model for Indeterminate group provides estimates of the intercept of 71.39 and an initial slope of -4.67. Compared to the Pass-75 and Fail-75 groups, these values confirm that the initial average response time was between the times for the clear passers and clear failers—at least for the first four blocks of items that this group took. As noted above, the Indeterminate group has a 6.96 percent decrease in response latencies from BLOCK0 to BLOCK4. The coefficients for the rate of change (λ) indicate the rate of change is not constant. The *t* statistic for the change from BLOCK1

to BLOCK2 is $t = 20.0_{(df=2319)}$, which is significant at $\alpha=0.05$. The change from BLOCK2 to BLOCK3 $t = 62.0_{(df=2319)}$, is also significant with 95 percent confidence. The graphic in Figure 4-5 illustrates these rates of change in the response times for adjacent blocks.

The SEM equations for the Indeterminate group examinees across all blocks covering 76 items to 265 items could not be calculated due to an unavoidable correlation between the rates of missing data and response latencies for that group. This correlation was induced by the decision rule that ended the test at different test lengths. Instead, a growth model was fit to the first eight blocks for the Indeterminate group. That is, the first eight blocks had relatively complete data for the Indeterminate group.

The Indeterminate group demonstrated a 19.0 percent rate of response change in the slopes from BLOCK0 to BLOCK7. The SEM model equations for the expected response times is:

$$E(X_1)=E(\tau_i)+\lambda_i E(\xi_j)+E(\delta_i) = \tau_i + \lambda_{ij}\,\kappa_j \qquad (7)$$

where κ_j represents the mean of ξ_j , j=1,..n observations and $E(\delta_i) =0$.

The estimated parameters, stated in model-based equation form are:

BLOCK0 $y_{t_0} = \lambda_1 \kappa_1 = (1.0)\ (71.47) = 71.47$

BLOCK1 $y_{t_1} = \lambda_{11} \kappa_1 + \lambda_{22} \kappa_2 = (1.0)\ (71.47) + (-0.05)\ (-13.05) = 72.12$

BLOCK2 $y_{t_2} = \lambda_{11} \kappa_1 + \lambda_{32} \kappa_2 = (1.0)\ (71.47) + (0.07)\ (-13.05) = 70.55$

BLOCK3 $y_{t_{27}} = \lambda_{11} \kappa_1 + \lambda_{42} \kappa_2 = (1.0)\ (71.47) + (0.25)\ (-13.05) = 68.20$

BLOCK4 $y_{t_4} = \lambda_{11} \kappa_1 + \lambda_{52} \kappa_2 = (1.0)\ (71.47) + (0.40)\ (-13.05) = 66.25$

BLOCK5 $y_{t_5} = \lambda_{11} \kappa_1 + \lambda_{62} \kappa_2 = (1.0)\ (71.47) + (0.52)\ (-13.05) = 64.68$

BLOCK6 $y_{t_6} = \lambda_{11} \kappa_1 + \lambda_{72} \kappa_2 = (1.0)\ (71.47) + (0.82)\ (-13.05) = 60.76$

BLOCK7 $y_{t_7} = \lambda_{11} \kappa_1 + \lambda_{82} \kappa_2 = (1.0)\ (71.47) + (1.0)\ (-13.05) = 58.42$

Table 4.5. Parameter estimates and standard errors of estimation for the latent growth model for Indeterminate group BLOCK0 through BLOCK7.

Parameter Description	Parameter	Estimate	Std. Error
Intercept	κ_1	71.47	0.50
Slope	κ_2	-4.67	0.42
Change in Slope for BLOCK1	$\lambda_{2,2}$	-0.05	0.02
Change in Slope for BLOCK2	$\lambda_{3,2}$	0.07	0.02
Change in Slope for BLOCK3	$\lambda_{4,2}$	0.25	0.02
Change in Slope for BLOCK4	$\lambda_{5,2}$	0.40	0.02
Change in Slope for BLOCK5	$\lambda_{6,2}$	0.52	0.02
Change in Slope for BLOCK6	$\lambda_{7,2}$	0.82	0.02
Change in Slope for BLOCK7	$\lambda_{8,2}$	1.00	

The model-fit statistics indicate an acceptable fit of the model to the data for the Indeterminate group (χ^2=376.15, df =25, $p<0.05$; RMSEA=0.07862). Figure 4-6 shows a graphic of the model-based equations interpreting the growth curve of test-taker response times.

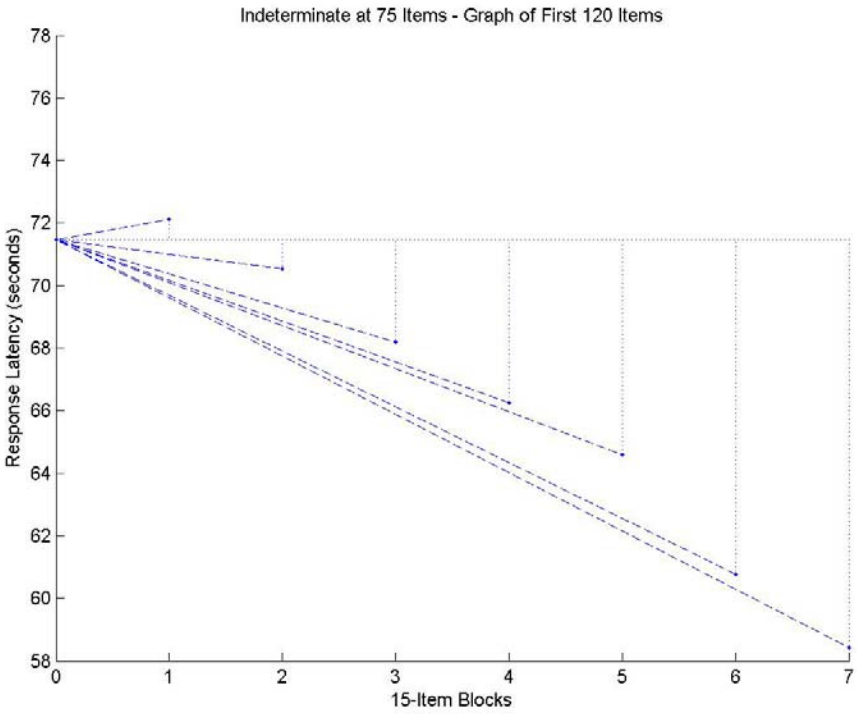

Figure 4-6. Model-based parameter estimates, model-fit statistics, and graphic of the model for the Indeterminate group on BLOCK0 to BLOCK7.

The model fit is acceptably good. The intercept is comparable to the intercept based on only BLOCK0 to BLOCK4 for this group of test takers, but the negative slope is substantially larger. There is an eighteen percent decrease in response latencies over the first seven blocks, versus a seven percent decrease on the first five blocks, as noted earlier. The coefficients for the rate of change (λ) indicate that the rate of change is not constant. Statistical *t*-tests conducted on the rate of change parameters for the first six items are shown in Table 4.5.

Table 4.6. Λ rate of change *t* statistic for the Indeterminate at 75 items (for the first 120 items taken)

BLOCK	*t* statistic	df
BLOCK1-BLOCK2	78.5	2319
BLOCK2-BLOCK3	117.5	2319
BLOCK3-BLOCK4	97.5	2319
BLOCK4-BLOCK5	78.5	2319
BLOCK5-BLOCK6	196.0	2319

A separate SEM growth model was fit to the data for a subset of 653 members of the Indeterminate group who completed all 265 items. This model provides a complete picture of the trajectory for those examinees that take the longest possible adaptive test (and also have some idea that they are "borderline" insofar as passing or failing). The results, including the model equations, fit statistics, and a graphic of the growth model, are presented in Figure 4-7.

The SEM model equations for the expected response times is:

$$E(X_1)=E(\tau_i)+\lambda_i E(\xi_j)+E(\delta_i) = \tau_i + \lambda_{ij}\,\kappa_j \qquad (8)$$

where κ_j represents the mean of ξ_j , j=1,..n observations and $E(\delta_i)$ =0.

BLOCK0 $y_{t_0} = \lambda_1 \kappa_1 = (1.0)\ (62.65) = 62.65$

BLOCK1 $y_{t_1} = \lambda_{11} \kappa_1 + \lambda_{2,2} \kappa_2 = (1.0)\ (62.65) + (-0.02)\ (-19.89) = 63.04$

BLOCK2 $y_{t_2} = \lambda_{11} \kappa_1 + \lambda_{3,2} \kappa_2 = (1.0)\ (62.65) + (0.07)\ (-19.89) = 61.25$

BLOCK3 $y_{t_{27}} = \lambda_{11} \kappa_1 + \lambda_{4,2} \kappa_2 = (1.0)\ (62.65) + (0.12)\ (-19.89) = 60.26$

BLOCK4 $y_{t_4} = \lambda_{11} \kappa_1 + \lambda_{5,2} \kappa_2 = (1.0)\ (62.65) + (0.20)\ (-19.89) = 58.67$

BLOCK5 $y_{t_5} = \lambda_{11} \kappa_1 + \lambda_{6,2} \kappa_2 = (1.0)\ (62.65) + (0.22)\ (-19.89) = 58.27$

BLOCK6 $y_{t_6} = \lambda_{11} \kappa_1 + \lambda_{7,2} \kappa_2 = (1.0)\ (62.65) + (0.44)\ (-19.89) = 53.89$

BLOCK7 $y_{t_7} = \lambda_{11} \kappa_1 + \lambda_{8,2} \kappa_2 = (1.0)\ (62.65) + (0.54)\ (-19.89) = 51.90$

BLOCK8 $y_{t_8} = \lambda_{11} \kappa_1 + \lambda_{9,2} \kappa_2 = (1.0)\ (62.65) + (0.57)\ (-19.89) = 51.35$

BLOCK9 $y_{t_9} = \lambda_{11}\kappa_1 + \lambda_{10,2}\kappa_2 = (1.0)\ (62.65) + (0.62)\ (-19.89) = 50.31$

BLOCK10 $y_{t_{10}} = \lambda_{11}\kappa_1 + \lambda_{11,2}\kappa_2 = (1.0)\ (62.65) + (0.75)\ (-19.89) = 47.73$

BLOCK11 $y_{t_{11}} = \lambda_{11}\kappa_1 + \lambda_{12,2}\kappa_2 = (1.0)\ (62.65) + (0.71)\ (-19.89) = 48.52$

BLOCK12 $y_{t_{12}} = \lambda_{11}\kappa_1 + \lambda_{13,2}\kappa_2 = (1.0)\ (62.65) + (0.74)\ (-19.89) = 47.93$

BLOCK13 $y_{t_{13}} = \lambda_{11}\kappa_1 + \lambda_{14,2}\kappa_2 = (1.0)\ (62.65) + (0.89)\ (-19.89) = 44.94$

BLOCK14 $y_{t_{14}} = \lambda_{11}\kappa_1 + \lambda_{15,2}\kappa_2 = (1.0)\ (62.65) + (0.91)\ (-19.89) = 44.55$

BLOCK15 $y_{t_{15}} = \lambda_{11}\kappa_1 + \lambda_{16,2}\kappa_2 = (1.0)\ (62.65) + (0.88)\ (-19.89) = 45.15$

BLOCK16 $y_{t_{16}} = \lambda_{11}\kappa_1 + \lambda_{17,2}\kappa_2 = (1.0)\ (62.65) + (0.92)\ (-19.89) = 44.35$

BLOCK17 $y_{t_{17}} = \lambda_{11}\kappa_1 + \lambda_{18,2}\kappa_2 = (1.0)\ (62.65) + (1.00)\ (-19.89) = 42.76$

The parameter values and their standard errors are reported in Table 4.6. All of the rate-change parameter estimates are significant from zero at α=0.05.

Table 4.7. Parameter estimates and standard errors of estimation for the latent growth model for indeterminate at 75 items BLOCK0 through BLOCK17.

Parameter Description	Parameter	Estimate	Std. Error
Intercept	κ_1	62.65	2.27
Slope	κ_2	-19.89	2.73
Change in Slope for BLOCK1	$\lambda_{2,2}$	-0.02	0.15
Change in Slope for BLOCK2	$\lambda_{3,2}$	0.07	0.14
Change in Slope for BLOCK3	$\lambda_{4,2}$	0.12	0.13
Change in Slope for BLOCK4	$\lambda_{5,2}$	0.20	0.13
Change in Slope for BLOCK5	$\lambda_{6,2}$	0.22	0.12
Change in Slope for BLOCK6	$\lambda_{7,2}$	0.44	0.09
Change in Slope for BLOCK7	$\lambda_{8,2}$	0.54	0.09
Change in Slope for BLOCK8	$\lambda_{9,2}$	0.57	0.09
Change in Slope for BLOCK9	$\lambda_{10,2}$	0.62	0.08
Change in Slope for BLOCK10	$\lambda_{11,2}$	0.75	0.08
Change in Slope for BLOCK11	$\lambda_{12,2}$	0.71	0.07
Change in Slope for BLOCK12	$\lambda_{13,2}$	0.74	0.07
Change in Slope for BLOCK13	$\lambda_{14,2}$	0.89	0.09
Change in Slope for BLOCK14	$\lambda_{15,2}$	0.91	0.09
Change in Slope for BLOCK15	$\lambda_{16,2}$	0.88	0.08
Change in Slope for BLOCK16	$\lambda_{17,2}$	0.92	0.09
Change in Slope for BLOCK17	$\lambda_{18,2}$	1.00	-

The model-fit statistics indicate an acceptable fit of the model to the data for this subset of the Indeterminate group on the full-length NCLEX (χ^2=10.81, df =7, p<0.15; AGFI=0.99; RMSEA=0.062). Figure 4-7 shows a graphic of the model-based equations interpreting the growth curve of test-taker response times.

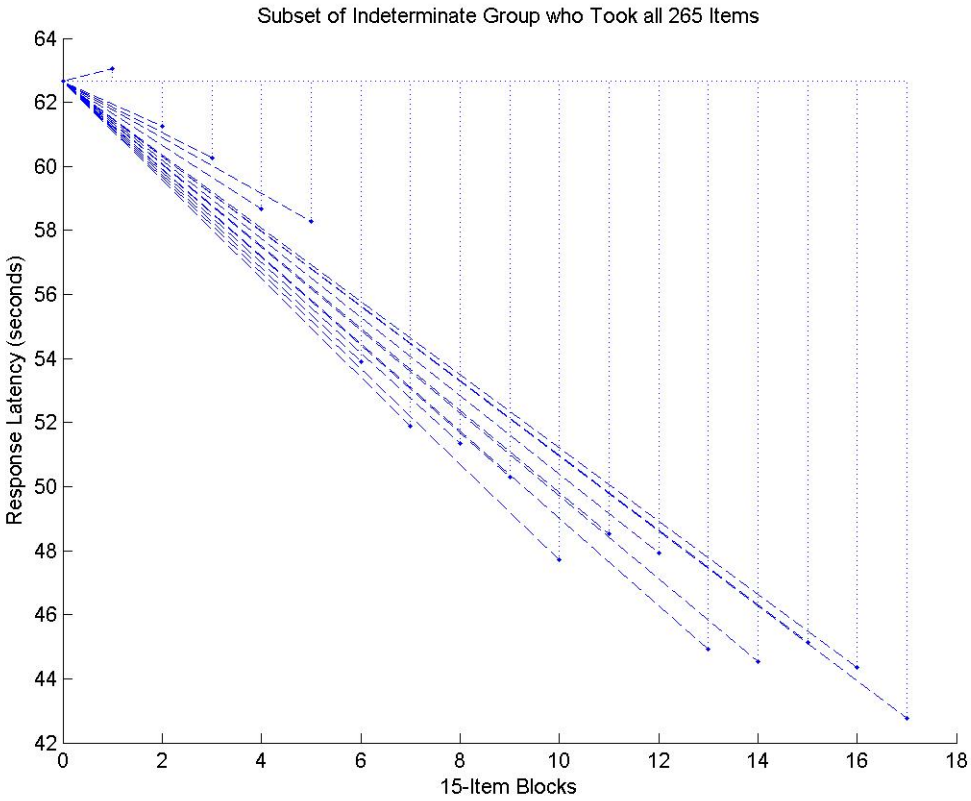

Figure 4-7. Trajectories for the subset of the Indeterminate group who took 265 items (maximum test length)

As indicated by Figure 4-7, this subset of the Indeterminate group that took all 265 items actually begins with shorter average response latencies than any of the other groups (intercept = 62.65). The initial slope is also significantly steeper (and negative) at -19.89. Two implications from these results are: (1) that lack of good pacing skills may be an issue for this Indeterminate group; and (2) that problematic pacing artifacts such as rapid guessing may occur fairly early in the test. Statistical *t*-tests conducted on the rate of change parameters for all the items are shown in Table 4.7.

Table 4.8. Rate of change ($\lambda_i - \lambda_{i-1}$): *t*-statistics for the Indeterminate group subset that took all 265 items

BLOCK	Rate of change t statistic	df
BLOCK1-BLOCK2	12.34	2319
BLOCK2-BLOCK3	7.33	2319
BLOCK3-BLOCK4	12.23	2319
BLOCK4-BLOCK5	3.20	2319
BLOCK5-BLOCK6	41.7	2319
BLOCK6-BLOCK7	22.1	2319
BLOCK7-BLOCK8	6.11	2319
BLOCK8-BLOCK9	12.23	2319
BLOCK9-BLOCK10	32.25	2319
BLOCK10-BLOCK11	-10.53	2319
BLOCK11-BLOCK12	8.42	2319
BLOCK12-BLOCK13	37.37	2319
BLOCK13-BLOCK14	7.87	2319
BLOCK14-BLOCK15	-7.05	2319
BLOCK15-BLOCK16	18.70	2319

Values greater than approximately 2.0 can be interpreted as being significant at α=0.05. Although a correct to the α-level for family-wise statistical error would raise that critical value, it would not change the interpretation in more than several instances. Overall, the *t*-test results in Table 4.7 imply that the pattern of declining response times for adjacent blocks is consistently significantly different than zero.

Trajectory Modeling Results for HLM

Hierarchical linear modeling (HLM) estimates three parameters in the two-level model: (1) the fixed effects, (2) the random effects level -1 coefficient, and (3) the variance-covariance components of the coefficients. The fixed effects are generalized

least squared estimates of the average group mean and average group slope. The random effects and the variance-covariance components are estimated using a maximum likelihood estimator.

The NCLEX-RN® data had to be dummy-coded using dichotomous variables, $x_{ij} \in \{0, 1\}$ as an indicator of group membership, in order to estimate the parameters for the two-level hierarchical linear model. Coding was done for the three proficiency groups of interest: (1) clear passers at 75 items (Pass-75) $x_{i1}=0$, $x_{i2}=0$; (2) clear failers at 75 items (Fail-75), $x_{i1}=0$, $x_{i2}=1$; and (3) indeterminate at 75 items (Indeterminate), $x_{i1}=1$, $x_{i2}=0$. Group is $x_{i1}=0$ if no information in item 76, $x_{i2}=1$ if information is present in item 76 or above and Group is weighted so PI2 is interpreted as the change in slope from the first set of item blocks.

The HLM analysis software, HLM2 (Raudenbush, Bryk, Cheong and Congdon, 2003) allows observations such as response times to be nested in persons. This nesting and the de facto unbalancing of the data due to the differential test lengths on an adaptive test are taken into account when estimating the variance and covariance components. An HLM model results in the following estimates. The level one (person/examinee) model is:

$$y_{ij} = \Pi_0 + \Pi_1 (\text{BLOCK}) + \Pi_2 (\text{GROUP}) + \varepsilon$$

where: Π_0= the examinee predictor is the group centered mean, the intercept

Π_1=the examinee pacing trajectory slope

Π_2=the examinee pacing trajectory slope after item 76

BLOCK=response latency by 15 item block

GROUP=examinees having information in Items 76 to 265

The level two (group) model is :

$\Pi_0 = \beta_{00} + \beta_{01}(X1) + \beta_{02}(X2) + r_0$

$\Pi_1 = \beta_{10} + \beta_{11}(X1) + \beta_{12}(X2) + r_1$

$\Pi_2 = \beta_{20} + r_2$

where:

β_{00}= the average of the Pass-75 intercepts

β_{01}=the average difference of Indeterminate intercepts when compared to Pass-75 intercepts

β_{02}= the incremental change from the first item block intercepts to the second item block intercepts

β_{10}=the average of the Pass-75 slopes

β_{11}=the average difference of the Indeterminate slopes when compared to Pass-75 slopes

β_{12}= the average difference of the Fail-75 slopes when compared to Pass-75 slopes

β_{20}= the incremental change in slope from the first item block to the second item block slope

X_1= dummy variable

X_2= dummy variable

For this application the HLM model is as follows:

Level-1 Model

Y = P0 + P1*(BLOCK) + P2*(GROUP) + E

Level-2 Model

P0 = B00 + B01*(X1) + B02*(X2) + R0

P1 = B10 + B11*(X1) + B12*(X2) + R1

P2 = B20 + R2

The full set of HLM results are presented in Table 4.8. The parameter estimates are included under the column labeled "Coefficient". Associated standard errors, *t*-statistics, and significance levels (i.e., *p*-values) are shown in the four right-most columns.

Table 4.9. Complete HLM Model Estimates

Fixed effect	Coefficient	Standard Error	Approx *t*-ratio	df	*p*(*t*)
INTRCPT1, P0					
INTRCPT2, B00	69.572532	0.783183	88.833	4032	0.000
X1, B01	2.394357	0.940278	2.546	4032	0.011
X2, B02	8.138534	1.185588	6.865	4032	0.000
BLOCK slope, P1					
INTRCPT2, B10	-1.318620	0.137999	-9.555	4032	0.000
X1, B11	-0.231800	0.162580	-1.426	4032	0.154
X2, B12	-0.216717	0.208904	-1.037	4032	0.300
GROUP slope, P2					
INTRCPT2, B20	-0.378287	0.098244	-3.850	4034	0.000

The condensed results, showing only the respective intercepts and slopes of the HLM models for each of the three proficiency groups, are shown in Table 4.9.

Table 4.10. Condensed HLM Intercept and Slope Estimates by Group

Group	Intercept	Slope
Pass-75	μ_0=69.572532	-1.318620
Indeterminate (at 75 items)	μ_1=71.966889	-1.550420
Indeterminate (> 75 items) (slope change at item 76)		-1.928707
Fail-75	μ_2=77.711066	-1.535337

The Table 4.9 results indicate that the Pass-75 group has the fastest initial pacing strategy and maintains that pace reasonably well (i.e., the intercept is 69.57 with a slope of -1.32). The intercept for the Indeterminate group indicates a somewhat slower initial pacing strategy than the Pass-75 group (intercept = 71.97) with corresponding greater rate of decline in their response times (slope = -1.55). That same group began responding even more rapidly after item 76, as indicated by the slope of -1.93. The Fail-75 group (Fail-75) had an intercept of 77.71 with a slope of -1.54, indicating an even slower initial pacing strategy than the Indeterminate group, with a rate of response-time decline approximately equal to the Indeterminate group on BLOCK0 to BLOCK4.

Where the SEM growth modeling provided a clearer picture of the (possibly nonlinear) changes in response trajectories across item blocks (but within proficiency groups) the HLM approach provides somewhat more direct between-group comparisons of the intercepts and slopes under a single model. These results may highlight an interpretive advantage of HLM over SEM growth modeling insofar as providing direct between-group comparisons of both the intercepts and slopes.

Table 4.10 provides the Level-1 (subject-level) standard deviations, variances, and associated significance tests. The variance components are subsequently reported as the diagonal elements of the asymptotic covariance matrix of the estimates (see Table 4.11).

Table 4.11. Level-1 standard deviations, variances, and significance tests for estimates of the intercepts and slopes

Random effect	Standard Deviation	Variance Component	df	Chi-square	P-Value
INTRCPT1, R0	23.81184	567.00387	1850	23141.71960	0.000
BLOCK slope, R1	3.35142	11.23199	1850	5196.24910	0.000
GROUP slope, R2	3.53436	12.49169	1852	5196.24910	0.000
Level-1, Error	8.84904	78.30550			

Table 4.11 shows the Level-1 covariance and correlation matrices for the parameter estimates (i.e., represent the Level-1 variance/covariance of the estimates of the intercepts and slopes). Variances are shown on the diagonal, covariances below the diagonal, and correlations above the diagonal. Strong correlations may indicate dependencies among the parameter estimates. For example, the correlation of -0.864 between the block and group facets in the model more-or-less highlights the confounding between proficiency group membership and variation in the response-time trajectories alluded to earlier. The correlation between the intercept and slope is (-0.576) indicating a

relationship between response latencies and change. In other words examinees with high response times dramatically increased their speed by the end of their examination.

Table 4.12 Level-1 variances, covariances and correlations

Parameters	Intercept	Block	Group
INTRCPT1, P0	567.004	-0.576	0.139
BLOCK, P1	-46.003	11.232	-0.864
GROUP, P2	11.684	-10.231	12.49169

Table 4.12 reports the reliability (stability) of the HLM intercept and slope estimates. The reliability coefficients quantify the stability of the parameter estimates. Reliability estimates for the variables in the model are calculated as a proportion. The proportion of total variance in parameter estimates that can be attributed to true parameter differences across subjects (Byrk, Raudenbush, & Cogdeon, 1992). The results indicate that the intercepts are highly reliable (.927). The slope estimates are less reliable (.669) for the first five item blocks and least reliable for items 76 to 265, taken in various numbers by the Indeterminate-group examinees (.486).

Table 4.13 Reliability of the HLM parameter estimates

Random level-1 coefficient	Reliability estimate
INTRCPT1, P0	0.927
BLOCK, P1	0.669
GROUP, P2	0.486

Finally, an intra-class correlation can be computed to compare the relative magnitude of these variance components by estimating the proportion of total variation attributable to between-subject variance. That intra-class correlation coefficient is .409, which implies that an estimated 41 percent of the total variation in pacing trajectories is attributable to differences in examinees.

The precision of the estimation of the intercept (the group means) depends on the sample size within each group. The precision of the estimation of the slope for each group depends on the sample size and the variability of response latency within the group (Raudenbush & Byrk, 2002). As noted above, the results indicate that the intercepts are quite reliable (.927), the slope estimates are less reliable (.669), and the slope change for the Indeterminate group after item 76 is the least reliable (.486). However, it is important to note that the estimates in this HLM model do not suggest a significant difference between the slopes of the three proficiency groups on the first five item blocks. There is a significant difference in the slope of Indeterminate at 76 to 265 items group from the Pass-75 group (t=-3.850, p<0.01), but that comparison is between the most able group and the mid-range Indeterminate proficiency group, after the latter group already completed 75 items. Substantively it is not clear, whether or not that comparison provides useful information.

Figure 4-8 provides a pictorial version of the HLM model. This figure demonstrates how the parametric HLM approach removes some of the fluctuation from the empirical graph of the means shown earlier in Figure 4-1. The capability of statistically testing the parameter estimates is also an advantage under HLM.

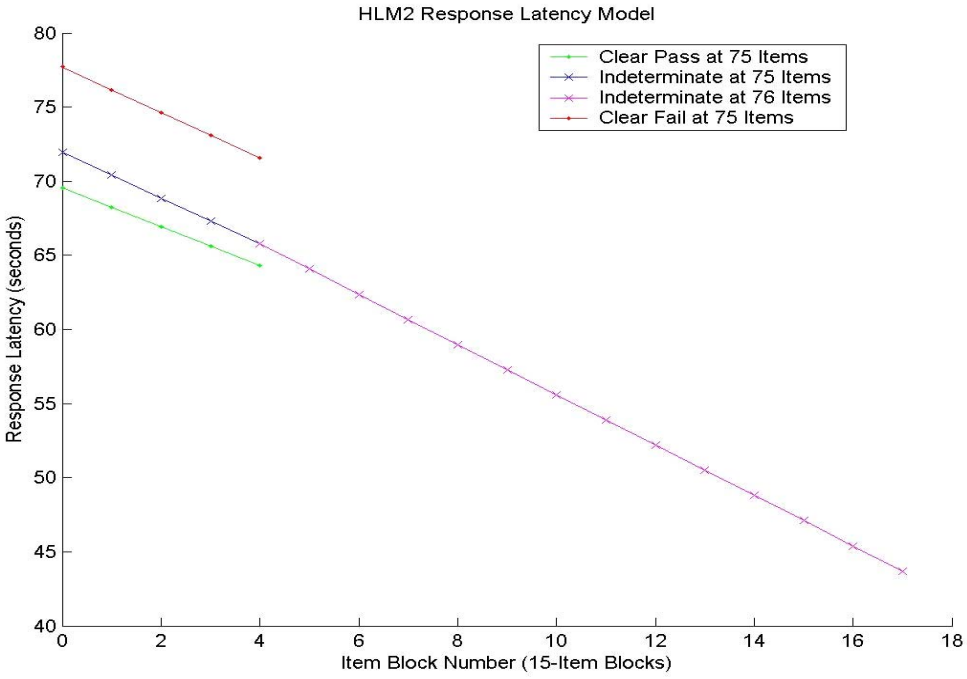

Figure 4-8. Graphic representation of the HLM response-time model.

Comparative Interpretations: SEM Growth Modeling Versus HLM

Both SEM-based growth modeling and HLM were used to devise models of likely pacing strategies for examinees taking a high stakes test. The intercepts of the two modeling techniques are similar, but that is because they are essentially based on the empirical means. The slopes are different in scale and relative magnitude. The slopes

differ because in SEM the latent factors have no inherent scale values. The scale of the slopes is set relative to a constraint of one placed on the final block. Although arbitrary, constraining the rate of change for the final item block allows the slopes at each block to be interpreted relative to the maximum amount of change for a particular proficiency group. On the downside, comparisons of the slopes across groups are not appropriate. In contrast, HLM works with the metric of the original variables. An advantage of HLM is that both slopes and intercepts can be compared between groups. However, the HLM model does not allow for the slope to change across item blocks, limiting its utility in detecting nonlinear growth trajectories. This becomes apparent by comparing Figure 4-8 (HLM model) to Figures 4-3 to 4-7.

Both approaches were also used to evaluate "performance trajectories", using percent-correct scores computed within item blocks. However, high correlations among the parameter estimates, very small reliabilities, and little apparent within person performance variation suggested serious estimation problems. In addition, the percent correct scores are somewhat confounded by differential item difficulty because of the CAT algorithm used for item selection. As a result, in the variance of the percent-correct scores tends to decrease over time, which further complicates the modeling process. Fortunately, as indicated by Figure 4-2, there is no substantial statistical evidence of any performance decline over the course of the NCLEX-RN®, despite the response latency trajectories demonstrated. That is, empirical trend plots of the percent-correct scores suggest that the percent-correct score trajectories do not change, even though almost all examinees tend to increase their pacing over the course of the examination.

CHAPTER V

CONCLUSIONS AND DISCUSSION

This study compares two methods of modeling response-time trajectories in a high-stakes computer-adaptive test (CAT): (1) hierarchical linear modeling (HLM) and (2) latent growth modeling using structural equation modeling (SEM). The study employed data from a sample of over 4100 examinees who took the NCLEX-RN® (National Council of State Boards of Nursing) during the 1999-2000 time period. The separate HLM- and SEM-based growth models were fit to data from the NCLEX-RN® examination to attempt to understand the response-time trajectories for three distinct groups at different proficiency levels: (1) examinees who clearly fail the NCLEX-RN® at 75 items (Fail-75); (2) examinees judged as indeterminate at 75 items and continuing to test up to 265 items (Indeterminate); and (3) examinees who clearly pass the NCLEX-RN® at 75 items (Pass-75).

The results clearly point to declining response-time trajectories for all three groups and further suggest that these proficiency-based groups differ from each other with respect to their average trajectories. The Pass-75 group demonstrated the fastest initial pacing strategies; however, those examinees appear to maintain their pace reasonably well. This result was evident from plots of mean response times across five blocks of items, as well as by both the SEM-based growth modeling and HLM. The Indeterminate group appears to adopt a slower initial pacing strategy than the Pass-75

group, as evidenced by the intercept terms in the trajectory models. That Indeterminate group appears to begin responding even more rapidly after item 76, however, there was no aggregate evidence suggesting that they had too little time. The Fail-75 group demonstrated the slowest initial pacing strategy—even slower than the Indeterminate group, but appeared to decline in their pacing approximately the same as the Indeterminate group on the first 75 items.

The use of two trajectory modeling approaches provided some useful advice for other researchers interested in this type of test-taking phenomena. The SEM-based growth modeling provided a clearer picture of the possibly nonlinear changes in response trajectories over the course of the test by allowing the changes in the slopes to be modeled. However, because of parametric constraints necessary to statistically identity the model, interpretation was hampered because the values of the parameter estimates were proportional to the constraints imposed. This was especially true for the between-proficiency-group comparisons. At best, only relative interpretations as to the magnitude of the intercepts and slopes were possible. In contrast, the HLM approach produces results that adhere to the scale of the original variables. The ensuing intercept and slope estimates are therefore directly interpretable as response times. HLM further provides somewhat more direct between-group comparisons of the intercepts and slopes under a single model. Considered jointly, these results suggest an interpretive advantage of HLM over SEM growth modeling insofar as providing direct between-group comparisons of both the intercepts and slopes.

The results suggest that both SEM-based growth modeling and HLM are useful methods for modeling response trajectories of various proficiency groups, with HLM probably preferable when looking specifically at between-group differences. Both modeling approaches indicated that there were statistically significant differences in the intercepts for the three proficiency groups and significant differences between the slopes for the Pass-75 and the Indeterminate groups (Hypothesis 1). Addressing both modeling approaches also demonstrated that the Pass-75 group maintained a somewhat consistent pacing trajectory over the course of 75 items. In contrast, the Indeterminate group never established a consistent pacing trajectory; rather, there was steady decline in the time taken per block of items, especially after the initial 75 items (Hypotheses 3 and 4).

Finally, the Fail-75 group established a pacing trajectory that was different from the other two groups, but maintained that trajectory across the length of the test (75 items). This finding addressed Hypothesis 5.

The finding of different pacing trajectories for the three proficiency-based groups is substantively intriguing, even though no evidence of test speededness was apparent. That is, both the SEM-based growth modeling and HLM approaches suggested that the intercepts (initial pacing) and the slopes (change in pacing) are different for examinees at different proficiency levels. Some researchers (e.g., Embretson and Prenovost, 2000) have theorized longer response times may indicate less effective processing. That is, response time (latency) represents a sum of processing speed, the number of processes attempted and other activities the examinee is engaged in while solving a complex task. Long response times represent processing difficulties and have been shown to be

positively associated with attentional failures. In any event, the capability to model the response trajectories seems to be an important first step toward better understanding the nature of the interaction between pacing skill and cognitive ability in particular domains. Ultimately, it may even be possible to provide pacing assistance mechanisms to facilitate examinees who lack proper pacing skills. However, while identifying pacing trajectories for a group of examinees is important, identifying an individual examinee's pacing trajectory may not ultimately impact their performance.

The overall conclusion is that this study found that all three examinee groups had decreased response latencies—that is, all examinees appeared to adopt an increased pacing strategy, even without any risk of running out of time. It may be that there is a natural "warm up" period for examinees. The test takers then consistently increase their pacing as they proceed through the test, regardless of proficiency level. It is interesting to note that in this study, as in a study by Bergstrom, Gershon and Lunz (1994), examinees did appear to spend more time on items they got incorrect than on items they got correct.

In light of this study and previous work by Bergstrom, Gershon and Lunz (1994), and Hadadi and Luecht (1998), a new definition of speededness may be in order. "Speededness" should be operationally redefined as a significant within-individual change in pacing trajectory, accompanied by a corresponding change in performance. This definition of speededness would allow testing organizations to employ modeling to identify individuals or groups of individuals whose performance declines in proportion to a change in their pacing trajectory (strategic or not). The testing algorithm might then,

within a confidence interval of the trajectories, be able to identify individuals at risk and recommend a “break “ or “time out” to contend with significant changes in pacing and performance.

Further research examining pacing trajectories in other large-scale high stakes tests would help confirm the usefulness of the pacing trajectory modeling techniques used here. This study did not examine the relationship between pacing and performance, *per se*. That is a topic for future research. The consideration of pacing as an auxiliary trait—intentionally measured or not—needs to be investigated. Hopefully, this study provides a well founded starting point for future research.

BIBLIOGRAPHY

Anastasi, A**. (**1976). *Psychological testing*. Macmillan: New York

Bejar, I. (1985). *Test speededness under number-right scoring: An analysis of the Test of English as a Foreign Language* (Report No. ETS-RR-85-11). Princeton, NJ: Educational Testing Service.

Bergstrom, B., Gershon, R. & Lunz, M. (1994) *Computerized adaptive testing exploring examinee response time using hierarchical linear modeling.* Paper presented at the annual meeting of National Council on Measurement in Education, New Orleans, LA.

Bontempo, B. (2000). *Assessing speededness using probabilistic models.* Unpublished doctoral dissertation, The University of Chicago, Chicago, IL.

Bontempo, B. & Julian, E. (1997). *Assessing speededness in variable-length computer adaptive testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Bridgeman, B. (2000). *Fairness in computer based testing: what we know and what we need to know.* (The GRE FAME Report). Princeton, NJ: Educational Testing Service.

Bridgeman, B. (2004, April). *Speededness as a threat to construct validity*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Bridgeman, B. & Cline, F. (2004). *Variations in mean response times for questions on the computer adaptive GRE general test implications for fair assessment.*(GRE No. 96-20P). Princeton, NJ: Educational Testing Service.

Bridgeman, B. & Cline, F. (2004). Effects of differentially time–consuming tests on computer adaptive test scores. *Journal of Educational Measurement,* 41 (2), 137-148.

Bridgeman, B., Cline, F. & Hessinger, J. (2003). *Effect of extra time on GRE Quantitative and verbal scores.* (GRE No. 00-03P). Princeton, NJ: Educational Testing Service.

Bridgeman, B., McBride, A & Monaghan, W. (2004). *Testing and time limits.* (R&D Connections). Princeton, NJ: Educational Testing Service.

Bryk, A., Raudenbush, S. & Congdon, R. (1988). *An introduction to Hierarchical Linear Models: computer program and user's guide* (2nd ed) Chicago:University of Chicago Department of Education.

Bryk, A. & Raudenbush, S. (1992). *Hierarchical Linear Models: applications and data analysis methods.* Newbury Park: Sage.

Bugbee, A. & Bernt, F. (1992). The effects of time constraints and mode of administration on test performance and perception. *Journal of Research on Computing in Education*, 25(2), 243-254.

Donlon, T.F. (1980). *An exploratory study of the implications of test speededness.* (ETS Research Report No. GREB-76-9P). Princeton, NJ: Educational Testing Service.

Donlon, T. (1973, November). *Establishing appropriate time limits for tests*. Paper presented at the meeting of the Northeast Education Research Association. Ellenville, NY.

Duncan, T., Duncan, S., Strycker, L., Li, F.& Alpert, A. (1999). *An Introduction to Latent Variable Growth Curve Modeling: concepts, issues and applications.* Mahwah, NJ: Earlbaum

Hadadi, A. & Luecht, R. M. (1998). Some methods for detecting and understanding test speededness on timed multiple-choice tests. *Academic Medicine*, 73(10), S47-S50.

Halkitis, P., Jones, P. & Pradhan, J. (1996). *Estimating testing time: the effects of item characteristics on response latency.* Paper presented at the annual meeting of the American Educational Research Association, New York, New York.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Henderson, W. (2004) *Speed as a variable on the LSAT and law school exams.* Law School Admission Council Research Report 03-03. Newtown: Law School Admission Council, Inc.

Henderson, W. (2004) The surprising and under theorized role of test taking speed. *Texas Law Review*, 82(4), 975-1052.

Hornke, L.(2000). Item response time in computerized adaptive testing. *Psicológica*, 21, 175-189.

Jones, M. G., Jones, B. D., Hardin, B., Chapman, L. Yarbrough, T. & Davis, M. (1999). The impact of high-stakes testing on teachers and students in North Carolina. *PHI DELTA KAPPAN*, 81(3), 199-203.

Jőreskog, K. & Sorbom, D., 2004, LISREL 8.7. Lincolnwood : IL, Scientific Software International.

Julian, E.R. & Bontempo, B.D. (1996, April). *Investigation into decision rules for NCLEX candidates who run out of time*. Paper presented at the annual meeting of the American Educational Research Association, New York, New York.

Linn, M. (2000). *Equity and knowledge integration*. (The GRE FAME Report). Princeton, NJ: Educational Testing Service.

Luecht, R. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement,* 20(4), 389-404.

Luecht, R., Hadadi, A., Swanson, D., & Case, S. (1998). A comparative study of a comprehensive basic sciences test using paper-and-pencil and computerized formats. *Academic Medicine* 73(10), S51-S53.

Luecht, R., & Nungester, R. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement,* 35 (3), 229-249.

Martinez, M. (1999). Cognition and the question of test item format. *Educational Psychologist*, 34(4), 207-218.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749.

Mislevy, R. (1996). Test theory reconceived. *Journal of Educational Measurement,* 33(4), 379-416.

Mislevy, R. & Bock, R. (1989). *PC-BILOG3:Item analysis and test scoring with binary logistic models.* Mooresville: Scientific Software.

NCLEX Examination using CAT (1997, October), 1-7.

NCLEX-RN® Examination test plan for the National Council Licensure Examination for Registered Nurses (2001, April), 3-10.

2005 NCLEX-RN® Examination Candidate Bulletin (n.d.), Retrieved August 2, 2005, from http://www.ncsbn.org/

Overton, R. & Harms, H. (1997). Adapting to adaptive testing. *Personnel Psychology*, 50(1), 171-186.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement,* 31(3), 200-219.

Plake, B. (1999). *A new breed of CATS: innovations in computerized adaptive testing.* Paper published by the University of Nebraska, Lincoln.

Raudenbush, S. & Bryk, A. (2002). *Hierarchical Linear Models: applications and data analysis methods.(2nd ed).* Newbury Park: Sage.

Raudenbush, S., Bryk, A., Cheong,Y. & Congdon, R. (2003). *HLM 5: Hierarchical Linear and Nonlinear Modeling,* version 5.45, Lincolnwood : IL, Scientific Software International.

Raykov, T. & Marcoulides, G. (2000). *A First Course In Structural Equation Modeling*. Malwah: Erlbaum.

Robert, R. & Stankov, L. (1999). Individual differences in speed of mental processing human and cognitive abilities: toward a taxonomic model. *Learning & Individual differences*, 11(1), 1-120.

Robin, F. (2002). *Investigating the relationship between test response behavior, measurement and person-fit.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Schnipke, D.L. & Scrams, D.J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, 34(3), 213-232.

Scrams, D.J., & Schnipke, D.L. (1997). *Making use of response times in standardized tests: are accuracy and speed measuring the same thing?* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.(ERIC document Reproduction Service No. ED409357)

Smith, R. (2000). *An exploratory analysis of item parameters and characteristics that influence item response time.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Stafford, R. (1971). The speededness quotient: a new descriptive statistic for tests. *Journal of Educational Measurement*, 8, 275-278.

Steiger, J. (1990). Structural model evaluation and modification: an interval estimation approach. *Multivariate Behavioral Research* 25(2), 173-180.

Stoolmiller, M. (1994) . Using Latent growth curve models to study developmental processes. In *The Analysis Of Change,* Gottman, J. M. (ed.), Mahwah: Earlbaum.

Stout, D. & Heck, J. (1998). Speed versus power testing: the impact of time constraints on introductory finance student test performance. *Financial Practice and Education* 7(1), 73-85.

Thissen, D. (1983). Time testing: an approach using item response theory. In D. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 179-203). New York: Academic Press.

van der Linden, W. (2005). *Linear Models for Optimal Test Design*. New York: Springer.

van der Linden, W., Scrams, D.& Schnipke, D. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement,* 23(3), 195-210.

van der Linden, W., Scrams, D., & Schnipke, D. (1998). *Using response time constraints in item selection to control for differential speededness in computerized adaptive testing.* Enschede (Netherlands). Research Report (ERIC Document Reproduction Service No. ED424266)

van der Linden, W. & van Krimpen-Stoop, E. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, 68(2), 251-265.

Wise, S. L., Kingsbury, G.G., Thomason, J. & Kong, X. (2004). *An investigation of motivation filtering in a statewide achievement testing program.* Paper presented at the annual meeting of the National Council of Measurement in Education, San Diego, CA.

Wise, S. L. & Kong, X. (2005). Response time effort: a new measure of examinee motivation in computer based tests. *Applied Measurement in Education*, 18(2), 163-183.

Yamamoto, K. (1995). *Estimating the effects of test length and test time on parameter estimation using the HYBRID model*. Technical Report (TR-95-2). Princeton, NJ: Educational Testing Service.

Yamamoto, K & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost & R. Langeheine (Eds) *Applications of latent trait and latent class models in the social sciences,* (pp. 89-98). Munster, Germany: Waxmann.