Assessing Speededness in Variable-Length Computer Adaptive Testing

Brian D. Bontempo
Ellen R. Julian

The National Council of State Boards of Nursing

March 15, 1997

A Paper Presented at the Annual Meeting of The National Council on Measurement in Education, Chicago 1997.

Assessing Speededness in Variable-Length Computer Adaptive Tests

Introduction

Speededness: The degree to which the amount of time allowed for test administration affects the rate at which examinees answer items.
This paper proposes that speededness be defined as how much faster examinees work on an examination with a particular time limit than they would have worked on an examination with an infinite amount of administration time. This definition does not include any assumptions about an impact on examinee performance. Instead, changes in examinee performance can be viewed as a possible effect of speededness and add an impetus for the creation of such a measure.

Assessing the speededness of a paper-and-pencil examination was relatively simplistic and inexact. In this mode of testing, there were at most two relevant variables available, the total time that each examinee spent on an examination and the total number of items that each examinee attempted. Assessments of speededness based only on these two variables failed to provide the information necessary, i.e., the time candidates spent on each item, to ensure a consistent degree of speededness across different items and thus different test forms.

This problem becomes complex with computerized adaptive tests (CATs) where every candidate is given a unique examination. This gets even more complex when the CAT is a variable-length CAT, i.e., a CAT where the number of items administered to an examinee varies depending on the distance of the examinee's ability estimates from the passing standard.

This research explored three different methods of analyzing speededness: (1) examinee response rate, (2) increase in examinee response rate within individual examinations, and (3) increase in examinee response rate across repeated examinations. The first method is a logical extension of traditional assessments of speededness, while the other two are relatively new. To provide a working example, these three methods were all performed on data from the NCLEX-RN™.

Literature Review

Gulliksen (1950) distinguished between examinations that measure only knowledge, called power tests, and those that also measure speed, called speeded tests. In pure power tests, examinees have a theoretically infinite amount of time to complete an examination. In the real world, an infinite amount of time is seldom available. "However, having a time limit does not necessarily mean that the test is speeded" (Schnipke, 1995). If only a small number of examinees do not finish, then the extent to which the time limit affected examinees' response speed (speededness---by definition), is small enough to be declared insignificant, and essentially the examination can be deemed a power test.

When conducting an examination with a time limit, there are two possible strategies for examinees that do not work fast enough naturally to complete the examination: some examinees work at their natural pace leaving some items unanswered and some hurry, even to the point of rapidly guessing at items as time nears expiration (Schnipke, 1995). It can be assumed that many examinees find a balance between these two extremes. Some examinees spend their energy (and time) on good performance while others answer as many items as possible. The unanswered items at the end of an incomplete examination are considered unreached items and present test makers with a missing data problem. There are three possible solutions. The missing data can be scored as incorrect. Data, usually random, can be imputed for the missing values. It is also possible to estimate the ability of the examinee based on the completed items, despite the resulting loss in precision (Julian and Bontempo, 1996; GRE, 1996).

Research has shown that an NCLEX-RN™ examinee's ability estimate generally increases from their first (failed) administration to the second (Gorham and Bontempo, 1996). An increase in response rate from the first administration to the second combined with either a stable or diminishing ability estimate may indicate that the examinee is answering questions at a rate above their capacity, resulting in ability estimates that are less than optimal. Of course an examinee's increased response rate and any change in ability may be due to several things besides their effort to keep a quicker pace. For example, an examinee may have taken a refresher course in the intervening time, increasing their recall, or they may have become more familiar with the computers and the administration process, thus increasing their response rate. On the other hand, as time elapses, an examinee may forget some of the material learned while enrolled in nursing school.

Placing time constraints on an examination also has implications for the IRT model. The speed at which examinees answer questions involves traits such as reading speed. These other facets are inherently different dimensions and, to the extent that these extraneous variables impact the performance of some examinees, violate model assumptions of unidimensionality and may result in erroneous estimates of both ability and item parameters (Oshima, 1994). If a test maker desires to derive measurement information from a time-constrained power test, an investigation into the effects of that time limit are necessary.

Traditional Assessment of Speededness

The traditional view of speededness is as a construct that is dichotomous in nature, i.e., an examination is either speeded or it is not. It must be clearly understood that whenever there is a time limit for an examination, speededness exists. By declaring that an examination with a time limit is unspeeded, a researcher is stating that the time limit does not significantly affect an examinee's speed.

One "rule of thumb" states that on a paper-and-pencil examination, 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 information about items that were answered at an accelerated speed.

Speededness in CAT

On paper-and-pencil examinations, it was difficult to know how much time it took a candidate to answer an item, but with computerized testing, this "examinee-item response time" is obtainable. Combining an examinee's response time across all items can yield that examinee's average item-response time, also called examinee speed or response rate. Conversely, combining an item's response time across all examinees can yield the item's average examinee response time, also called item length.

In traditional paper-and-pencil situations, where a group of examinees is given the same set of items, the amount of time that the an individual is expected to use to complete the examination, the expected completion time, is the sum of all item lengths. In CAT, each examination has a unique set of questions. Hence, the expected completion time is different for every examination and thus every examinee. When a variable-length CAT is used, the expected completion time is not only dependent on which items are used but also on how many.

The NCLEX-RN™ Examination

The NCLEX-RN™ examination is constructed by the National Council of State Boards of Nursing for use by all of the United States jurisdictions as a step in the licensure process for registered nurses. The examinations were converted to computerized adaptive testing (CAT) in April, 1994. Each year, over 120,000 examinees take the NCLEX-RN™ examination. Once an examinee has completed the minimum of number of items (75, consisting of 60 scored and 15 unscored), items are administered until one of the following conditions is met: (1) the examinee's ability estimate is more than 1.65 times the standard error of measurement (SEM) away from the passing standard, (2) the examinee has reached the maximum number of items (265, 250 scored and 15 unscored), or (3) the examinee has reached the four hour and forty-five minute time limit. (Five hours are allowed for the entire examination process, including tutorials and a mandatory 10 minute break. The typical examinee has approximately four hours and forty-five minutes for answering items.)

Approximately 3 percent of the RN examinees are unable to complete their examination within that time frame. However, more than 55% of the examinees are administered only the minimum number of items. These are examinees who perform considerably above or below the passing standard, and whose ability is more than 1.65 times the SEM away from the passing standard when the minimum number of items is completed.

Seldom does an examinee run out of time (ROOT) before completing the minimum number of items. Therefore, only those who must take the longer examinations are at risk of running out of time. To determine how many examinees would have been at risk of running out of time if they had been required to take the maximum-length examination, examinees' response rate to the items they saw was extrapolated into a 265-item examination. This revealed that, had these early finishing examinees been required to take the maximum number of items and had they maintained a similar speed, over one-third of them would have needed more than the maximum time to complete their examination and, therefore, would have run out of time.

Methodology

Data

Basic sample: All examinees (86,005) taking the NCLEX-RN™ between April 1, 1996 and September 30, 1996 comprised the Basic sample. The total examination time and the number of items taken by these examinees was be used to determine examinee speed. All subsequent examinee samples are drawn from this group of examinees.

Item sample: The items (1,543) that comprised the operational pool from April 1, 1996 to September 30, 1996 made up the Item sample. The operational pool is the bank of items that are 'actively' used to test examinees taking the NCLEX™ during the given time frame.

Intra-examination sample: Some of the examinees who used most of the available time may have rushed to complete the remaining items in the allotted time. These are the people who were most likely to be affected by the examination's time limit. The Intra-examination sample included examinees whose examination consisted of 215 or more questions (the maximum is 265) and more than 4 hours and 10 minutes (the maximum is 4 hours and 45 minutes). These values were chosen because they encompass the 5% of the population most likely to have rushed. A random sample of 303 such examinees (of the approximately 4,500 examinees who meet these criteria) comprised the Intra-examination sample. This sample did not include candidates who ran-out-of-time.

Repeater sample: Of the examinees who fail, many take the examination again after the required three-month waiting period. They are called repeaters. There were approximately 650 examinees who ROOTed and failed in the April-June Quarter. This is the group of examinees who would have had been able to retake the examination before September 30, 1996, the final date that the study's item pool was operational. Of the 650, 126 first-time examinees came back for a second attempt before September 30, 1996. These examinees comprised the Repeater sample.

Procedure

This research explored three different methods of measuring speededness: (1) examinee response rate, (2) examinee response rate increase near expiration, and (3) examinee response rate increase across repeated examinations. The first method, examinee response rate, is grounded in traditional criterion adapted directly from the Swineford rule. The last two methods, examinee response rate increase near expiration and examinee response rate increase across repeated examinations, sought to determine whether an examinee spent less time on an item or set of items than expected based on previous performance.

Examinee response rate: The total time that each Basic-sample examinee utilized was divided by the number of items completed. This is an estimate of the examinee's response rate. The mean, 80th percentile, and 99th percentile of all examinees' speed, was extrapolated into an expected completion time for a 265-item examination. The mean, the 80th percentile, and the 99th percentile was also multiplied by 199 (75% of 265) rendering an expected completion time for 75% of the items. The three speeds were also used to calculate the number of items that the three types of examinees would be expected to complete in the four hours and forty five minutes of allotted time. It should be noted that these response rate estimates are conservative because they include the time that examinees spent on items which may have been rushed.

Examinee response rate increase within individual examinations: A complete history of examinee-item response times was compiled for each examinee in the Intra-examination sample. Examinees in this sample may have rushed on some of their later items. This was determined by calculating an examinee's relative response rate to a group of items:

$$\bar{S}_c = \frac{\sum_{i=1}^{I}\left(T_{ic} - \bar{T}_i\right)}{I} \qquad (1)$$

where T_{ic} = examinee c response time to item i, T_i = item length for item i (the average time spent on the item by all examinees), I = total number of number of items (or subset of items) taken by examinee c, $\bar{s}_c$ = relative response rate of examinee c.

Each examinee's relative response rate was calculated using information from the first 60 scored items. More than half (55%) of the examinees taking the NCLEX™ examination finish at this point. Many of those examinees who continue to be administered items after this point, consciously change their cognitive strategy and speed up because they believe that they are going to be administered a full length, 265 item examination. This scenario is very beneficial for this study because one can assume, although this assumption is subject to criticism, that examinee response rate for the first 60 scored items serves as an estimate of how long an examinee might use on each question if they were given an infinite amount of time to complete the examination.

For each examinee, $\bar{s}_c$ was also calculated for all after the first 60 as well as for the examinee's last 50 items. If the difference between $\bar{s}_c$ for the first 60 items and $\bar{s}_c$ for the remaining 190 (or last 50) is significantly greater than 0, then we can reject the hypothesis that the examination's time limit did not have an impact on examinees' response rate.

$$\bar{S}_{c(First60)} = \frac{\sum_{i=1}^{60}\left(T_{ic} - \bar{T}_i\right)}{60}$$ (The relative response rate of examinees to the first 60 scored items. This serves as our reference point) (2)

$$\bar{S}_{c(Last50)} = \frac{\sum_{i=I-50}^{I}\left(T_{ic} - \bar{T}_i\right)}{50}$$ (The relative response rate of examinees to the last 50 items) (3)

$$\bar{S}_{c(Remaining190)} = \frac{\sum_{i=61}^{I}\left(T_{ic} - \bar{T}_i\right)}{I-60}$$ (The relative response rate of examinees to all items administered after the minimum length examination) (4)

$$\Delta_{(Last50)} = \bar{S}_{c(First60)} - \bar{S}_{c(Last50)}$$ (5)

$$\Delta_{(Remaining190)} = \bar{S}_{c(First60)} - \bar{S}_{c(Remaining190)}$$ (6)

(The change in response rate from the first 60 items to either the last 50 items or the remaining 190 items)

Examinee response rate increase across repeated examinations taken by the same examinee: Many examinees taking a variable-length CAT for the first time, work at a natural speed for the first set of items because they anticipate that the time limit will be a minimal concern. Upon returning for a second attempt, these examinees may work at speeded rate starting from the very first question. If these examinees maintain this speeded rate consistently throughout the examination, then the Intra-examination procedure, will not be able to detect their speeded behavior. In order to see if repeat examinees work at a speeded rate, $\bar{s}_c$ for all of the items in a single examination was calculated for each examination in the Repeat sample. The difference between $\bar{s}_c$ for the first examination ($\bar{S}_{c1}$) and $\bar{s}_c$ for the second ($\bar{S}_{c2}$) was compared for individual examinees as well as for the entire group.

Results

Examinee response rate: The distribution of examinee response rates can be seen in Figure 1. The mean response rate was 61.8 seconds/item (see Table 1). The 80th percentile rank of response rate was 74.6 seconds/item while the 99th percentile was 114.8 seconds/item. Thus, we would expect the average examinee to finish a 265 item examination in 4 hours and 33 minutes, 80 percent of the examinees to finish in 5 hours and 30 minutes, and 99% of the examinees to finish in 8 hours and 27 minutes. We would also expect the average examinee to finish a 199 item examination in 3 hours and 25 minutes, 80% of the examinees to finish the examination in 4 hours and 7 minutes, and 99% of the examinees to finish in 6 hours and 21 minutes. Conversely, in the NCLEX™ examination's four hour and forty five minute time limit, the average examinee would be expected to answer 277 items, 80% of the examinees would answer 229 items, and 99% of the examinees would answer 149 items. We would expect 63% of the examinees to finish a 265 item examination in 4 hours and 45 minutes. We would also expect 91% of the examinees to finish 75% of the 265 item examination.

Figure 1. Distribution of mean response rate for all examinees in the Basic sample (N=86,005)

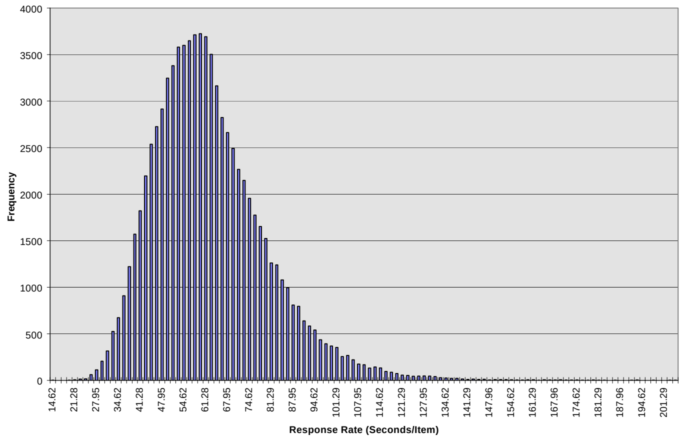


Table 1. Examinee Response Rate for the Basic Sample (N=86,005)

	Individual Item *(Seconds/Item)*	Expected Completion Time for a Hypothetical 199 Item Examination *(Hours:Minutes)*	Expected Completion Time for a Hypothetical 265 Item Examination *(Hours:Minutes)*	Expected # of Items completed in 4:45 Time Limit *(Items)*
Fastest Examinee	14.7	0:49	1:05	1163
Median Examinee	59.3	3:17	4:22	288
Mean Examinee	61.8	3:25	4:33	277
80th Percentile Examinee	74.6	4:07	5:29	229
99th Percentile Examinee	114.8	6:21	8:27	149
Slowest Examinee	228.0	12:36	16:47	75

Examinee response rate increase within individual examinations: The mean of $\Delta_{(Remaining190)}$ was 12.5 seconds/item and the mean of $\Delta_{(Last\ 50)}$ was 18.6 seconds/item (see Table 2). Both of these means were significantly different than 0, $\underline{t}$ (303) = 19.7 and 21.7 respectively, p < .05, thus supporting the notion that the sample of examinees did increase their response rate after the minimum number of items had been completed. In fact, 88% of the sample worked faster on the remaining 190 and 91% worked faster on the last 50. The correlation between $\overline{S}_{c(First60)}$ and $\Delta_{(Remaining190)}$, $\Delta_{(Last\ 50)}$ was high (.86 and .92 respectively) indicating that slow examinees sped up while those who were not slow did not speed up. This relationship was probed further. The $\overline{S}_{c(First60)}$ was found to be a significant predictor of both $\Delta_{(Remaining190)}$, $\underline{F}(1,301)$ = 1617, p < .01 and $\Delta_{(Last\ 50)}$, $\underline{F}(1,301)$ = 845, p < .01.

Table 2. The Change in Response Rate for examinees in the Intra-examination Sample (in seconds/item)

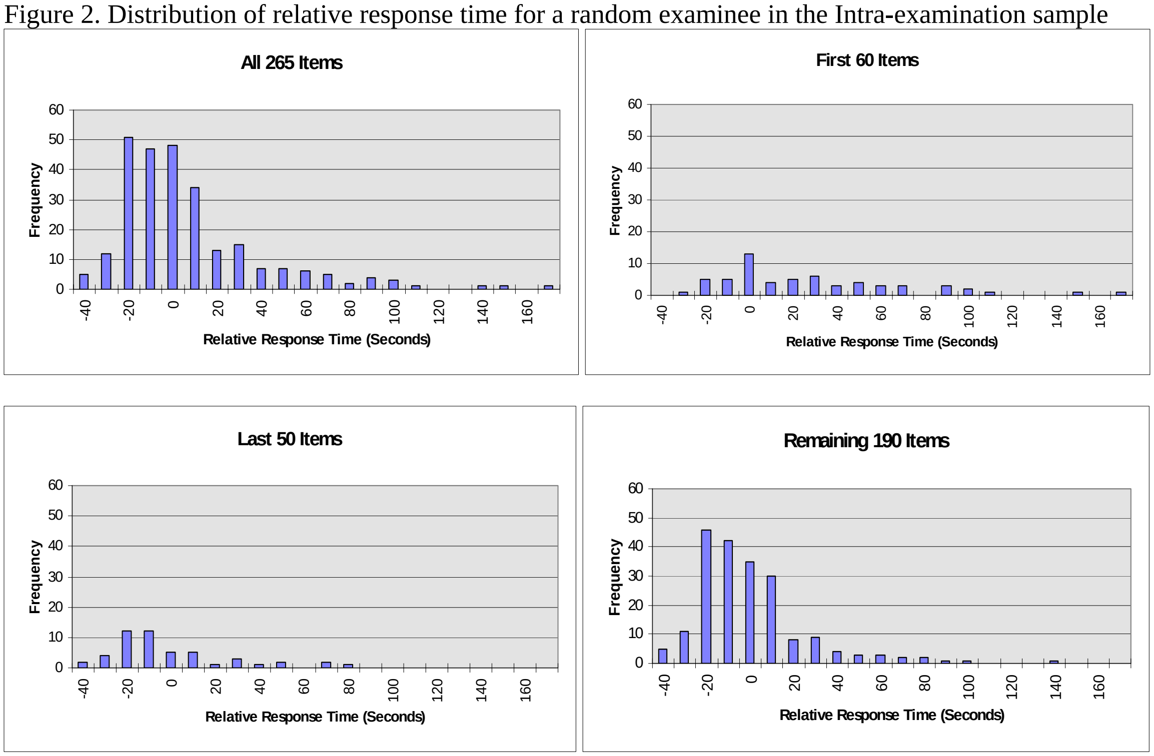
	Mean	Std Dev	min	max
$\Delta_{(Remaining190)}$	12.5	11.1	-14.0	62.2
$\Delta_{(Last\ 50)}$	18.6	14.9	-25	91.4

In an effort to determine whether an individual examinee significantly sped up, the distribution of an examinee's relative response times to the first 60 items was compared to the distribution of the same examinee's relative response times for both the remaining 190 and the last 50 items. A Saitherwaite's t-test was used to compare the first distribution to the latter two distributions. Sixty-six percent of the examinees in the sample had significantly different (α=.05) relative response times for the remaining 190 items while 77% of the examinees had significantly different (α=.05) relative response times for the last 50 items (see Table 3). In using the more conservative of these two results to generalize to the population of examinees in the Basic sample, one can estimate that at least 4,308 examinees (5%) were significantly affected by the time limit (4,308 examinees = 2,231 ROOTs + 66% of the 3,011 examinees who took more than 215 items and used more than 4 hours and 10 minutes). It should be noted that the distributions of relative response times for individual examinees are not normal (see Figure 2). These distributions tend to be positively skewed and to have outliers. These outliers are items on which examinees spent extraordinary amounts of time. Although the distributions in Figure 2 indicate that these extraordinarily long items occurred more often in the first 60 items than in the latter parts of the examination, it is unclear if this is a universal phenomena.

Table 3. The Percentage of Examinees in the Intra-examination Sample who Sped Up

	Percent of Examinees who sped up	Percent of Examinees who significantly sped up
on the Remaining 190 Items	88%	66%
on the Last 50 Items	91%	77%

Figure 2. Distribution of relative response time for a random examinee in the Intra-examination sample



Examinee response rate increase across repeated examinations taken by the same examinee: There were 126 examinees who ran-out-of-time on their first administration and came back for a second attempt while the same item pool was in use. For this group, the mean decrease in relative response rate from the first administration to the second was 8.7 seconds/item (see Table 4). This difference was significant as measured by a paired comparison t-test, $\underline{t}$ (125) = 6.83, $p < .01$. Ninety-seven examinees (77%) answered questions at a quicker pace on the second administration. A Saitherwaite's t-test revealed that 71 examinees (56%) answered questions at a significantly quicker pace (α=.05). Only 12 (10%) examinees ran-out-of-time again on their second attempt. One hundred and four examinees (83%) passed on their second attempt. The correlation between $\overline{S}_{c1}$ and $\overline{S}_{c1} - \overline{S}_{c2}$ was .53. This relationship was significant, $\underline{F}(1,124) = 51$, $p < .01$, thus indicating that slower examinees tended to speed up more than faster examinees. A chi square test could not ascertain a relationship between speeding up and passing.

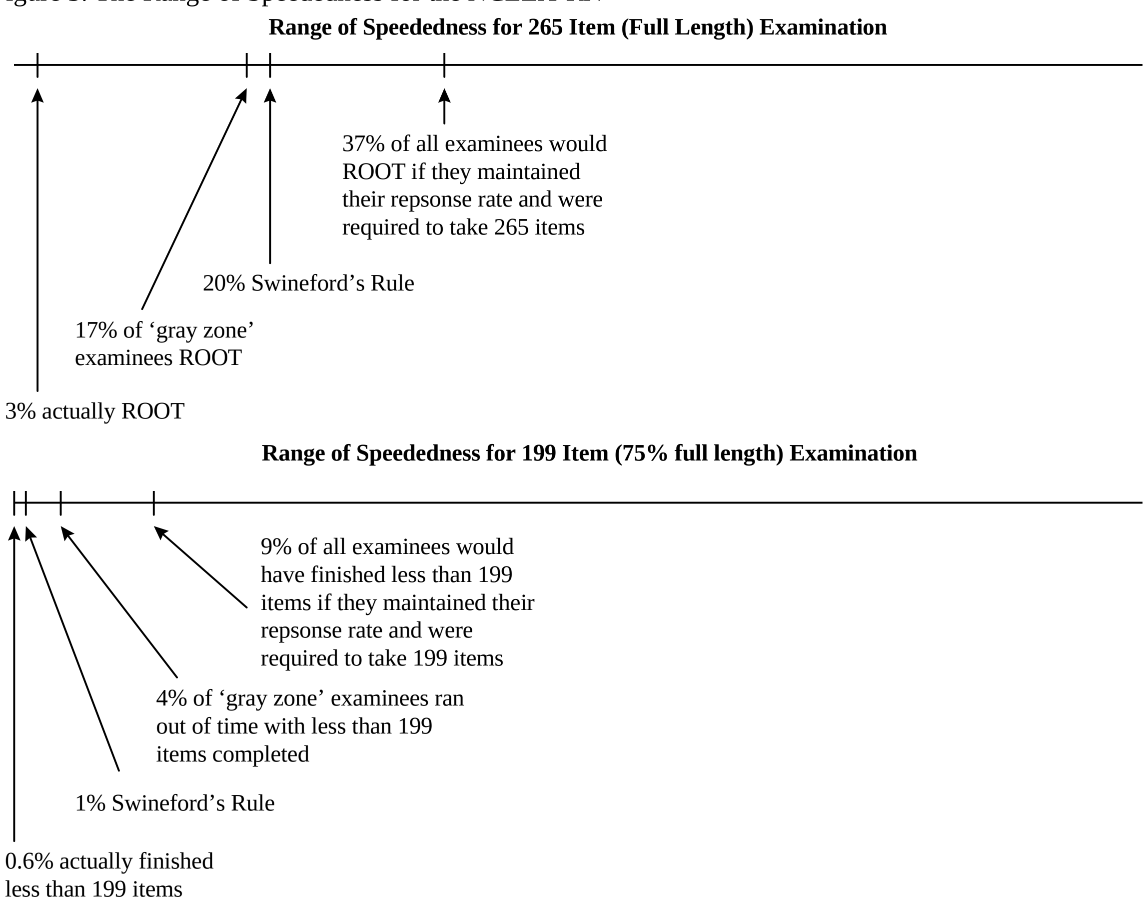
Table 4. Descriptive Statistics for Repeater Sample (in Seconds/Item)

	Mean	Std Dev	min	max
$\bar{S}_{c1}$	20.3	14.7	6.2	97.7
$\bar{S}_{c2}$	11.6	13.9	-20.9	87.1
$\bar{S}_{c1} - \bar{S}_{c2}$	8.7	14.3	-34.2	48.3

Discussion

The first analysis, examinee response rate, revealed that the four hour and forty-five minute NCLEX™ as a 265 item fixed-length CAT is speeded. Of course, the NCLEX™ is a variable-length CAT and not all examinees are required to take all 265 items. Therefore, it might be better to approach this as a range of speededness, taking the most liberal estimate and the most conservative estimate as endpoints. On one hand, only 3% of the examinees actually ran-out-of-time, but 37% would have run-out-of-time if they had maintained their response rate and were required to take the maximum number of items. Another piece of information, that might help center this range, is the proportion of 'gray zone' examinees, those examinees whose ability is in the area requiring the maximum number of items, who ran out of time. For the NCLEX™, 17% of the 'gray zone' examinees ran-out-of-time. Looking at the other half of the Swineford rule, 0.6% of the examinees ran-out-of-time with less than 199 items (75% of a full length examination); 9% of the examinees would have completed less than 199 items had they maintained their response rate and been required to take at least that many items. Finally, 4% of the 'gray zone' examinees ran-out-of-time before they had answered 199 items. Figure 3 illustrates this more clearly.

Figure 3. The Range of Speededness for the NCLEX-RN™



One policy implication that might be taken from this analysis is the implementation of two separate time limits for variable-length CATs. One time limit could be established for the minimum length portion of the examination while another time limit could be established for the rest of the examination.

The results from the first analysis still leave one unanswered question that is essential in understanding speededness. That question is: Under a given time limit, how much faster do examinees' work than they would if they were given an infinite amount of time? The second analysis addressed this question. This analysis capitalized on the variable-length nature of the NCLEX™ which, for first-time examinees, facilitates a slower paced testing strategy for the first 75 items. It was presumed that upon reaching the 76th item, many examinees change strategy and speed up so that they will have enough time to answer the remaining items. This presumption was confirmed for two-thirds of the examinees in the Intra-examination sample. These examinees sped up significantly, thus indicating that the time limit did have an effect on the rate at which they answered items. This analysis estimated that at least 5% of the examinee population was significantly affected by the time limit. Still, the sample represented only the portion of the population that was most inclined to have sped up. Further analysis will need to be done on all non-minimum length examinations to see just how much of the population was significantly affected. This percentage will be one estimate of the degree of speededness in the examination. There are other possible estimates of speededness that could be made from the relative response time information. For example, one could weight the number of examinees that were significantly affected by the degree to which these examinees were affected (Schnipke, 1997).

Before going any further, a closer look needs to be taken at outliers in item response time. These outliers are items on which examinees spent extraordinary amounts of time. If a rationale for the elimination of outliers in the analysis of speededness can be developed, then the distribution of relative response times for an examinee may reach normality. This would allow for standardization of relative response time and provide more robust assessments of increases in response rate.

Even with all of the merits of the second analysis, it will never be able to detect those examinees who worked at a speeded rate for the entire examination. The third analysis aimed to locate some of these individuals. It revealed that as a group, repeaters worked significantly faster on their second attempt. The analysis also revealed that about half of the repeat examinees significantly sped up. Still, it is unclear as to whether this increase in speed was due to a change in cognitive strategy or outside factors such as an increase in ability.

Conclusion

There are two aspects of this study that make it unusual. The first is that this study approaches speededness as a construct entirely independent of any impact on ability estimates. To restate, the definition of speededness is the degree to which the amount of time allowed for test administration affects the rate at which examinees answer items. The second unusual aspect of this study is that speededness is approached as a continuous construct rather than a dichotomy.

A next logical step in research is to create a measure of speededness based on a probabilistic Item Response Theory model. This model can be created and calibrated for a variable-length CAT where response times to items taken under less limiting time constraints is easily available. Due to the nature of item selection in CAT, calibration of all items in a given item pool will be possible. Once the items are calibrated, then changes in examinee response time can be more accurately and universally applied. Once this measure is developed, then it can be used to assess the impact of speededness on examinee performance.

In a perfect world, examinees would be given an infinite amount of time to complete a power test making speededness an irrelevant issue. However, in the real world, a determination of an appropriate time length is necessary. Often times, this decision is made based on financial and administrative concerns, rather than psychometric considerations. This research utilizes new information available only in computerized testing, to measure speededness. Beyond that the study relates this information back to already established ‘rules of thumb’ in very straightforward ways. It is hoped that these methods will help ground examination time length decisions for variable-length CATs in more concrete psychometric rationale.

References

Gorham, J. L. and Bontempo, B. D. (1996). *Repeater patterns on NCLEX ™ using CAT versus NCLEX ™ using paper-and-pencil testing*. Paper presented at the Annual Meeting of the American Educational Research Association, New York.

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

Julian, E. R. and Bontempo, B. D. (April, 1996). *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.

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

Schnipke, D. L. (March, 1995). *Assessing speededness in computer based test using item response times*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Schnipke, D. L. and Scrams, D. J. (June, 1996). *Modeling Response Times in Testing with a Two-State Mixture Model: A New Approach to Detect Speededness*. Paper presented at the Annual Meeting of the Psychometric Society, Banff, Alberta, Canada.

Schnipke, D. L. and Scrams, D. J. (1997). Modeling Response Times in Testing with a Two-State Mixture Model: A New Method of Measuring Speededness. In Press, *Journal of Educational Measurement*.

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