THE UNIVERSITY OF CHICAGO

ASSESSING SPEEDEDNESS USING PROBABILISTIC MODELS

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE SOCIAL SCIENCES

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF EDUCATION

BY

BRIAN D. BONTEMPO

CHICAGO, ILLINOIS

JUNE 2000

TABLE OF CONTENTS

LIST OF TABLES

(List of Tables continued)

LIST OF FIGURES

(List of Figures continued)

ABSTRACT

The study proposed a new method for the assessment of speededness. This new method was based on the utilization of probabilistic models to measure the speed and ability of examinees in timed and untimed conditions.

In order to measure speed, item duration was calibrated using a ten-point rating scale model. Item duration was fixed for both the timed and untimed condition whereby the measurement of examinee speed was conducted objectively in both settings.

In order to measure ability, item difficulty was calibrated using the Rasch dichotomous model. Item difficulty was anchored for both the timed and untimed condition whereby the measurement of examinee ability was conducted objectively in both settings.

This method was implemented to assess the speededness of the NCLEX-RN® examination. The method proved successful and found that the speed of the examinees was faster during the timed conditions than in the untimed conditions. The ability of the examinees was unexpectedly higher in the timed condition than in the untimed condition.

Additional inquiry found that examinees performed better when they increased their speed. If examinees increased speed too much, their performance dropped off.

CHAPTER 1

INTRODUCTION & LITERATURE REVIEW

When test developers create a power test (Gulliksen, 1950) and place a time limit on test administration, there is the potential for that time limit to effect the rate at which examinees answer items. When the time allowed for administration is not adequate, some examinees may rush through the examination. This kind of test is traditionally called a speeded test (Swineford, 1956). Until recently, speededness has been elusive to researchers due to the inaccessibility of item response times, which are now available through computerized testing. Using item response times, this research re-defines the notion of speededness, poses a method by which speededness can be investigated objectively, and utilizes that method to assess the speededness of the NCLEX-RN examination.

Gulliksen (1950) distinguished between examinations that measure only knowledge, called power tests, and those that also measure cognitive processing speed, called speed tests. In power tests, a test maker utilizes only the responses provided by examinees to measure the latent trait of interest. The effects of all other factors that contribute to measurement error, such as guessing, anxiety, motivation, and response

speed are treated as insignificant. By claiming to have a power test, a test maker is declaring that the effect of the time limit on examinee performance is insignificant. However, when the administration time for a power test is not adequate, then examinees may increase their response rate. This increase in response rate may detract from optimal performance. Tests that suffer from these conditions are called speeded (Swineford, 1956). Traditionally, speededness has been thought of as a dichotomy. A power test is either speeded or unspeeded.

In creating the notion of speededness, Swineford also established a way in which this classification could be determined. If all examinees (99%) reach 75% of the items and all of the items are reached by 80% of the examinees, then the test may be considered unspeeded (Swineford, 1956). There are many testing situations where it is impossible to determine how many items were answered by each examinee. All that is available is the number of examinees that completed the examination. In these situations, another criterion can be used. If 95% of the examinees complete the examination, then it may be deemed unspeeded. These criteria are norm-referenced and fail to probe the examination for any item level information such as examinee response rate.

The Swineford criteria also presume that by answering an item, an examinee has had ample time to perform optimally on the item. "When tests are number-right scored (i.e., no points detracted for incorrect responses), examinees are likely to rapidly guess on items rather than leave them blank" (Scnipke & Scrams 1997). Using item level

information in the form of item response times and correct/incorrect responses, Schnipke and Scrams have developed a method of detecting the items in which examinees engage in rapid-guessing behavior. Schnipke and Scrams recommend that test developers interested in speededness report the percentage of examinees that do not reach certain items and the percentage of examinees that engage in rapid-guessing behavior on those items. This assessment of speededness is valuable because it uses item level information in the form of Z scores to assess speededness.

Still, the notion of speededness and its assessment is to date incomplete. The notion of speededness as defined by Scnipke and Scrams is "the extent to which time limits affect examinees' test performance." This definition is good for it emphasizes the notion of speededness as a linear construct rather than a dichotomy as defined by the Swineford rule. Yet, it fails to tease out the different aspects of the speededness phenomenon. Speededness is a two-pronged phenomenon. Speededness is the extent to which an inadequate time limit affects examinee response rate and the affect of that increase in response rate on examinee performance. It is possible to have a time limit that causes examinees to hurry without detracting from their performance.

Setting the definition of speededness on the shelf, there is still plenty of room for improving the method by which test evaluators measure the speededness of an examination with a given time limit. At this point, no method of assessing speededness is capable of probing those examinees that work at an accelerated rate yet do not engage in rapid-guessing behavior. No method is applicable to adaptive testing. Lastly, no

method assesses the effect of the time limit on the rate of response or the performance of an individual examinee.

CHAPTER 2

METHODOLOGY

Assessing Speededness using Probabilistic Models

This section outlines a new method for assessing the speededness of an examination that is both old and new. The method employs traditional aspects of science by utilizing basic experimental design and common item equating procedures. The method employs relatively new techniques such as using item response times collected from a computer to create objective measures using probabilistic models. It is surprising that this method has not been used before.

The method assesses speededness as a linear construct, rather than as a dichotomy. The method's applicability is wide, for it is capable of assessing the speededness of both linear and adaptive tests. However, the method is only applicable to computerized situations where item response times are easily and accurately obtainable. The method uses item response times as well as candidate responses as data sources. It employs the Rasch dichotomous and rating scale model (Rasch, 1980; Andrich, 1979; Masters, 1982) in analyzing these data and provides a result that displays the extent to which the time limit affected an individual candidate's response

rate (or speed) as well as performance. Lastly, it aggregates this information to provide an overall estimate of the speededness of an examination.

The method is basic. Set up a repeated measures experiment where the examination administration time (control = unlimited time, treatment = actual administration time) is the independent variable and the two dependent variables are measures of examinee performance. Specifically, the dependent variables are examinee ability and speed. Nonetheless, there are some considerations that need to be taken when designing the experiment, collecting the data, and analyzing the results.

In order to make ability and speed comparable across treatment conditions, special considerations must be taken in developing the test forms. First, the test forms must be equatable. Second, since person performance is compared across treatment conditions, test form equating cannot be done via common person equating procedures. Rather, it must be accomplished via common item equating procedures. Lastly, no examinee can be administered the same item more than once. These conditions make it necessary to build at least 6 test forms. A possible form design is illustrated in Table 1.

TABLE 1
POSSIBLE FORM DESIGN FOR ASSESSMENT OF SPEEDEDNESS

	Control (Untimed)		**Treatment (Timed)**	
Examinee Group #1	Form 1		Form 2	
	Subform A	Subform B	Subform C	Subform D
Examinee Group #2	Form 3		Form 4	
	Subform A	Subform C	Subform B	Subform D
Examinee Group #3	Form 5		Form 6	
	Subform A	Subform D	Subform B	Subform C

After the forms have been designed, they should be administered to the examinees via computer so that accurate raw data can be collected. The raw data that need to be collected come in two forms, item results (correct/incorrect) and item response times.

Once the data have been collected, the next step is to create objective measures of each item's difficulty and duration. I should note that item duration is defined to be the objective measure of how long it takes examinees to answer the item. It is not a measure of the number of words in the item's stem. Using the item response times collected under controlled conditions, the rating scale model (Andrich, 1978; Masters, 1982) should be used for devising measures of each item's duration. Using the item results collected under controlled conditions, the Rasch dichotomous model (Rasch, 1980) should be used for devising measures of each item's difficulty. Other Rasch based techniques for improving the measurement system such as conducting misfit analyses should also be conducted.

Using the item's difficulty and duration, the next step is to create objective measures of examinee ability and speed under each experimental condition. Specifically, the difficulty and duration of each item should be anchored using the results from the above calibration. Then, the data collected under control (untimed) conditions should be used to calculate estimates of each examinee's untimed ability and untimed speed using both the Rasch dichotomous model and rating scale model. Lastly, the data collected under treatment (timed) conditions should be used to calculate estimates of each examinee's timed ability and speed. The same Rasch models and the same item difficulties and durations that were used in the untimed calculation should be used in this, the timed, calculation as well. Table 2 illustrates the raw data and variables.

TABLE 2.
EXPERIMENTAL DESIGN FOR THE ASSESSMENT OF SPEEDEDNESS

Independent Variable	**Control (Unlimited Time)**	**Treatment (Time Limit)**
Raw data collected from each examinee	1. Item response time for each of the untimed items 2. Result (right/wrong) for each of the untimed items	1. Item response time for each of the timed items 2. Result (right/wrong) for each of the timed items
Dependent Variable calculated for each examinee	1. Objective measure of untimed speed 2. Objective measure of untimed ability	1. Objective measure of timed speed 2. Objective measure of timed ability

Assessing the Speededness of the NCLEX-RN Examination

In order to illustrate this method, the speededness of the National Council Licensure Examination for Registered Nurses (NCLEX-RN® examination) was

investigated. Because this examination is a variable-length computerized adaptive test (a CAT with a varying number of items), the experimental design for the assessment of speededness was slightly more complicated than the basic design outlined previously.

Before getting into the details of the experimental design, some introduction to the examination is necessary. The National Council of State Boards of Nursing developed the NCLEX-RN® examination for the purpose of assessing and rendering pass/fail decisions on the competency of nursing licensure candidates to practice safe and effective entry-level nursing. In 1994, the NCLEX-RN® examination was converted to a variable-length computerized-adaptive testing modality. Approximately, 120,000 people take NCLEX-RN® examination each year.

The NCLEX-RN items are put through a great deal of scrutiny as they are written, pretested, and used in operation (Hubert and Gorham, 1998). All of the items are written using strict format guidelines that encompass phrasing, grammar, and appropriate usage of language. They are all four-response multiple-choice items that are reviewed by item writing editors, sensitivity review panels, and four different committees of nursing experts. During pretesting, items must achieve a sample size of greater than 400 for calibration. In addition, items maintain a point biserial correlation coefficient greater than 0.10 and each distractor must have at least a single response. Items must also maintain an absolute Z score of fit that is less than four. Items that have average response times greater than 2 minutes are also removed. Items are tested for DIF semi-annually. In addition, the items are reviewed regularly with special attention

paid to old items and items that have been highly exposed. In summary, the impact of non-content related effects have been minimized for every item. Psychometrically speaking, the items are of high caliber and quite similar.

The psychometric policies and procedures of the NCLEX-RN® examination are fairly typical of a variable-length adaptive test. The examination employs the Rasch model for both item difficulty calibration and examinee ability estimation. The difficulty of each operational item is calibrated using data collected on the item when it was administered as a pretest or experimental item. Each test is built using items from the operational item pool. The first item of each test is selected so that it has a difficulty near the cutscore. Upon completing the first item, an examinee's ability is estimated. Using the estimate of ability and the calibrated item difficulties, the next item is selected so that the examinee will have a 50% probability of getting the item correct. This process repeats itself for each item resulting in a test that is targeted to the examinee's ability. Examinees who perform well receive difficult questions, and examinees who perform poorly receive easy items. Once an examinee has answered the minimum number of operational items (60), the examination is stopped after either the cutscore is out of the bounds of the 95% confidence interval around the examinee's ability estimate or the examinee has reached the maximum number of operational items (250). If the final estimate of an examinee's ability is above the cutscore, (s)he passes. If not, (s)he fails. However, if an examinee fails to reach one of these two stopping points before the maximum administration time has expired (5 hours), a more stringent

pass/fail decision rule is applied. In this case, the last 60 estimates of an examinee's ability must be above the cutscore. The passing rate for these examinees is about 20% lower than the passing rate of similar examinees who do not run out of time. Even if the time limit has little affect on the rate and performance of examinees who run out of time, the pass/fail decision rule clearly disadvantages examinees who run out of time.

Based on the stopping rules of the examination, those examinees who were required to take the maximum number of items were the ones with ability estimates near the pass/fail cutscore, approximately ¼ of the examinees. In essence, because these examinees were close to the cutscore, their ability needed to be estimated to a greater degree of precision than those examinees who were either far above or far below the cutscore. However, by administering these examinees more items, the test developer may have put these examinees at risk of being effected by the time limit.

If the time limit was generous enough to allow all examinees to take their time and complete the examination regardless of how many items they were administered, then the number of items an examinee was administered would be of little concern. In the case of the NCLEX-RN, if every examinee were given only the minimum number of items, then the time limit for the examination would have been very generous and speededness would not have been a problem at all. Each year, less than 5 examinees fail to complete the minimum number of items.

However, if the time limit was generous enough to allow all examinees to take their time completing the minimum number of items, yet not adequate to allow

examinees to complete the maximum number of items, then differing test taking strategies may have ensued. This study aimed to detect those who employed one test taking strategy. This research detected those examinees who anticipated that they'd receive the minimum number of items and therefore initially took their time. That is, until they were administered more than the minimum number of items at which point they changed strategy and increased their speed. This is a reasonable strategy to expect since over half of the examinees completed their examination by taking only the minimum number of items.

Data

The data used in this study consisted of all first-time examinees taking the examination from April 1, 1998 to September 30, 1998. In total, 63,780 examinees took the NCLEX-RN from the operational item pool that was in the field during this period. Of these, the data from 2 of the original data files failed to import into the system. These files contained approximately 5,000 examinees testing between July 29th and August 4th and between September 9th and 16th. Otherwise, these data were similar to the 58,784 examinees that were successfully imported into the system. Conclusions derived from the sample of 58,788 should accurately represent the conclusions derived from the population of examinees taking the test during the time period.

This sample was divided into 3 sub-samples. The first sub-sample contained the 4,449 examinees who completed and failed the examination in the 5 hour time frame taking less than 120 items. The second sub-sample contained the 38,682 examinees

who completed and passed the examination in the 5 hour time frame taking less than 120 items. The third sub-sample was the group most likely to be effected by the limit. This sub-sample, the grey zone sub-sample, contained the 15,653 examinees who either completed the examination in more than 120 items or who failed to complete the examination within the five hour time limit. Some of these examinees passed and others failed the examination.

Method

The assessment of speededness for the NCLEX-RN examination was a ten-step process. The first five steps were conducted as part of the typical operational processes of the examination while other five involved in depth calculation. The ten steps were as follows:

1. Built equatable test forms
2. Administered the test under untimed conditions
3. Collected item response information during untimed conditions
4. Administered the test under timed conditions
5. Collected item response information during timed conditions
6. Created objective measures of item duration
7. Calculated two measures of each examinee's speed, untimed speed and timed speed
8. Created objective measures of item difficulty

9. Calculated two measures of each examinee's ability, untimed ability and timed ability
10. Calculated examinee change scores

The first step was to build equatable test forms. Since, the NCLEX-RN® examination is based on a Rasch calibrated item bank, each examinee received a unique test form that was linkable to any other test form that was created with items from the calibrated item bank. Additionally, any part of an examination was linkable to another part of the examination although the standard error of measurement varied greatly depending on the number of items in each part and the 'targeting' of those items to the ability of the examinee. Further information on the equatability of tests taken in a computerized adaptive testing environment can be found in (Wright & Stone, 1979).

The second step was to administer the test under untimed (control) conditions. Since many first-time examinees anticipated that they'd receive the minimum number of items, they may have worked at a relaxed pace until they realized that they had surpassed the minimum number of items. In essence, until an examinee had completed more than the minimum number of items, (s)he perceived that they were being administered the test under very generous time limits. For the purposes of this study, the first part of each examination (the items taken up to the minimum number of items) was administered under conditions that were essentially untimed.

The third step was to collect item response information during untimed (control) conditions. For every examinee in the sample, the item response time and the dichotomous correct or incorrect result were collected for all of the items up to the minimum number of items.

As the fourth step, the test was administered under timed (treatment) conditions. For all examinees taking more than the minimum number of items, time limit considerations changed shortly after the examinees realized that they had taken more than the minimum number of items. Once an examinee realized that he/she has taken more than the minimum number of items, he/she may have changed response rate in an effort to complete the test in the remaining time. For the purposes of this study, the second part of the examination (all items taken after the minimum number of items) was administered under timed conditions.

The fifth step was to collect item response information during the timed (treatment) condition. For every examinee taking more than 120 items, the grey zone sub-sample, the item response time and the dichotomous correct or incorrect response was collected for all of the items after the minimum number of items. This was the entire sample of examinees providing enough timed information to calculate reasonable estimates of timed ability and timed speed.

The sixth step was more a process than a step. The goal was to create objective measures of item duration. This process was broken into several parts.

a) Developed a universal speed rating scale

i) Calculated the natural log of response time for all item response times – Since the distribution of item response time was skewed, it was necessary to normalize the data by taking the natural logarithm of the data. This process is common in item response time research (See Schnipke, 1999).

ii) Calculated the speed rating scale score for each item – Previous research (Bontempo, 1997) has shown that the NCLEX-RN examination data fit a ten point rating scale model. This same model was used to convert the natural log of item response times into speed scores (0-9). Table 3 displays the conversion of raw response time data into units on the speed rating scale. The relationship between item response time and speed rating scale was inverse. Fast responses received high scores.

TABLE 3.
RESPONSE TIME DATA CONVERSION TABLE

Time (Seconds)		Ln(Time)		Speed
Lower	Upper	Lower	Upper	Rating Scale
0.0	7.4	0.0	2.0	9
7.4	12.2	2.0	2.5	8
12.2	20.1	2.5	3.0	7
20.1	33.1	3.0	3.5	6
33.1	54.6	3.5	4.0	5
54.6	90.0	4.0	4.5	4
90.0	148.4	4.5	5.0	3
148.4	244.7	5.0	5.5	2
244.7	403.4	5.5	6.0	1
403.4	α	6.0	α	0

iii) Derived initial rating scale step calibrations – Speed rating scale step calibrations were calculated for each sub-sample using the MESA software Winsteps version 2.98 (Linacre, 1999). The data from the grey zone sub-sample were split into two parts, the early part of the examination and the late part of the examination. In total, four different estimates of the steps' duration were calibrated.

iv) Assessed the rating scale – In addition to assessing the fit and order of the step calibrations, the stability of the rating scale steps across the six estimates was assessed.

v) Decided on a set of bench step calibrations – After assessing the rating scale, a step anchor file was constructed that fixed the step calibrations to a constant value for all future calibrations.

b) Developed measures of item duration

i) Derived initial estimates of item duration – Using the step anchor file and the same sets of data from the step calibrations, four estimates of duration were derived for each item in the pool.

ii) Assessed the item duration estimates – Using Rasch statistics, the fit of each item to the model was assessed. The stability of the item duration estimates across the six estimates was also investigated.

iii) Assigned a bench duration estimate to each item – An item anchor file was built that fixed each item's duration at a constant value for all future calibrations.

The seventh step was to calculate two measures of each examinee's speed, untimed speed and timed speed. Using the anchored item duration estimates, an untimed speed estimate was calculated for each examinee in the grey zone sub-sample using the data collected during the early part of the examination. Using the anchored item duration estimates, a timed speed estimate was calculated for each examinee in the grey zone sub-sample using the data collected during the late part of the examination.

The eighth step was to create objective measures of item difficulty. Since the NCLEX-RN examination maintains a high level of psychometric scrutiny, there was no reason to suspect that the difficulty of the items would vary from sample to sample. However, there was reason to suspect that the difficulty of the items might change under speeded conditions. For each item response from the grey zone examinees, the dichotomous correct/incorrect score was used to calculate two different difficulty estimates of the 1803 items in the pool, one using the early data and one using the late data. The stability of the estimates of item difficulty across these two datasets was assessed. A bench difficulty was assigned to every item and written into an item anchor file.

The ninth step was to calculate two measures of each examinee's ability, untimed ability and timed ability. Using the anchored item difficulty estimates, an untimed ability estimate was calculated for each examinee in the grey zone sub-sample using the data collected during the early part of the examination. Using the anchored item difficulty estimates, a timed ability estimate was calculated for each examinee in

the grey zone sub-sample using the data collected during the late part of the examination. Since person fit statistics are of little use in a CAT environment (Stone, 1994), the fit statistics were not investigated.

The final step was to calculate examinee change scores. The change in ability and speed were calculated for each examinee. See Table 4 for the derivation of change scores. This change infers the change in ability and speed due to the treatment or the impact of the time limit on an individual examinee's speed and performance. The change scores were also aggregated together as the distribution of change scores. These distributions displayed the overall impact of the time limit on speed and performance, or the overall speededness of the examination.

TABLE 4.
EXAMINEE CHANGE SCORES DERIVATION

Early part of the exam (Untimed or Control Conditions)	**Late part of the exam** (Timed or Treatment Conditions)	**Change Scores**
Rasch Measure of Speed	Rasch Measure of Speed	Difference in Rasch Measures of Speed
Rasch Measure of Ability	Rasch Measure of Ability	Difference in Rasch Measures of Ability

Additional Inquiry

In addition, the following hypotheses were tested:

Hypothesis A. Examinee speed is faster during the late part of the examination than during the early part of the examination.

Hypothesis B. Change in speed is inversely related to early speed.

Hypothesis C. Examinees perform better when they aren't working under tight time constraints. (Average examinee performance is better during the early part of the examination than during the late part of the examination.)

Hypothesis D. Change in examinee speed is inversely related to change in performance. (For the group of examinees that work faster during the late part of the examination than the early part of the examination, change in speed is inversely related to change in ability.)

Hypothesis E. Examinee speed is constant within each part of the examination.

Hypotheses A and C were be tested by determining if the change score distributions were different than zero. Hypotheses B and D were tested by plotting the two variables and calculating the correlation between them. This last hypothesis was investigated using regression techniques. Two regression functions were derived for each examinee, one for each part. Each function was created by regressing the item response times onto the numeric sequence of administration. If the slope of the function neared zero then the examinee's speed for that part of the examination was approximately constant. The slopes of all of the examinees' speed functions for each part were aggregated together to determine if the overall examinee speed was constant within each part of the examination.

In addition, the characteristics of speededness were investigated. Specifically, the characteristics of items and examinees that changed from the untimed to the timed

part of the examination were documented. The gender, ethnicity, and language skills of examinees changing in speed and ability were investigated. And, the content area of misfitting or unstable items was investigated.

Finally, the behavior of rushed examinees was investigated further. The individual item responses were probed to see if there was a relationship between item response time and result.

CHAPTER 3

ANALYSES AND RESULTS

Assessing the Speededness of the NCLEX-RN® Examination

This section documents the results of the various calibrations that were conducted in steps six to ten of the methodology used to assess the speededness of the NCLEX-RN® examination. This section also contains the results of the additional inquiry.

The Speed Rating Scale

Once the raw data had been converted into speed rating scale scores, the rating scale was calibrated using the grey zone sub-sample data from the early part of the examination. The rating scale statistics for the early grey sub-sample are provided in Table 5. Table 5 shows that the calibrations were ordered. The steps all fit quite well except for the two end steps (0 and 9) which still displayed an adequate fit. These were responses where the examinee either rapidly guessed or took an extraordinary length of time. Both of these types of responses were unexpected and should not fit the model

well. It was encouraging to note that of the 939,166 responses analyzed in this calibration, only 595 were these types of responses.

TABLE 5.
SUMMARY OF RATING SCALE STATISTICS FOR THE EARLY GREY SUB-SAMPLE

CATEGORY LABEL	OBSERVED COUNT	MEASURE AVERGE	MEASURE EXP.	COHERENCE M->C	COHERENCE C->M	INFIT MNSQ	OUTFIT MNSQ	STEP CALIBRATN	
0	453	-3.036	-3.59	58%	3%	1.62	1.67	NONE	time>403
1	2958	-2.147	-2.26	38%	8%	1.11	1.12	-4.77A	time>245
2	25062	-1.415	-1.45	41%	5%	1.05	1.05	-3.96A	time>148
3	127557	-0.869	-0.87	43%	21%	1.02	1.02	-2.78A	time>090
4	306982	-0.337	-0.32	45%	57%	0.99	0.99	-1.48A	time>055
5	298357	0.204	0.211	43%	57%	0.99	0.99	-0.03A	time>033
6	142803	0.762	0.758	42%	24%	0.98	0.98	1.22A	time>020
7	31999	1.362	1.323	42%	4%	0.95	0.95	2.53A	time>012
8	2853	2.058	1.917	36%	0%	0.90	0.90	4.03A	time>007
9	142	2.174	NONE	0%	0%	1.33	1.31	5.25A	time<007

In order to certify that the step calibrations remained stable across different samples as well as under different conditions of speededness, three other calibrations were conducted using data from the three other sub-samples. By comparing the rating scale's stability across the clear failers and the clear passers, the stability of the scale across different samples was certified. In order to test the scale's stability across conditions of speededness, the step calibrations from the early and late part of the examination for the grey zone sub-sample were compared. A plot of the step estimates from each of these calibrations seen in Figure 1. It is evident by this plot that the rating scale worked well. The end steps displayed some variation. However, this variation was expected due to the unexpected nature of these responses. The step calibrations

from the early grey sample were sufficient to use for this study. These step calibrations were anchored for all future calibrations and analyses.

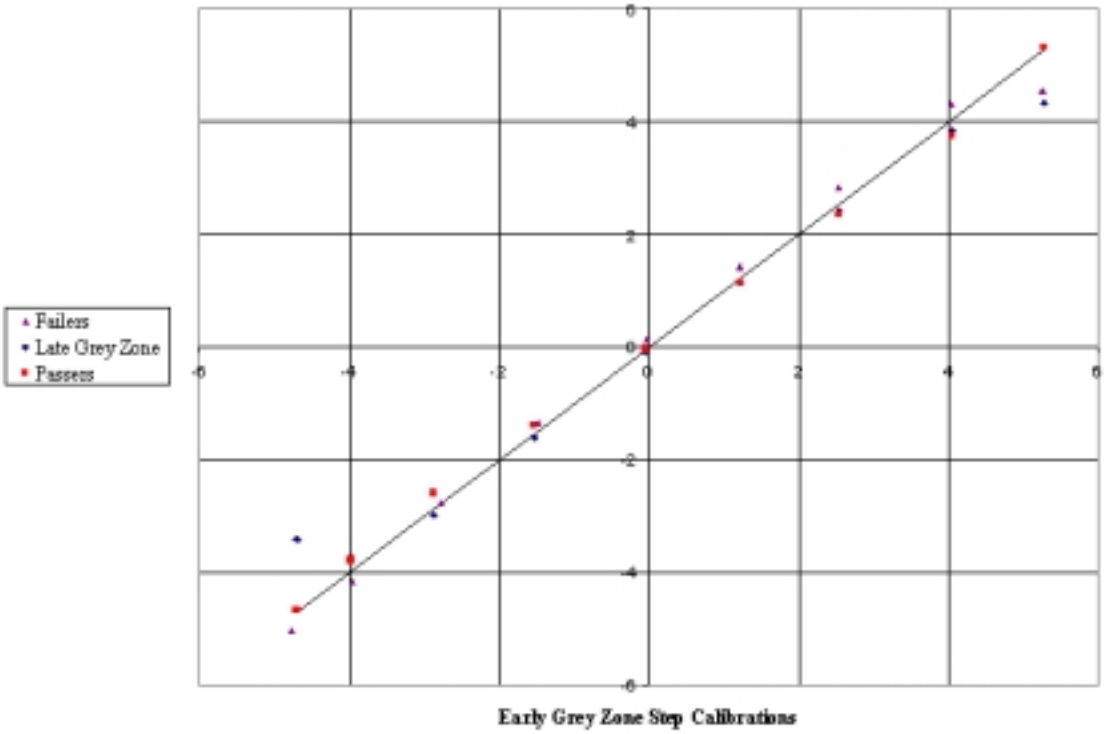

Figure 1. Step Calibrations for Failers, Late, and Passers plotted against Early Step Calibrations

The Item Duration Estimates

Having developed a solid rating scale, the item duration estimates could be calibrated and tested. This was done first for the early grey sub-sample. Of the 1803 items that were available, 9 of the items were not taken by a single examinee. The summary statistics for the items are displayed in Table 6. Overall, the items separated well (7.20 to 7.47) and fit the model (Mean Infit & Outfit Mean Square=1.00).

TABLE 6.
SUMMARY OF ITEM DURATION ESTIMATES FROM THE EARLY GREY SUB-SAMPLE

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	2374.8	523.5	0.00	0.07	1.00	-0.2	1.00	-0.2
S.D.	2056.5	459.6	0.68	0.05	0.24	3.1	0.24	3.1
MAX.	12195.0	2666.0	4.14	0.80	2.46	9.9	2.46	9.9
MIN.	6.0	2.0	-2.84	0.02	0.01	-9.9	0.01	-9.9

REAL RMSE	0.09	ADJ.SD	0.67	SEPARATION	7.20	ITEM	RELIABILITY	0.98
MODEL RMSE	0.09	ADJ.SD	0.67	SEPARATION	7.47	ITEM	RELIABILITY	0.98
S.E. OF ITEM MEAN	0.02							

In order to test the stability of these item duration estimates, the calibrations were conducted on the three other sub-samples. The estimates of item duration from the three sub-samples were plotted against the item calibrations from the early grey sample. Items that had a sample size of less than 50 were excluded from the plot. The item duration estimates of all items with sample size greater than 50 are shown in Figure 2. Figure 2 reveals that the item duration remained stable across samples and

across conditions of speededness. The mean difference in duration between the item difficulty estimates from the early and late grey sample was -0.07 and the standard deviation was 0.12. Really, none of the items of items changed in duration.

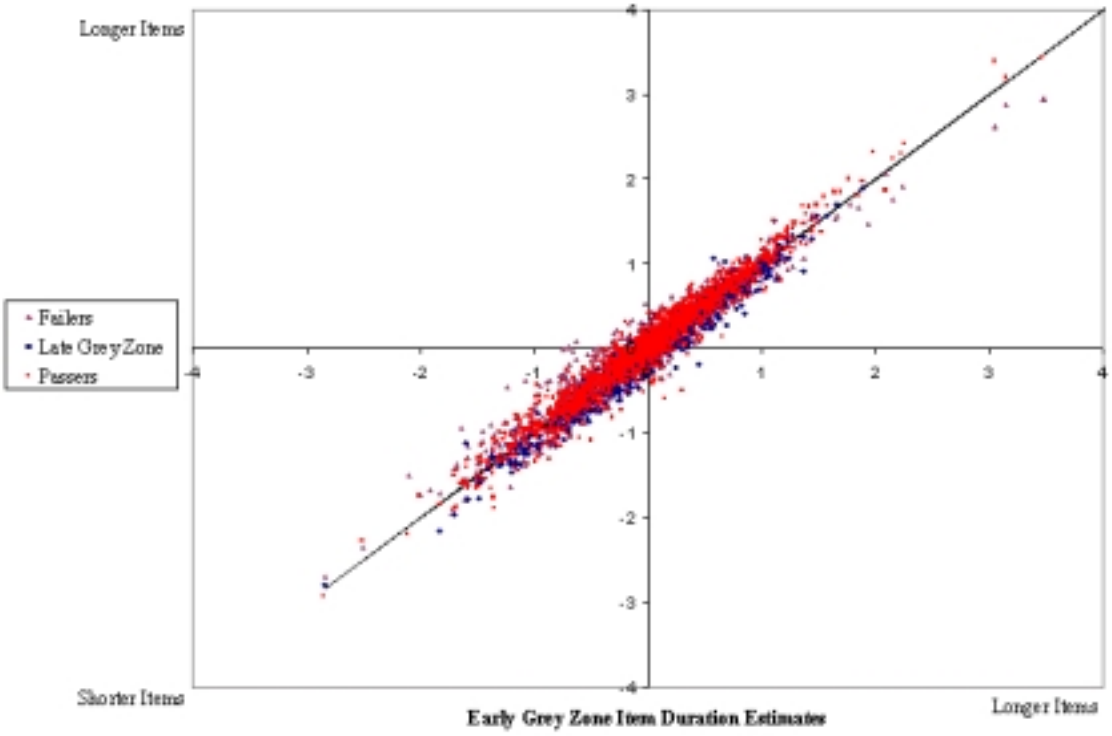

Figure 2. Item Duration for Failers, Late, and Passers plotted against Early Item Duration

In summary, both the speed rating scale and item duration estimates from the early grey sub-sample were stable and functioning quite well. Having determined this, both the step calibrations and item duration estimates were anchored using the data from this sample. By anchoring these calibrations, examinee speed could be measured objectively in both the timed and untimed parts of the examination.

Examinee Speed

The speed of each of the 15,653 examinees in the grey zone sub-sample were calculated twice, once using the data from the early part of the examination and once using the data from the late part of the examination. The summary statistics for the speed of the grey zone examinees in the early part of the examination are displayed in Table 7. The mean speed was 0.04 logits with a standard deviation of 0.73 logits. Overall, the examinees fit the model although some displayed a degree of misfit. These were examinees that were behaving inconsistently. Some items were answered slower than expected and others were answered faster than expected. Overall, these examinees tended to be the slower examinees. The mean speed of the 185 examinees with infit mean square greater than 2.0 was –0.59 logits, much lower than the speed of the examinees who fit the model. The examinees were easily differentiated as the separation exceeded four.

TABLE 7.
SUMMARY OF EARLY SPEED

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	272.2	60.0	0.04	0.15	1.00	-0.2	1.00	-0.2
S.D.	33.1	0.1	0.73	0.00	0.34	1.8	0.34	1.8
MAX.	395.0	60.0	2.90	0.16	4.09	9.7	4.08	9.7
MIN.	92.0	46.0	-3.76	0.14	0.24	-6.2	0.24	-6.2
REAL RMSE	0.16	ADJ.SD	0.72	SEPARATION	4.52	PERSON RELIABILITY		0.95
MODEL RMSE	0.15	ADJ.SD	0.72	SEPARATION	4.82	PERSON RELIABILITY		0.96
S.E. OF PERSON MEAN	0.01							

The speed of an examinee in logits was useful for conducting research but lacked some conceptual grounding without reference to common units of time. Using the data, a table was constructed that allowed for the conversion of logit speed into average response time per item. This average item response time was more useful than the average item response time per item using the raw data. This was because the raw data calculation did not account for the difference in the duration of the items that each examinee received. The conversion table was summarized in graphical form in Figure 3. The mean item response time in the early part of the examination was 52 seconds per item. This graph is log-linear. The linear representation of this graph is shown in Figure 4, which displays the ln of average item response time plotted against the Rasch measure of speed. The ln of average item response time was 4.6.

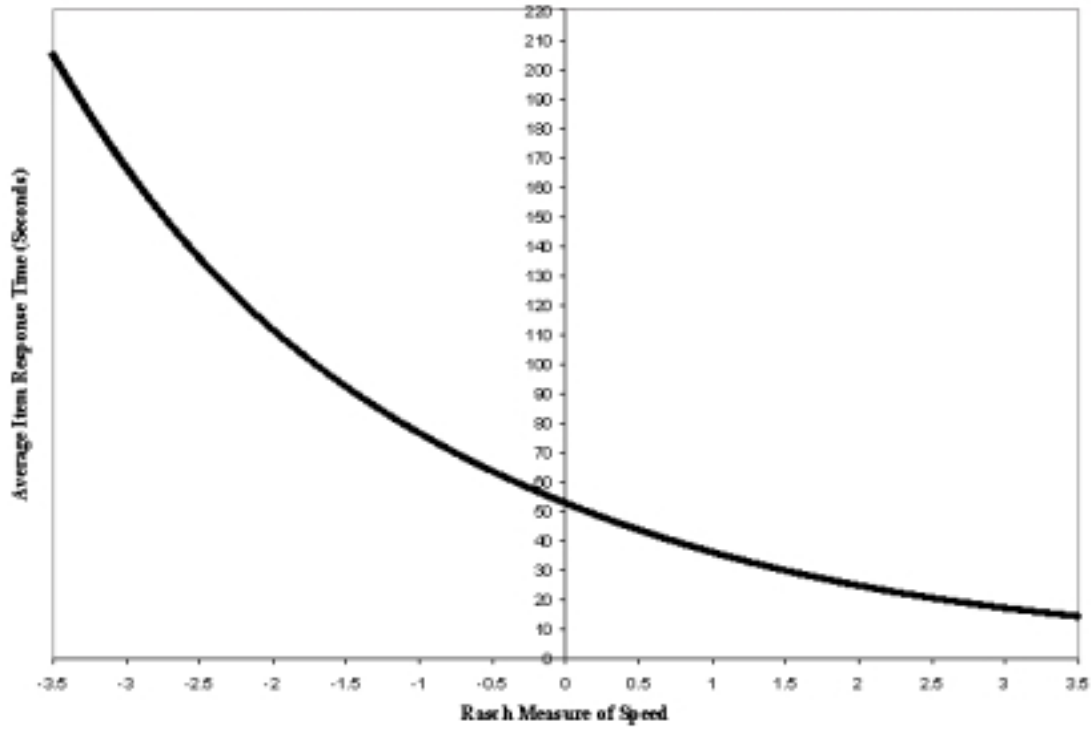

Figure 3. Average Item Response Time in Seconds by Rasch Measure of Speed

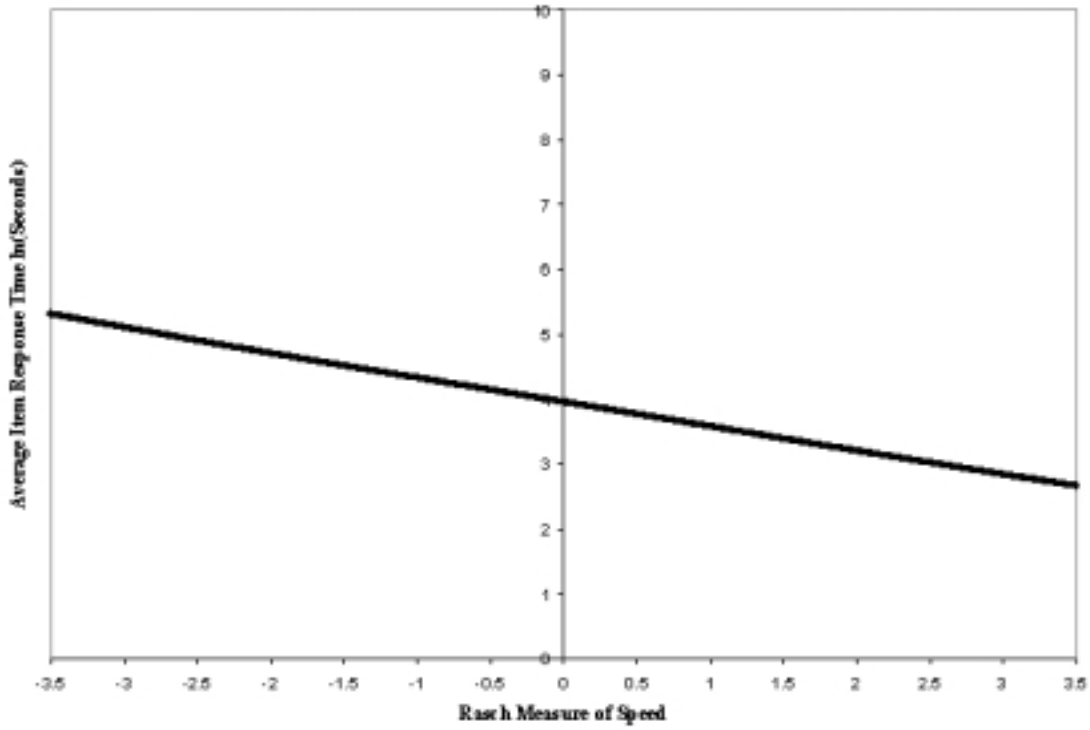

Figure 4. Average Item Response Time in ln(Seconds) by Rasch Measure of Speed

The speed of examinees in the grey zone sub-sample was also calculated for late part of the examination. The descriptive statistics for the distribution of speed in the late part of the examination are displayed in Table 8. The mean late speed was 0.55 logits with a standard deviation of 0.61 logits. Overall, the examinees continued to fit the model. The mean infit and outfit mean square were 1.01. However, some examinees did not fit the model at all. The maximum infit mean square was 9.90. These examinees were again slower than average. The mean late speed of the 178 examinees that had an infit mean square greater than 2.0 was 0.31 logits. Thirty-three of these examinees misfit in the early part of the examination as well. The examinees differed greatly in speed for the separation exceeded five.

TABLE 8.
SUMMARY OF LATE SPEED

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	755.8	152.6	0.55	0.10	1.01	-0.2	1.01	-0.2
S.D.	249.3	45.4	0.61	0.02	0.36	2.5	0.36	2.5
MAX.	1481.0	190.0	4.54	0.28	9.90	9.9	9.90	9.9
MIN.	33.0	16.0	-3.00	0.08	0.34	-8.1	0.34	-8.1

REAL RMSE	0.11	ADJ.SD	0.60	SEPARATION	5.70	PERSON RELIABILITY 0.97
MODEL RMSE	0.10	ADJ.SD	0.60	SEPARATION	6.07	PERSON RELIABILITY 0.97
S.E. OF PERSON MEAN	0.00					

The Item Difficulty Estimates

The dichotomous correct or incorrect result data from the early grey sub-sample were used to make initial estimates of the difficulty of the items in the pool. A summary of these estimates is found in Table 9. Seven items were answered correctly by everyone. Three were answered incorrectly by everyone. Nine items were not administered to anyone. Overall, the items fit the model quite well. The maximum infit mean square was 1.22 and the maximum outfit mean square was 2.24. None of the items with a sample size of greater than 50 had an infit or outfit mean square of greater than 1.5. In addition, the item separation exceeded 5 indicating that the items had a great deal of variation in difficulty. As expected, the items functioned quite well.

TABLE 9.
SUMMARY OF ITEM DIFFICULTY ESTIMATES FROM THE EARLY GREY SUB-SAMPLE

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	270.1	526.3	0.00	0.16	1.00	-0.1	1.01	-0.1
S.D.	227.9	459.3	1.29	0.15	0.03	0.9	0.08	0.9
MAX.	1382.0	2666.0	4.67	1.08	1.22	3.3	2.24	3.2
MIN.	1.0	5.0	-4.15	0.04	0.76	-3.7	0.43	-3.7
REAL RMSE	0.22	ADJ.SD	1.28	SEPARATION	5.82	ITEM	RELIABILITY	0.97
MODEL RMSE	0.22	ADJ.SD	1.28	SEPARATION	5.88	ITEM	RELIABILITY	0.97
S.E. OF ITEM MEAN	0.03							
WITH 10 EXTREME ITEMS	=	1794 ITEMS		MEAN	-0.01	S.D.	1.33	
REAL RMSE	0.26	ADJ.SD	1.30	SEPARATION	5.03	ITEM	RELIABILITY	0.96
MODEL RMSE	0.26	ADJ.SD	1.30	SEPARATION	5.06	ITEM	RELIABILITY	0.96

Because of the in depth scrutiny of the NCLEX item review, there was no need to certify the stability of the difficulty estimates across different samples of examinees. Still, the stability of items across conditions of speededness was tested. The data from the late part of the examination were used to calibrate the items a second time. Since there were several items in the late sub-sample that were not taken by any examinees, it was necessary to link the calibrations by setting the mean item difficulty to 0.292 logits. This was the mean difficulty from the early calibration of the 775 items that had a sample size of greater than 50 in both settings. Figure 5 shows the difficulty estimates of the 775 items. Overall, the items behaved in a stable manner. Nonetheless, the very easy items got easier and the very hard items got harder. In addition, the easy items got harder and the hard items got easier. This is explainable. In the late part of the examination, the adaptive algorithm targeted the difficulty of the items to the ability of the examinees better than in the early part. This allowed for greater differentiation between the difficulty of similar items. As a result, the extremes spread apart while the items in the middle got squished up. More information on this phenomenon, common in adaptive testing, is found in (Bergstrom and Lunz, 1994). The early difficulty estimates of the items were sufficient for use and were thus anchored for all future analyses.

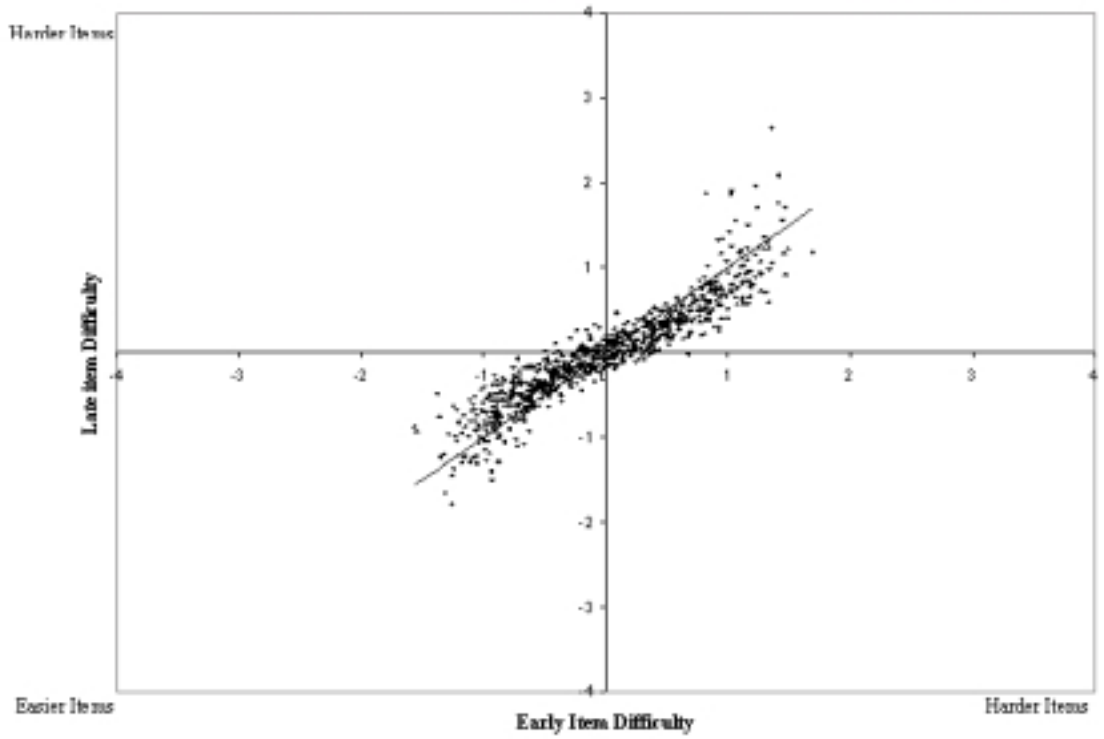

Figure 5. Late Item Difficulty plotted against Early Item Difficulty

Examinee Ability

The ability estimates of the early grey examinees are summarized in Table 10. The mean early ability was 0.39 logits with a standard deviation of 0.43 logits. Typical of an adaptive test, the examinees fit quite well. The maximum infit mean square was 1.38 while the maximum outfit was larger at 4.03. Other research on examinee fit in adaptive testing (Bradlow, 1997) has shown that misfitting examinees are ones that experience warm-up effects. The separation was only one meaning that there was not much variance in the ability estimates of the early grey examinees.

TABLE 10.
SUMMARY OF EARLY ABILITY

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	30.8	60.0	0.39	0.27	1.00	-0.1	1.00	-0.2
S.D.	3.4	0.1	0.43	0.01	0.07	0.9	0.10	0.8
MAX.	43.0	60.0	1.36	0.32	1.38	2.8	4.03	5.4
MIN.	16.0	46.0	-2.65	0.26	0.81	-3.4	0.79	-3.3

REAL RMSE	0.28	ADJ.SD	0.33	SEPARATION	1.20	PERSON RELIABILITY 0.59
MODEL RMSE	0.27	ADJ.SD	0.33	SEPARATION	1.23	PERSON RELIABILITY 0.60
S.E. OF PERSON MEAN	0.00					

The ability estimates of the examinees in the late part of the examination are summarized in Table 11. The mean ability was 0.48 logits with a standard deviation of 0.41 logits. Typical of an adaptive test, the examinees fit very well. The maximum infit and outfit mean square were less than 1.28. The separation of the examinees was greater than two indicating that there was enough variation in the examinees to differentiate them into about two groups.

TABLE 11.
SUMMARY OF LATE ABILITY

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	77.3	152.6	0.48	0.17	1.02	0.6	1.02	0.6
S.D.	22.3	45.4	0.41	0.03	0.04	1.0	0.04	1.0
MAX.	113.0	190.0	1.40	0.52	1.21	3.9	1.28	3.9
MIN.	6.0	16.0	-0.72	0.15	0.89	-3.1	0.88	-2.9

REAL RMSE	0.18	ADJ.SD	0.37	SEPARATION	2.04	PERSON RELIABILITY 0.81
MODEL RMSE	0.18	ADJ.SD	0.37	SEPARATION	2.07	PERSON RELIABILITY 0.81
S.E. OF PERSON MEAN	0.00					

Examinee Change Scores

The change in speed was calculated for each examinee by subtracting the early speed from the late speed. The distribution of change in speed is seen in Figure 6. The mean change in speed was 0.50 logits, an expected increase in speed. This change was significantly different from zero. This supports Hypothesis A; examinee speed is faster during the late part of the examination than during the early part of the examination.

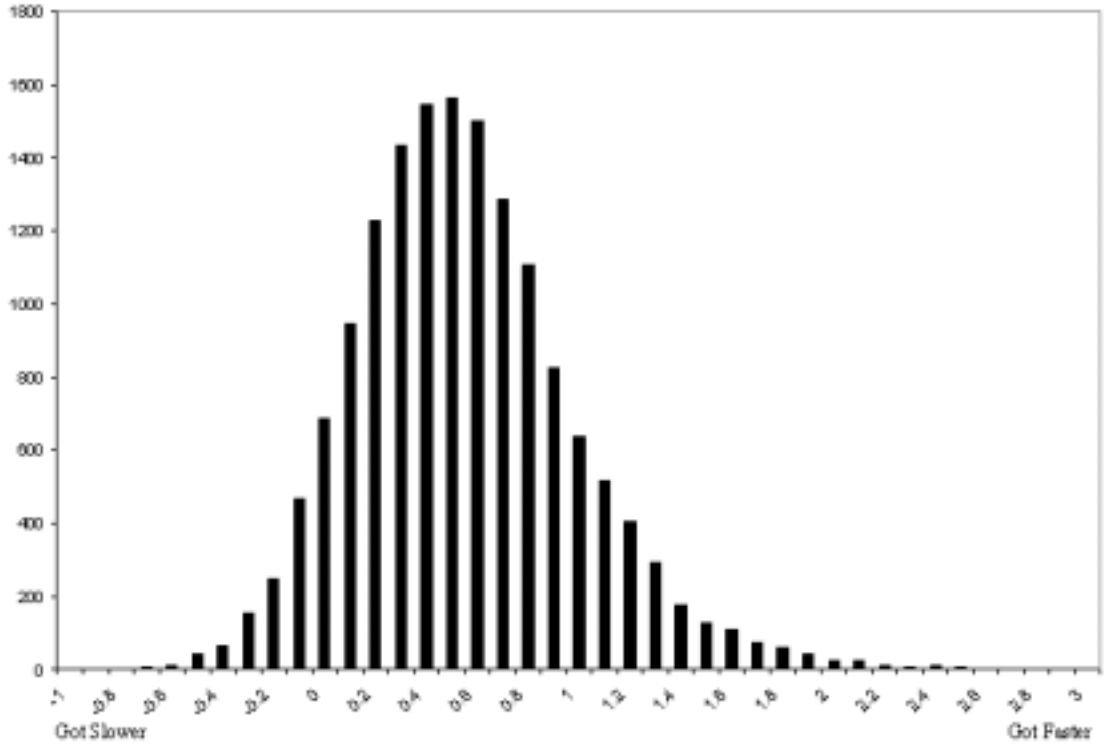

Figure 6. The Distribution of Change in Speed in logits

The change in ability was calculated for each examinee by subtracting the early ability from the late ability. The distribution of change in ability is seen in Figure 7. The mean change in ability was 0.09 logits, an unexpected increase in ability. This contradicts hypothesis C; examinees actually perform better under timed conditions.

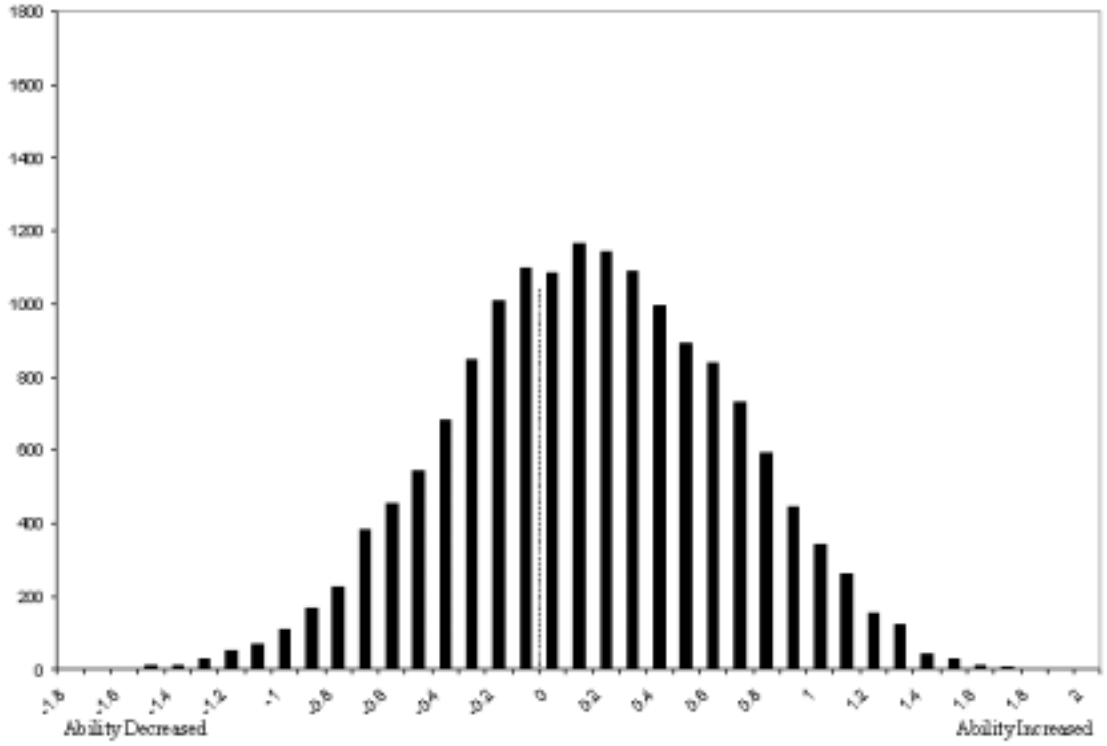

Figure 7. The Distribution of Change in Ability in logits

These two distributions summarize the overall speededness of the examination. The time limit caused examinees to speed up ½ a logit, a significant change both practically and statistically speaking. However, this increase in speed did not result in

an overall decrease in ability. Rather, it increased the ability of the examinees by almost a 1/10 of a logit.

Additional Inquiry

Having successfully assessed the speededness of the examination, the relationship between the behavior of examinees in untimed conditions and timed conditions was investigated deeper. First, the relationship between early speed and late speed was investigated. Then, the relationship between early ability and late ability was investigated. Finally, the relationship between speed and ability was assessed.

Each examinee's early speed was plotted against his(her) late speed. This is shown in Figure 8. This figure shows us again that examinees sped up. It also shows us that speed in the early part was strongly correlated with speed in the late part of examination ($p=0.80$).

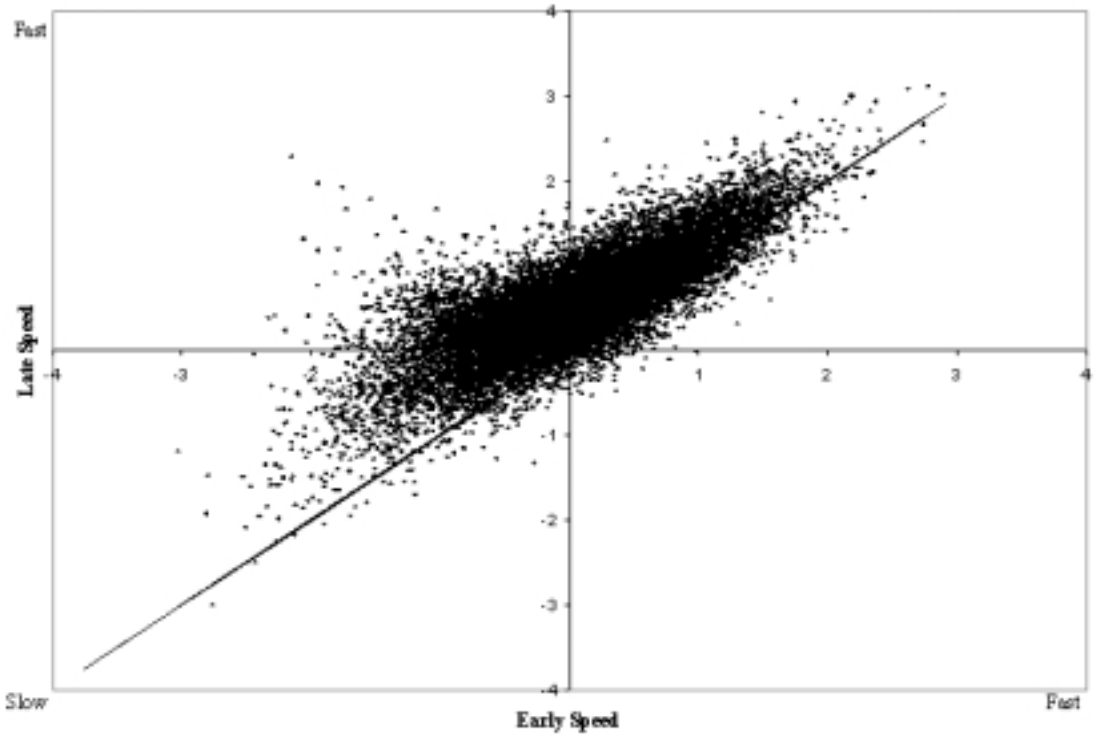

Figure 8. Late Speed plotted against Early Speed

The relationship between change in speed and early speed was also investigated. The change in speed was plotted against early speed. This is seen in Figure 9. The correlation between early speed and change in speed was -0.55. This figure supports Hypothesis B; change in speed is inversely related to early speed.

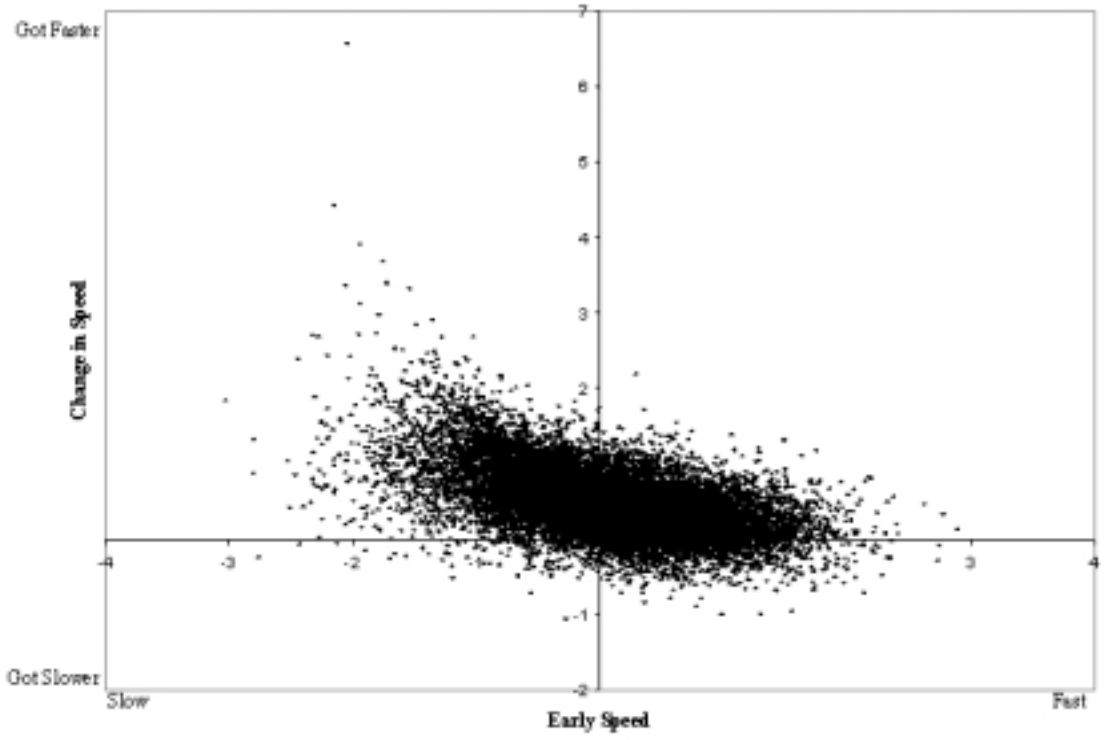

Figure 9. Change in Speed plotted against Early Speed

The relationship between early ability and late ability was investigated. Figure 10 displays the late ability plotted against the early ability of these examinees. The correlation between early ability and late ability was 0.22, a weak positive relationship. It is evident by the plot that many examinees improved and many got worse.

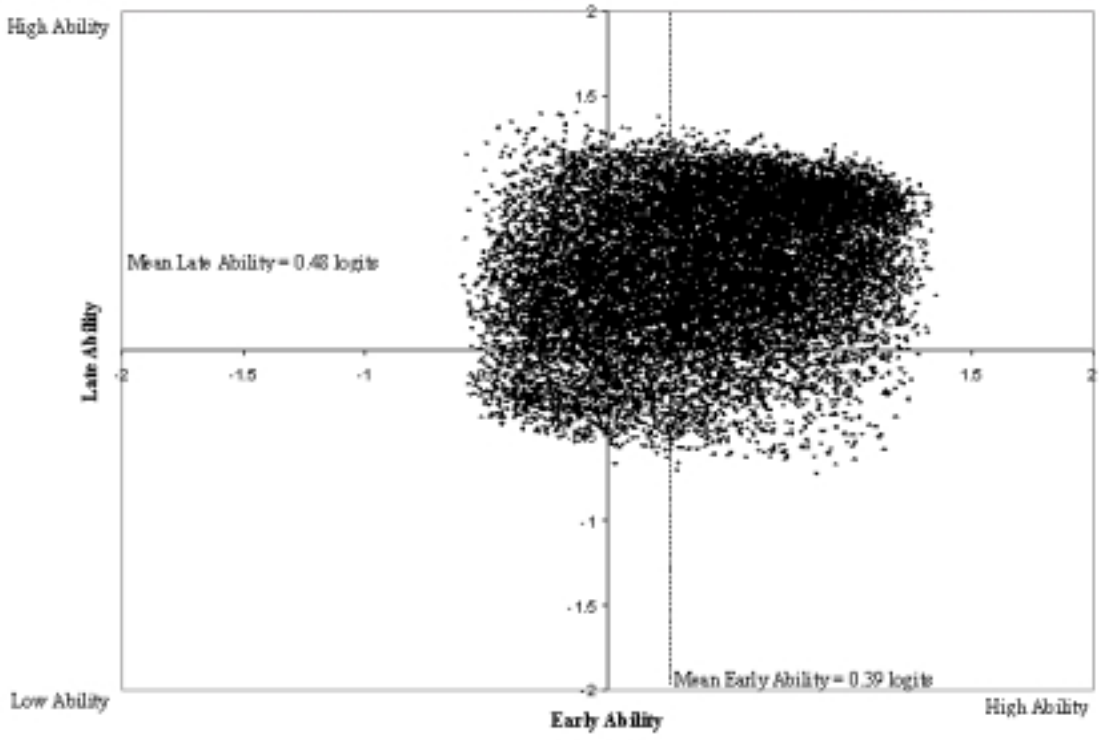

Figure 10. Late Ability plotted against Early Ability

The relationship between speed and ability was investigated. Figure 11 displays the early ability plotted against early speed. The correlation was 0.06. It is evident that there was no relationship between speed and ability.

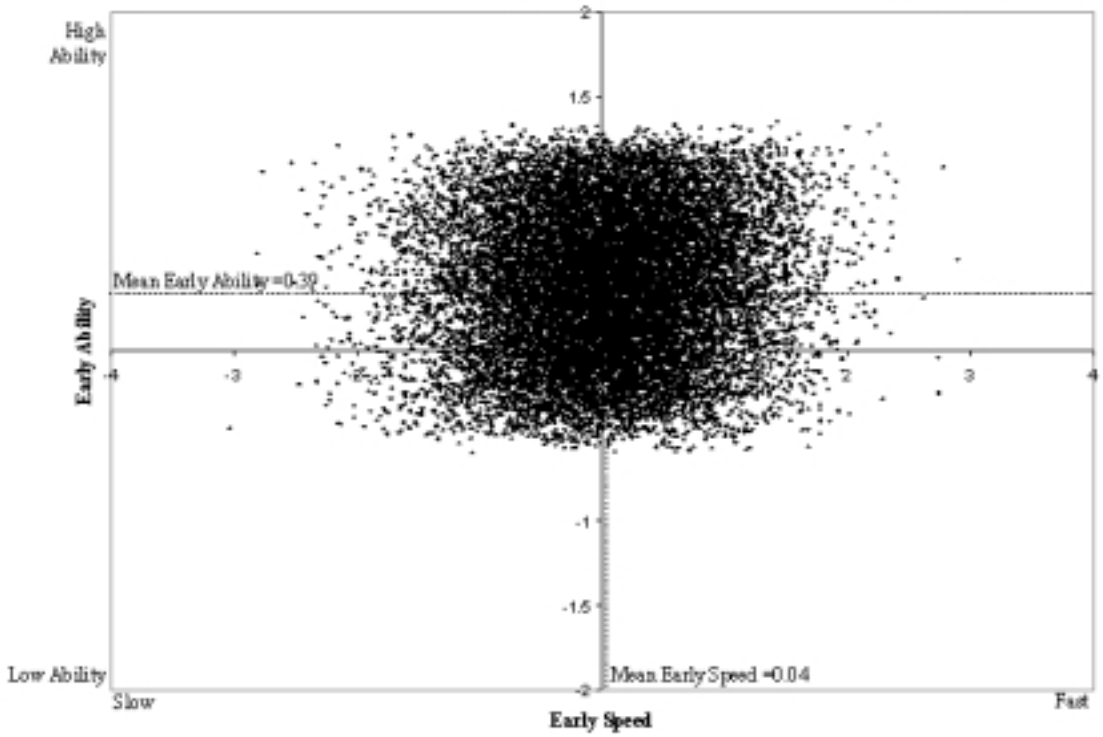

Figure 11. Early Ability plotted against Early Speed

The relationship between speed and ability was investigated for the late part of the examination. Figure 12 displays late ability plotted against late speed. The correlation was 0.01. Again, there was no relationship between speed and ability.

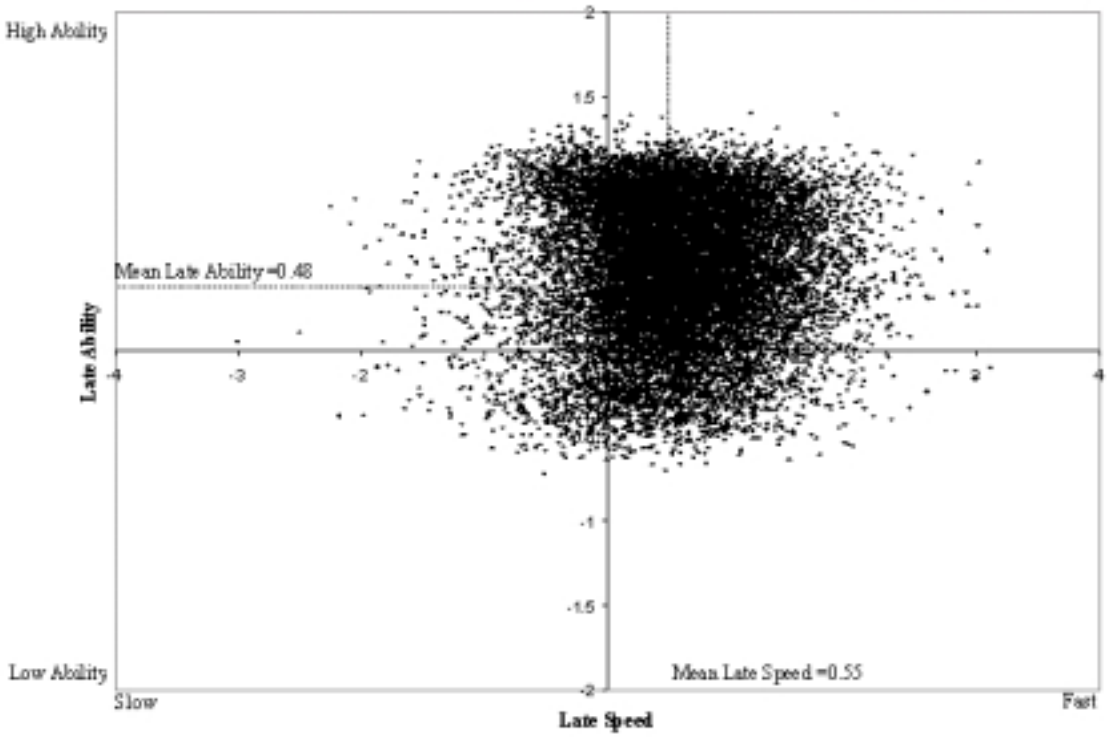

Figure 12. Late Ability plotted against Late Speed

This led to an investigation of the relationship between change in speed and change in ability. Figure 13 displays the change in ability plotted against the change in speed. The figure shows that there was, on average, no change in ability as a result of a change in speed. The combined correlation was –0.05. This rejects hypothesis D; change in speed is not related to change in ability.

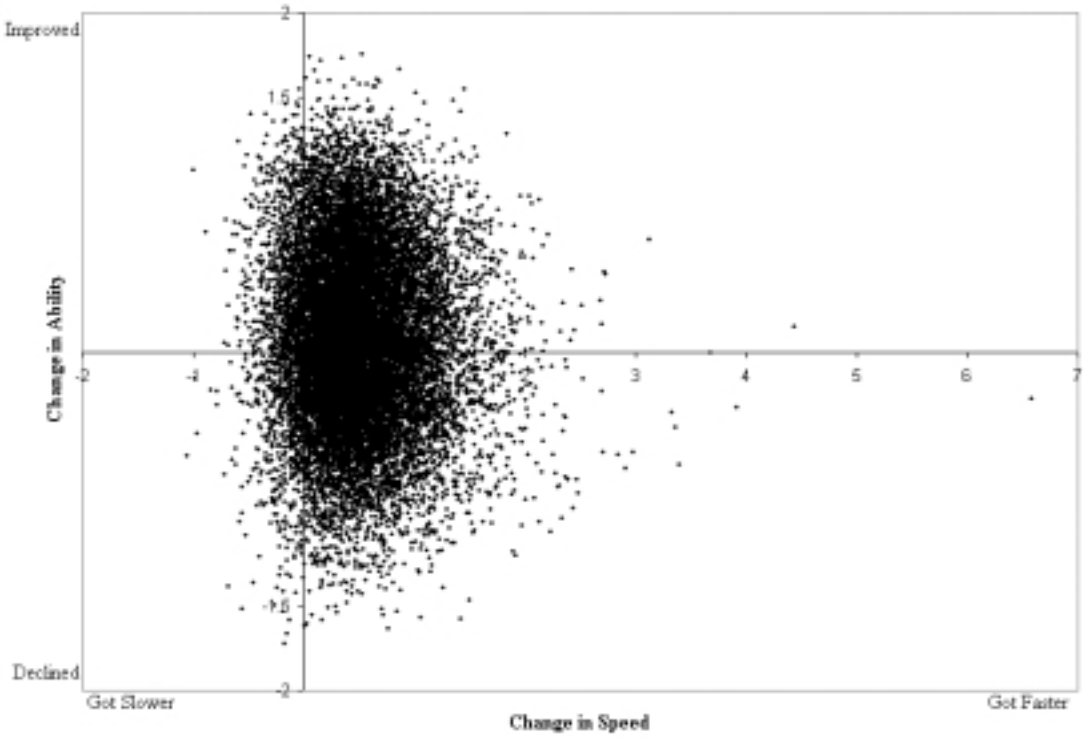

Figure 13. Change in Ability plotted against Change in Speed

The consistency of speed within section was assessed. For each examinee, the original item response times were regressed onto the item sequence numbers. The distribution of the slope of the regression lines for each examinee in the early part of the examination is shown in Figure 14. The mean was -0.08 with a standard deviation of 0.35. For the most part, examinees maintained a consistent speed during the early part of the examination.

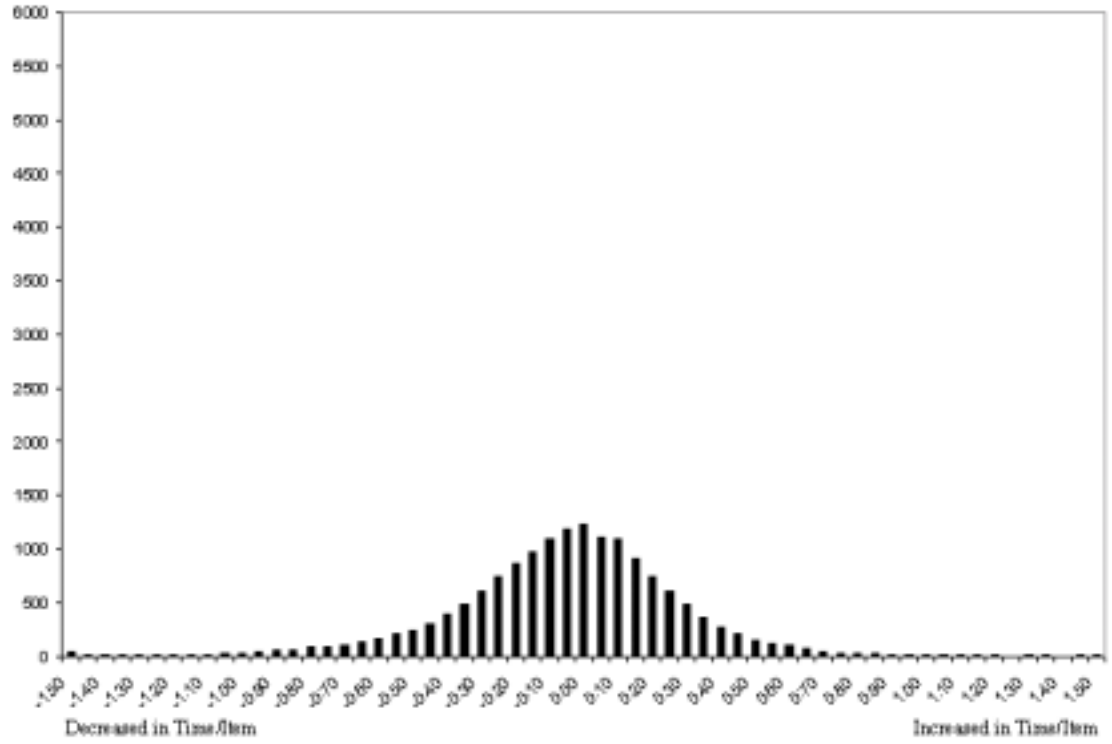

Figure 14. Distribution of Early Regression Slopes

The data from late part of the examination were used to calculate the same regression slopes. The distribution of these slopes is displayed in Figure 15. The mean was –0.09 with a standard deviation of 0.14. This mean is similar to the mean of the early regression slopes. However, the variance of the slope in the late part of the examination was smaller. Examinee speed was more consistent in the latter part of the examination than in the early part. This supports hypothesis E; examinee speed was consistent within each part of the examination. The skew in this distribution is worthy of comment. The skew shows that examinees were more likely to speed up within the latter part of the examination than to slow down. Rapid guessing behavior largely contributed to this trend. There were 2,380 responses that were made in less than 7.4 seconds, thus receiving a speed rating of nine. Over 1,100 examinees made at least one of these responses and 135 examinees made more than 10 of these responses.

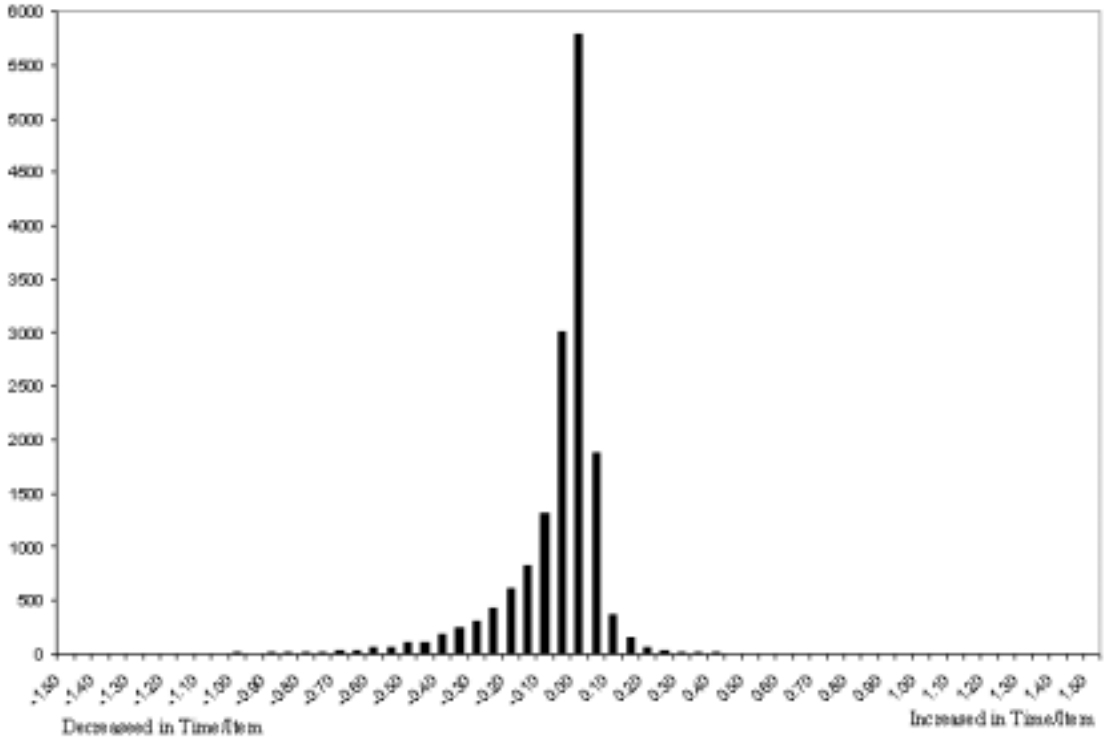

Figure 15. Distribution of Late Regression Slopes

Speed and Ability by Examinee Demographics

Examinee demographics were investigated. In particular, the ethnicity, gender, ESL status, and educational program type were probed. Overall, ethnicity and English language proficiency did have an impact on the speed and ability of the examinees. Gender and type of nursing education program did not.

The mean speed by ethnicity is shown in Table 12. This table provides further support for the conclusion that slower examinees changed speed more than faster examinees. The Whites and Native Americans worked the fastest both in the early and late part of the examination. As expected, their speed changed the least. The Hispanics worked 0.3 logits slower than the Whites in the early part of the examination and changed speed more. The Pacific Islanders were 0.4 logits slower than the Whites in the early part of the examination and changed speed more than the Hispanics. The Blacks were 0.5 logits slower than the whites and changed speed the most of all. The Asians were the slowest, 0.6 logits slower than the whites in the early part of the examination and changed speed almost as much as the Blacks did.

TABLE 12.
MEAN SPEED BY ETHNICITY

	N	Early Speed		Late Speed		Change in Speed	
		Mean	St Dev	Mean	St Dev	Mean	St Dev
Asian Indian	115	-0.46	0.76	0.15	0.59	0.58	0.45
Asian Other	695	-0.44	0.75	0.16	0.63	0.59	0.55
Black	1490	-0.33	0.70	0.27	0.57	0.60	0.47
Hispanic	684	-0.15	0.69	0.39	0.57	0.54	0.49
Native American	129	0.11	0.76	0.55	0.64	0.44	0.45
Pacific Islander	135	-0.24	0.78	0.32	0.63	0.57	0.52
White	11826	0.16	0.70	0.63	0.59	0.48	0.42
Missing	579	-0.32	0.81	0.25	0.67	0.57	0.50
Total	15653	0.04	0.73	0.55	0.61	0.50	0.44

The ability of the examinees by ethnicity was calculated. This is displayed in Table 13. The Whites and Native Americans were the most able in both the early and late part of the examination. These two sub-groups also increased in ability more than the other sub-groups did, 0.11 and 0.13 logits respectively. The Hispanics were only 0.01 logits less able than these two sub-groups in the early part of the examination. However, they improved only 0.05 logits in ability. The Pacific Islanders had an early ability of 0.37 logits, 0.03 logits less than the Whites. This group did not improve in ability from early to late. The Blacks had an early ability of 0.36 logits, 0.04 less than the Whites. This group improved less than the Whites, 0.02 logits respectively. The Asians were the weakest, 0.28 and 0.34 logits in the early part and 0.35 and 0.32 logits in the late part. The Asian Others decreased in ability by 0.02. The Asian Indians improved 0.06 logits.

TABLE 13.
MEAN ABILITY BY ETHNICITY

	N	Early Ability		Late Ability		Change in Ability	
		Mean	St Dev	Mean	St Dev	Mean	St Dev
Asian Indian	115	0.28	0.53	0.35	0.42	0.06	0.60
Asian Other	695	0.34	0.43	0.32	0.44	-0.02	0.58
Black	1490	0.36	0.44	0.38	0.41	0.02	0.53
Hispanic	684	0.39	0.42	0.43	0.41	0.05	0.52
Native American	129	0.40	0.43	0.53	0.41	0.13	0.57
Pacific Islander	135	0.37	0.43	0.38	0.44	0.00	0.53
White	11826	0.40	0.43	0.51	0.40	0.11	0.52
Missing	579	0.34	0.43	0.40	0.41	0.06	0.54
Total	15653	0.39	0.43	0.48	0.41	0.09	0.52

The mean speed and ability were investigated by English language proficiency. Mean speed by ESL status is displayed in Table 14. Examinees whose primary language was English were much faster than non-native speakers. Non-native speakers that professed to speak English well were 0.50 logits slower than native speakers and non-native speakers that did not profess to speak English well were 0.68 logits slower. Non-native speakers changed speed 0.10 logits more than native speakers.

TABLE 14.
MEAN SPEED BY ESL STATUS

	N	Early Speed		Late Speed		Change in Speed	
		Mean	St Dev	Mean	St Dev	Mean	St Dev
Primary	14180	0.10	0.72	0.59	0.60	0.49	0.43
Well	1050	-0.40	0.74	0.20	0.60	0.59	0.51
Second	377	-0.58	0.68	0.00	0.60	0.59	0.54
Missing	46	-0.31	0.94	0.36	0.75	0.67	0.63
Total	15653	0.04	0.73	0.55	0.61	0.50	0.44

The mean ability by ESL status is displayed in Table 15. The mean ability of native speakers was 0.39 logits, 0.03 logits higher than the mean ability of non-native speakers that professed to speak English well. This was 0.06 logits higher than non-native speakers that did not profess to speak English well. The ability of native speakers increased by 0.1 logits while the ability of non-native speakers did not change.

TABLE 15.
MEAN ABILITY BY ESL STATUS

	N	Early Ability		Late Ability		Change in Ability	
		Mean	St Dev	Mean	St Dev	Mean	St Dev
Primary	14180	0.39	0.43	0.49	0.40	0.10	0.52
Well	1050	0.36	0.44	0.35	0.43	-0.01	0.56
Second	377	0.33	0.44	0.34	0.43	0.01	0.55
Missing	46	0.20	0.43	0.34	0.31	0.14	0.50
Total	15653	0.39	0.43	0.48	0.41	0.09	0.52

The mean ability and speed were investigated by nursing program type. The mean speed by educational program type is displayed in Table 16. The Diploma examinees worked only slightly faster (0.01 logits) than the Associate examinees which worked only slightly faster (0.02 logits) than the Baccalaureate examinees. The Diploma examinees changed speed (0.04 logits) more than the other two sub-groups.

TABLE 16.
MEAN SPEED BY PROGRAM TYPE

	N	Early Speed		Late Speed		Change in Speed	
		Mean	St Dev	Mean	St Dev	Mean	St Dev
Diploma	612	0.08	0.70	0.62	0.59	0.54	0.44
Associate	9071	0.07	0.73	0.57	0.61	0.50	0.44
Baccalaureate	5316	0.05	0.72	0.54	0.59	0.50	0.44
Missing	654	-0.42	0.80	0.14	0.71	0.56	0.56
Total	15653	0.04	0.73	0.55	0.61	0.50	0.44

The groups were almost identical in ability and change in ability. The mean ability by educational program type is displayed in Table 17. The mean of the Diploma examinees was 0.41 logits only 0.02 logits higher than the other two. There was no significant effect of program type on the behavior of the examinees.

TABLE 17.
MEAN ABILITY BY PROGRAM TYPE

	N	Early Ability		Late Ability		Change in Ability	
		Mean	St Dev	Mean	St Dev	Mean	St Dev
Diploma	612	0.41	0.41	0.50	0.39	0.10	0.52
Associate	9071	0.39	0.43	0.49	0.40	0.10	0.52
Baccalaureate	5316	0.39	0.43	0.49	0.40	0.10	0.52
Missing	654	0.32	0.45	0.28	0.43	-0.05	0.58
Total	15653	0.39	0.43	0.48	0.40	0.09	0.52

Lastly, the speed by gender was investigated. The means speed by gender is displayed in Table 18 and the mean ability by gender is displayed in Table 19. Females worked faster than males by 0.12 logits. The two sub-groups changed speed in a similar fashion. The ability by gender was also investigated. No differences were found between males and females.

TABLE 18.
MEAN SPEED BY GENDER

	N	Early Speed		Late Speed		Change in Speed	
		Mean	St Dev	Mean	St Dev	Mean	St Dev
Female	13605	0.06	0.74	0.56	0.61	0.50	0.44
Male	1990	-0.06	0.71	0.43	0.58	0.49	0.44
Missing	58	-0.13	1.00	0.43	0.87	0.56	0.59
Total	15653	0.04	0.73	0.55	0.61	0.50	0.44

TABLE 19.
MEAN ABILITY BY GENDER

	N	Early Ability		Late Ability		Change in Ability	
		Mean	St Dev	Mean	St Dev	Mean	St Dev
Female	13605	0.39	0.43	0.48	0.41	0.09	0.52
Male	1990	0.38	0.43	0.47	0.40	0.10	0.52
Missing	58	0.29	0.43	0.38	0.35	0.10	0.53
Total	15653	0.39	0.43	0.48	0.40	0.09	0.52

Analysis of the Items

In order to understand the items more completely, some additional analyses were conducted. The fit of the items to the duration scale was probed. There was no motivation to probe the fit of the items to the difficulty scale since all items fit the model. In addition, the relationship between difficulty and duration was probed.

The fit of individual items to the model was investigated. The infit and the outfit of individual items were virtually the same. Figure 16 displays the early outfit mean square plotted against the early infit means square. It is clear, that these two indicators of fit are virtually identical. The items fit the same way regardless of whether they were given to slow or fast examinees. This was also true of the fit of the items in the late part of the examination.

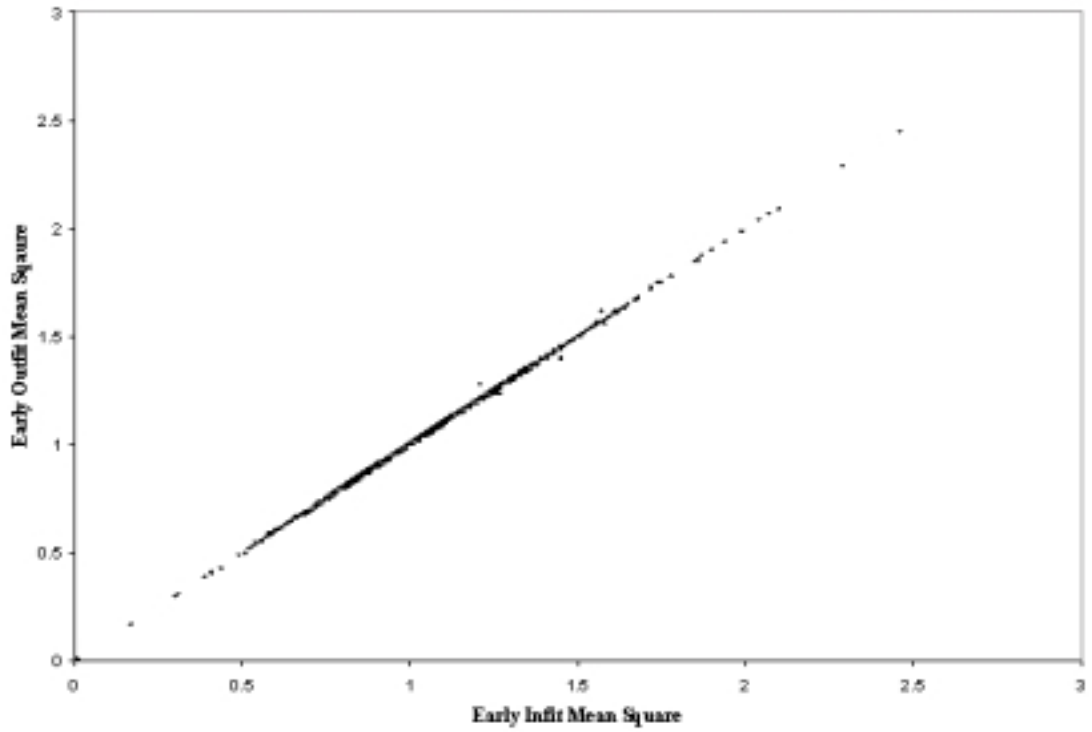

Figure 16. Early Item Outfit Mean Square plotted against Early Item Infit Mean Square

The fit of the items to the duration scale was similar in both the early and late part of the examination. Figure 17 displays the infit of the items to the duration scale in the late part plotted against the infit of the items in the early part. These are only the items that had a sample size of greater than 50 in both samples. Only 27 items having a sample size of greater than 50 had an Infit or Outfit Mean Square greater than 1.5. The content area of the items is seen in Table 20. The majority of the items came from Physiological Integrity the area with the most items. None of the items were from Psychosocial Integrity. The mean duration of these items was -0.62 logits indicating that the misfitting items were shorter than average. The sample size of these items varied, some had low sample size, and others had high.

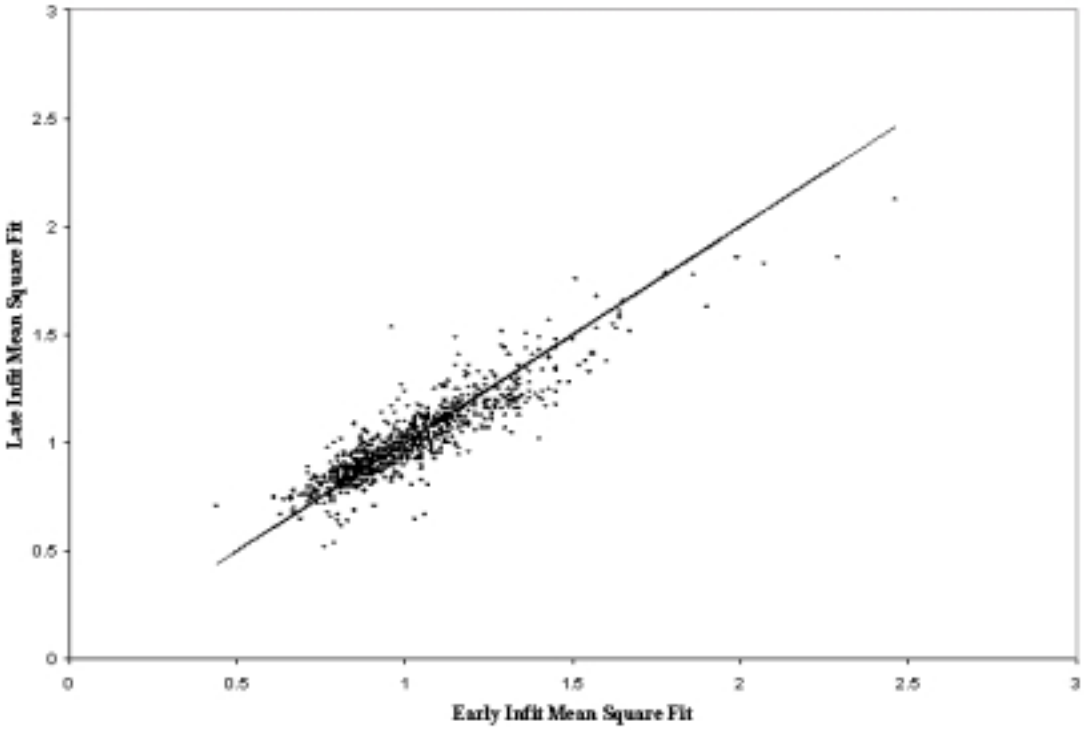

Figure 17. Late Item Mean Square Fit plotted against Early Item Mean Square Fit

TABLE 20.
CONTENT AREA OF ITEMS THAT MISFIT ON DURATION

Content Area	**Frequency**
Safe, Effective Care Environment	
Management of Care (10%)	0
Safety and Infection Control (8%)	1
Health Promotion and Maintenance	
Growth Development Through the Life Span (10%)	2
Prevention and Early Detection of Disease (8%)	1
Psychosocial Integrity	
Coping and Adaptation (8%)	0
Psychosocial Adaptation (8%)	0
Physiological Integrity	
Basic Care and Comfort (10%)	3
Pharmacological and Parenteral Techniques (8%)	6
Reduction of Risk Potential (15%)	9
Physiological Adaptation (15%)	5
Total	27

The relationship between fit and duration was investigated. Figure 18 is a plot of the early infit mean square against the early item duration estimate. There was an inverse relationship between fit and duration. Longer items tended to fit better than shorter items.

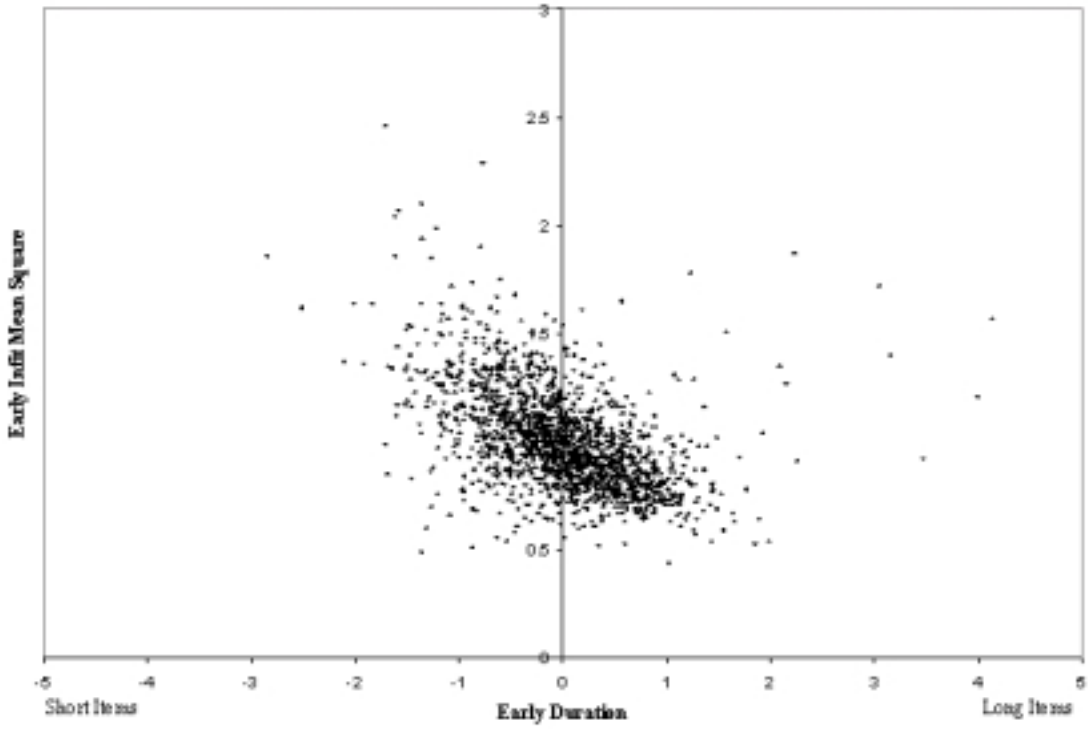

Figure 18. Early Item Infit Mean Square plotted against Early Item Duration

This same relationship was also evident in the late part of the examination. Figure 19 shows the late item infit mean square plotted against the late item duration. Again, longer items tended to fit better than short items.

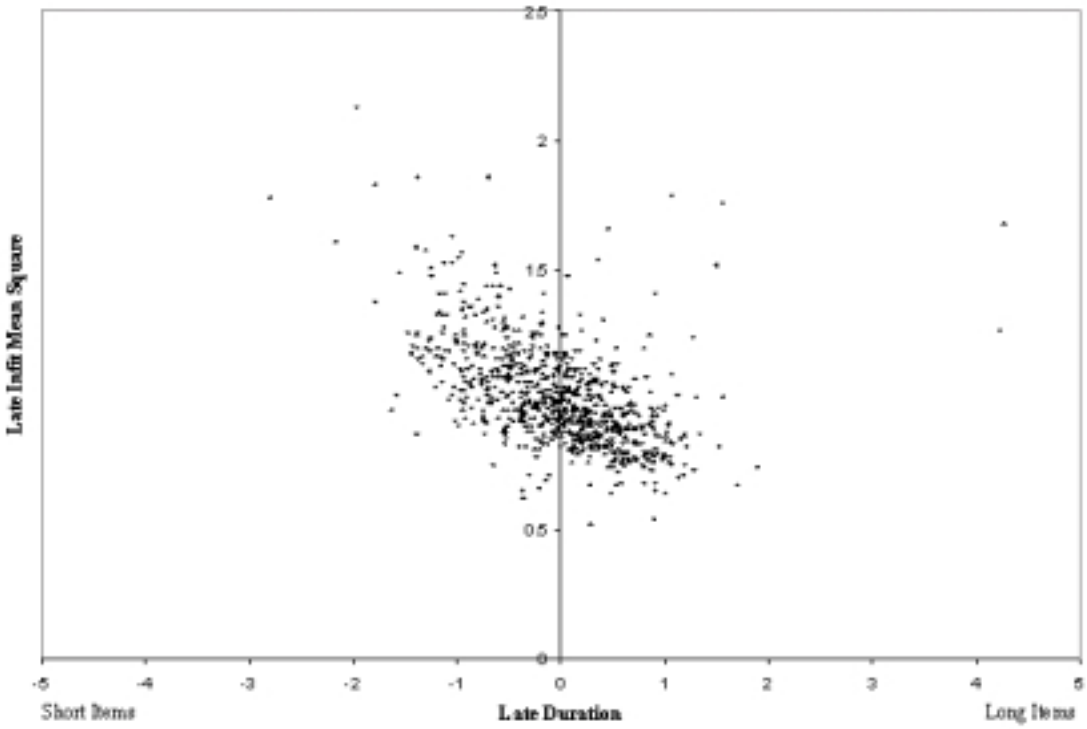

Figure 19. Late Item Infit Mean Square plotted against Late Item Duration

The relationship between duration and difficulty was investigated. Figure 20 shows the early item difficulty plotted against early item duration. There is very little relationship between the two.

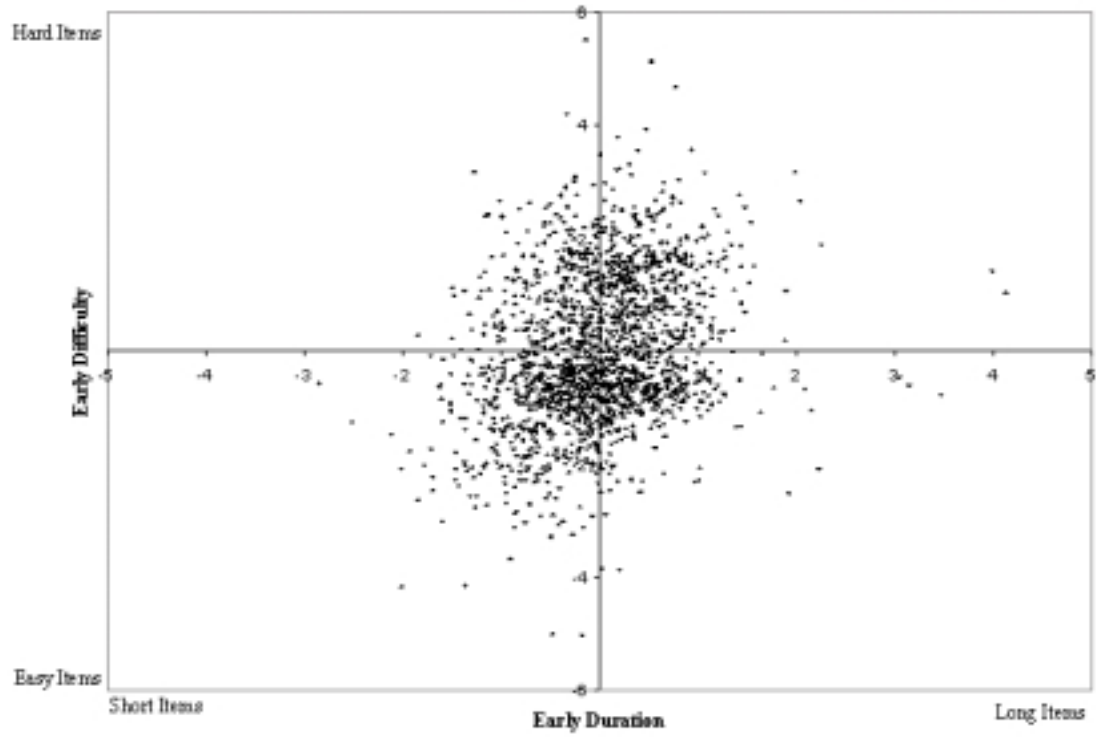

Figure 20. Early Item Difficulty plotted against Early Item Duration

This same relationship was investigated for the late part of the examination. Figure 21 displays the difficulty of the items in the late part of the examination plotted against the duration of the items in the late part of the examination. Again, there was very little relationship.

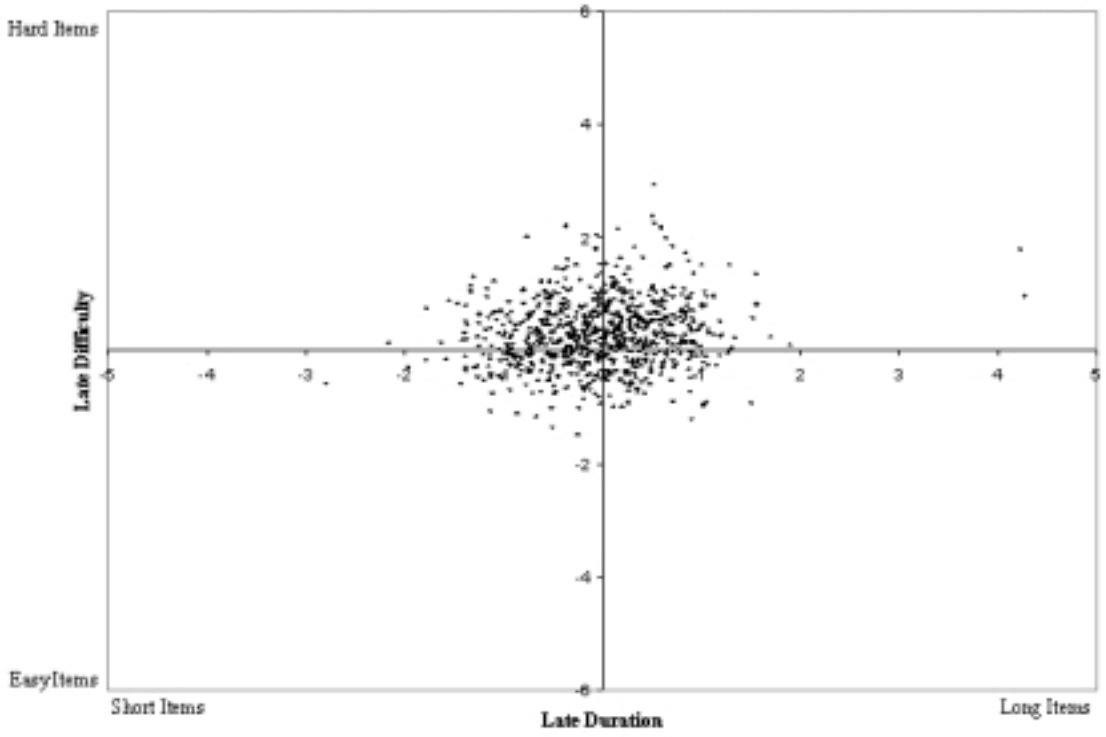

Figure 21. Late Item Difficulty plotted against Late Item Duration

Analysis of Residuals

The relationship between item response time and result was investigated on the individual response level. This was done using the residual of response time and the residual of the result (correct/incorrect). The residual is the difference between the expected value and the observed value. The expected value for result was determined based on the ability of the examinee and the difficulty of the item. Since the NCLEX-RN is an adaptive test targeted at 50% probability on each item, result residuals were around 0.5 and -0.5. Positive residuals were items that were answered correctly and negative values, incorrect. The expected value of response time was determined based on the speed of the examinee and the duration of the item. Positive residuals were items answered quicker than expected, negative residuals were items answered slower than expected.

Figure 22 shows the residuals of result plotted and against the residuals of response time for all response that had a residual of response time greater than two. These are all of the responses that were quicker than expected. The data points on the left were responses that were incorrect and those on the right were correct. Responses that were faster than expected, yielded correct responses more often than incorrect responses. In fact, when the residual of response time was between two and three, 61% of the responses were correct regardless of the difficulty of the item. However, items that were answered at least three logits quicker than expected resulted in only 40%

correct responses. The difficulty of an item for a particular examinee had little effect on the rate at which the examinee answered the item.

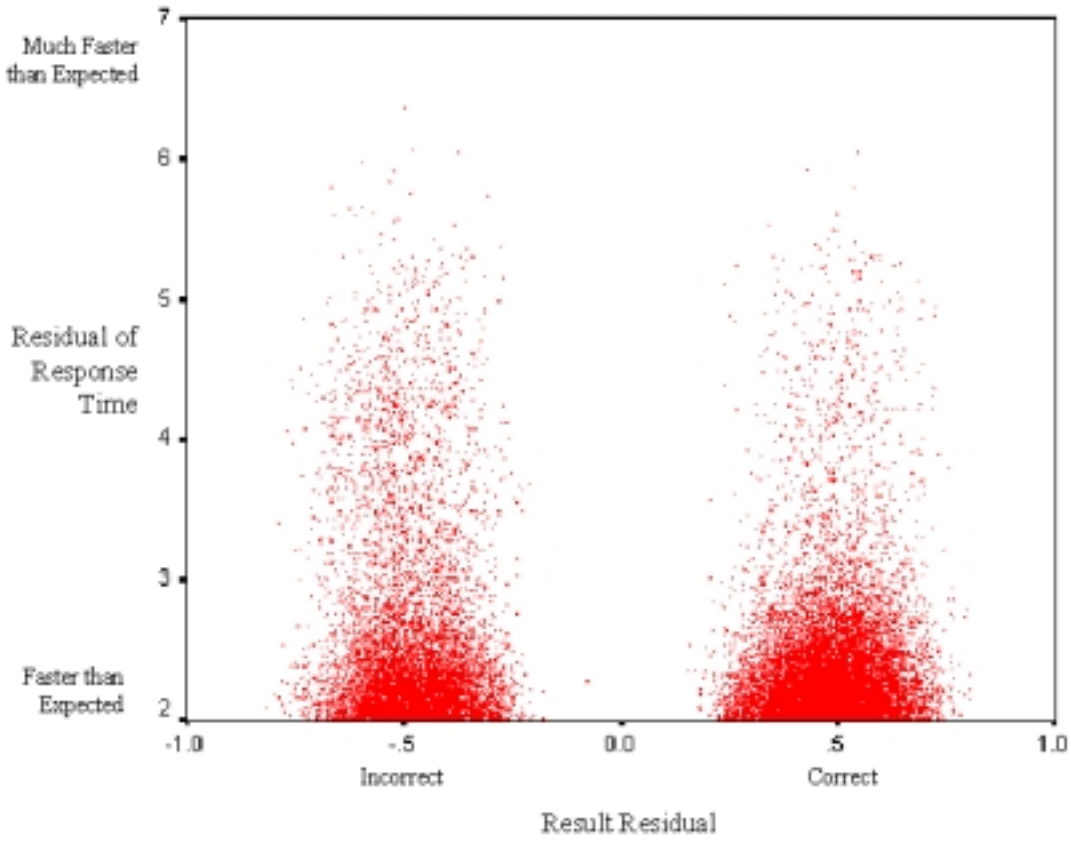

Figure 22. Residual of Response Time plotted against Residual of Results

CHAPTER 4

CONCLUSION

The method proposed in this study for assessing speededness proved successful. The technique used for developing measures of speed was potent. The ten-point speed rating scale performed well and item duration remained stable across settings. This allowed for an objective assessment of examinee speed across settings. By objectively assessing examinee speed under different time limit considerations, the extent to which time limits effected individual behavior was accurately evaluated.

The speededness of the NCLEX-RN® examination was tolerable. As a group, the examinees who took more than 120 items, sped up. However, this did not effect their overall performance in a negative way. Nonetheless, there were some examinees that declined in performance dramatically. Future research should probe other factors such as fatigue to find out what contributed to these examinees' decline.

The additional inquiry showed that examinees do work at a consistent enough pace to provide useful measures of speed. The examinees who performed the most inconsistently were examinees that rapidly guessed towards the end of their examination. Their speed rating scale values for the final few items were nine. These

examinees spent less than 7.4 seconds on the last few items. This resulted in a negative slope of item response time and a high degree of misfit. The test developers could prevent this behavior by implementing a minimum time per item, such as 10 seconds per item. This would force examinees to try on each item rather than guess. Operationally, this is feasible with today's computer technology.

The additional inquiry also revealed that on the individual response level, unexpectedly quick responses, did not result in decreased success. In fact, on unexpectedly quick items, success was higher than the expected 50%. That is, until the response was made so quickly that it was essentially a rapid guess. Then, the success rate dropped off.

The demographics of examinees that were effected by the time limit were not surprising. The most significant factor in predicting slow speed was familiarity with the English language. Non-native speakers were slower. For this group, increases in speed did not result in decreased performance. Still, non-native speakers did not increase in performance as native speakers did.

Ethnicity also proved to be a predictor of speed and ability. Blacks and Asians were slowest and weakest. These two groups also changed speed the most. And, they did not increase their ability. Future research probing these groups further may hold interesting findings.

The relevance of this study to the field of education is primarily test based. Cost and resources will always precipitate the need for time limits on tests. As long as there

are time limits, there is the potential for these time limits to effect examinees. This study makes it easier and more accurate for test developers to build more objective tests by providing a method that comprehensively assesses the impact of the time limit on examinees.

Additionally, this study provides some support for the notion that time limits can cause examinees to speed up without detracting from their performance. A more thorough investigation into the impact of the time limit on different sub-groups is necessary before this notion should be endorsed. This finding holds promise for test makers who need to develop more stringent time limits because of other operational issues. In addition, this study reinforces the obvious; rapid guessing behavior does detract from performance. Test developers spend resources informing examinees of this effect. Still, there will always be some examinees who will engage in this type of behavior. The suggestion that test developers implement minimum response times will prevent this negative behavior in the future.

In conclusion, the intent of the research was to test a new method for assessing the phenomenon of speededness. This method was easy to implement and provided results that were easy to interpret. It is questionable whether or not many test designers will have the capability to collect this kind of data. Nonetheless, most new examination programs beta test their examinations before making final decisions about examination specifications such as the time limit. Collecting this kind of data during the beta stage and analyzing it using the methods contained within is reasonable to ask of test

developers. By investigating examination time limits with this method, test developers will further their pursuit of making tests as objective as possible while also keeping the costs of such examinations to a minimum.

REFERENCES

Andrich, D. (1978). A rating formulation for ordered response categories. Psychometrika, 2, 581-594.

Berstrom, B. and Lunz, M. (1994). The equivalence of item calibrations and ability estimates across modes of administration. In M. Wilson (Ed.), Objective Measurement: Theory into Practice Volume 2, Norwood, NJ: Ablex Publishing.

Bontempo, B. D. (1997). Creating objective measures of examinee speed and item length for the NCLEX-RN® Examination. A paper presented at the Midwest Objective Measurement Seminar, Chicago.

Bradlow, E. (1997). Bayesian identification of outliers in computerized adaptive tests. Unpublished Research. The National Council of State Boards of Nursing.

Cook, T. D. and Capmbell, D. T. (1979). Quasi-Experimentation: Design and Analysis Issues for Field Settings. Boston: Houghton-Mifflin.

Gulliksen, H. (1950). Theory of Mental Tests. Hillsdale, NJ: Lawrence Erlbaum.

Hubert, C. and Gorham, J. (1998). Technical report : NCLEX-RN® and NCLEX-PN® Examinations using computerized adaptive testing. Unpublished Statistical Report. The Chauncey Group International, LTD.

Linacre, M. (1999). A User's Guide to Winsteps. Chicago, MESA Press.

Rasch, G. (1980). Probabilistic Models for Some Intelligence and Attainment Tests. Chicago: MESA Press.

Schnipke, D. L. and 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.

Stone, G. (1994). The historical development of fit and its assessment in the computer adaptive testing environment. A paper presented at the annual meeting of the Midwest Educational Research Association, Chicago.

Swineford, F. (1956). Technical manual for users of test analysis. Statistical Report 56-42. Princeton, NJ: Educational Testing Service.

Wright, B. D. & Masters, G. N. (1982). Rating Scale Analysis. Chicago, MESA Press.

Wright, B. D. & Stone, M. H. (1979). Best Test Design. Chicago, MESA Press.