Investigation into Decision Rules for NCLEX™ Candidates Who Run Out of Time

Ellen R. Julian, PhD
Brian D. Bontempo

The National Council of State Boards of Nursing
April 1996

Paper presented at April 1996 AERA annual meeting, New York.

Investigation into Decision Rules for NCLEX™ Candidates Who Run Out of Time

Background

The NCLEX-RN™ and NCLEX-PN™ are 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 and practical/vocational nurses. The examinations were converted to computerized adaptive testing (CAT) in April, 1994. Each year, over 120,000 candidates take the NCLEX-RN and over 60,000 candidates take the NCLEX-PN. The maximum test length for the RN examination is 265 items (250 scored and 15 unscored), and 205 items for the PN (180 scored and 25 unscored) (unless otherwise noted, all references to numbers of items will be to the number of scored items only). Both examinations have five-hour time limits. Approximately 3 percent of the RN candidates do not complete their examination within that timeframe. Fewer PN candidates run out of time because of the shorter test length. This research will focus on the RN candidates.

Candidates taking the NCLEX-RN are administered items until (1) the candidate's ability estimate is more than 1.65 standard errors of measurement (SEMs) away from the passing standard, (2) the candidate has reached the maximum number of items, or (3) the candidate has reached the five-hour time limit. Once an ability estimate has reached a point where it is unlikely (to a specified degree, 1.65 SEMs) that this estimate would change to the other side of the passing standard upon retesting, the examination is terminated and a simple decision rule is employed: was the final ability estimate above the passing standard? If so, they pass. If not, they fail. In the Rasch model, the statistical precision or SEM is a direct function of the number of items taken and how well those items are targeted to the candidate's ability. In the NCLEX, each item is targeted to the candidate's ability in an identical fashion making the statistical precision dependent, for the most part, on the number of items taken. At the maximum number of items, the precision of the final ability estimate has been deemed acceptable and is approximately equal to that of the former linear paper-and-pencil examination. For candidates taking 250 items, the same decision rule is applied.

But when a candidate runs out of time, they have not achieved an ability estimate that is 1.65 SEMs away from the passing standard nor have they answered the maximum number of items. This implies that the precision of their final ability estimate is less than the acceptable precision. By definition, they are within a 'gray zone.' Therefore, a different decision rule, called a ROOT (ran-out-of-time) rule, is applied to these candidates.

ROOT Rules

Because a primary goal of the licensure examination process is to protect the public, it can be argued that candidates should fail when it is not clear that they are competent, in which case all ROOTs would fail. However, a secondary objective of the licensure examination process is to pass candidates who have demonstrated sufficient competence. Some of the candidates who run out of time have performed above the passing standard, and their true competence level appears to be above passing, albeit close to it. 'ROOT-rules' are designed to allow these candidates to pass. No ROOT rule would pass a candidate whose final ability estimate is below passing; ROOT rules are only applied to candidates whose final ability estimate is above passing.

A good ROOT rule would provide a smooth transition in the decision process from candidates with complete examinations to those who ROOT. Given two identical above-passing ability estimates, the

ability estimate derived from a higher number of items has a higher statistical probability of being the "true" ability of the candidate, and that candidate of being a "true" passer. Therefore, a candidate who has answered 250 items should have almost the same probability of passing at 249 items and a much higher probability than at 60 items. (No one would be allowed to pass, under any ROOT rule, who had not completed at least a minimum-length examination of 60 items.)

The mere existence of a ROOT rule is an acknowledgment that, in the absence of a satisfactory amount of statistical evidence, the candidate's performance must provide other types of assurance that their "true" ability is above passing. One goal of this study is to find a ROOT rule that provides all candidates with the same "true" ability with the same probability of passing, regardless of how long they spend answering the average item. Simply responding slowly to questions does not constitute a lack of nursing ability. There are many passing candidates whose examination ends at the minimum 60 items who answer questions so slowly that they would have run out of time if longer tests were required. These candidates should not be penalized for answering slowly. They pass because they have performed well enough to demonstrate that there is only a remote possibility that their "true" ability is below passing.

An additional source of evidence that the final ability estimate is an accurate estimate of the candidate's "true" ability is the consistency of their performance. A good ROOT rule should distinguish among candidates with similar (above passing) final ability estimates whose entire examination performance record would suggest that one candidate's "true" ability is above passing, while the other one's is below. This goal can be translated as requiring that the final (above passing) ability estimate be an accurate reflection of the entire examination performance. In the absence of a useful person fit indicator for CAT (Bradlow and Zeger, 1995), other methods of summarizing performance must be used. This goal requires utilization of the information contained in a candidate's examination record, not just in a final ability estimate.

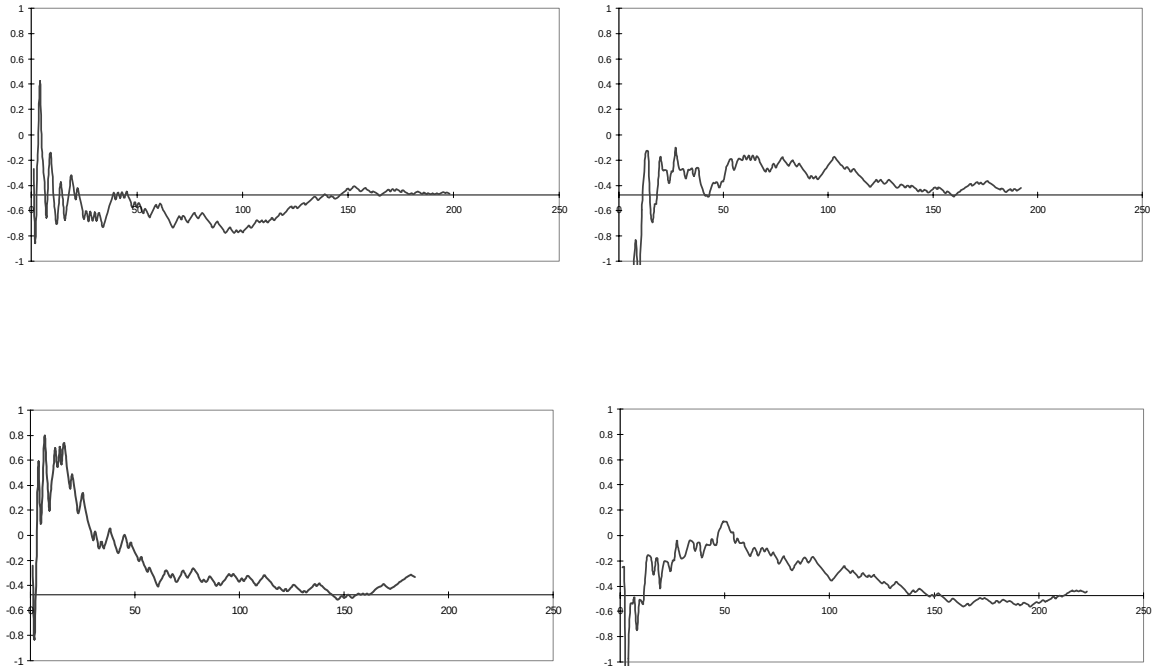
Current 'ROOT Rule'

After each item the computer reestimates the candidate's competence level, using the information from all of the items answered to that point. The current ROOT-rule, referred to as the Last 60 rule, passes a candidate if each of the last 60 ability estimates made before the candidate ran out of time was above passing. If a candidate runs out of time and, at any point during the last 60 items, the estimate of their competence level was below passing, they do not pass.

The Last 60 rule focuses on the criterion of performance consistency. For candidates whose performance has remained consistent across the full spectrum of examination content (covered approximately three times in 60 items), we have greater confidence that their final estimate reflects their true competence level, and would not change dramatically with further testing. Candidates who have progressed further into the examination have a modest advantage over those who ROOT after fewer items, because the change in the ability estimates after each item answered is smaller later in the test, resulting in more stable ability estimates.

Approximately 57% of those who take the maximum-length examination pass. Only 36% of all ROOTs currently pass. Staff who work with ROOTs report a perception that the current rule is doing a thorough job of failing all candidates whose "true" ability seems to be below the passing standard. However, they are concerned that the Last 60 rule may be failing some who should pass (see Figure 1 for examples of ROOT candidates who failed because of the Last 60 rule).

Figure 1. Examples of ROOT Candidates Who Failed Because of the Last 60 Rule



Alternative Rules

Final ability estimate. For comparison purposes, decisions simply based on the final ability estimate are included in tables and discussions. This rule essentially is no rule; it entails treating ROOTs as if they had reached the maximum number of questions. Because the final ability estimate is our best estimate of the candidate's competence, it provides an important anchor for comparison of other rules' results.

Last XX rule. Informal investigations during the NCLEX Beta Test (Zara, 1994; Way, 1994) made it apparent that a simple change to the number in a "Last XX" rule would not accomplish the goal of failing candidates who apparently should fail and passing those who apparently should pass. Because of the candidates whose only dip below passing is at the very end of the test, the XX in the new rule would have to be a very small number, with the consequence of passing a large number of candidates whose ability had been below passing throughout most of the examination. This possibility will not be investigated further.

Length-Adjusted Passing Standard (LAPS). When a candidate completes a maximum-length examination, the pass/fail decision is based solely on the final ability estimate, because the level of statistical precision achieved by that length examination has been judged acceptable. For shorter examinations, the precision is less. The LAPS rule would adjust the ability estimate required for a ROOT to pass upwards by the difference between the desired SEM (the SEM after 250 items) and the candidate's final SEM. Therefore, for a ROOT at 249 items, only a slightly higher ability estimate would be required in order to pass, whereas, for a ROOT at 100 items, a substantially higher ability estimate would be required.

This rule incorporates both the goal of utilizing the statistical precision of final estimates and maintaining a smooth transition from decisions made for complete examinations to those made for ROOTs. LAPS's

disadvantage may be that it allows candidates to pass who have achieved an acceptable final ability estimate but have performed inconsistently.

Indisputable pass. In the course of designing the resolution process for candidates whose tests ends early through no fault of their own, a "rough-and-ready" definition of an "indisputable pass" (IndisPass) has been developed. These are candidates who would have passed, even if they had answered all remaining items incorrectly (IndisPass-wrong). 'All remaining' is defined as the difference between the item number at which their test ended and the maximum length test of 250 items. In essence, the information missing from a candidate's examination is imputed, as incorrectly answered, yielding a complete, maximum-length examination (Little & Rubin, 1987). Only candidates who have completed more than 223 items can pass under this rule.

A (perhaps more realistic) variation on this rule (IndisPass-random) would impute the outcome if all remaining items were answered by choosing one of the four answer options at random (in which case it would be assumed that one-fourth of the items would be answered correctly). This rule could only pass those candidates who completed more than 196 questions.

The Last 60 Rule fails some candidates who would qualify as indisputable passes. These are often candidates who ROOTed with one or two items remaining and whose abilities were almost high enough to have stopped the examination, but who had dipped below passing sometime within the last 60 items. However, not many candidates qualify as indisputable passes, so that many of the candidates who pass with the Last 60 rule would fail under an indisputable pass rule.

Combination rules. A combination rule would incorporate both the final estimate of a candidate's ability, and the consistency of the performance that produced it. The Last 60 rule could be combined with either of the indisputable pass rules, the IndisPass-random or IndisPass-wrong. The candidate would pass if either rule resulted in a pass decision. A combination rule could be communicated this way: "Candidates would pass if they had been consistently above passing, or if they couldn't fall below passing even if they answered every remaining item wrong."

Methodology

Data:
Existing data files contained only ability estimates and SEMs after the last completed item. An array of ability estimates and SEMs from the entire examination was also needed for the candidates being subjected to further analysis. Software constraints prevented acquisition of the actual SEM after each item. However, the expected SEM is a reasonable estimate of the actual because, with the Rasch model, the SEM is dependent on only the number of items administered and how well those items are targeted to the candidate's ability.

Consideration of the indisputable pass definition required an array of the expected change in ability estimates after each additional item was administered, assuming it were perfectly targeted. Preliminary research determined that it was not necessary to have separate arrays for ability increments after a correct response and decrements after an incorrect response.

Samples and Methodology:

- *Basic sample:* All 3954 NCLEX-RN ROOTs from the first year of CAT implementation constitute the basic sample for the descriptive analyses.
    - Each of the decision rules was applied to all of these candidates to determine the pass or fail outcome of their test. Candidates whose outcomes differ depending on the passing rule received special attention.

- *Hypothetical ROOTs sample*: A random sample of 228 candidates who completed maximum-length examinations within the five hours were selected. The range of final ability estimates for this group is therefore restricted to a narrow band around the passing standard. A complete examination history of the calculated ability estimate after each completed item was acquired for each candidate in this group. Their response strings were propagated into multiple hypothetical ROOTs with different test lengths.
  - The proposed ROOT rules were applied to each hypothetical ROOT and the outcomes compared with the candidate's actual pass/fail outcome after 250 items.

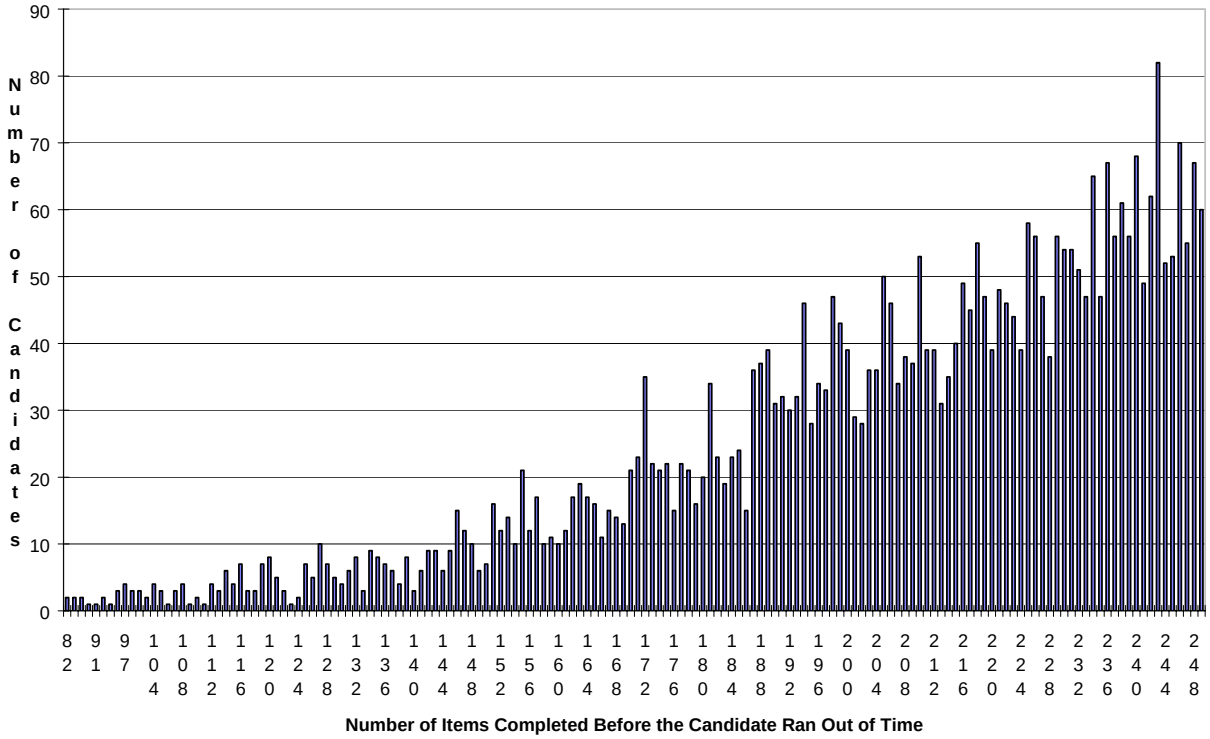
- *Policy-capturing sample*: A subset of 56 ROOT candidates with final ability estimates above passing was selected at random, and their complete examination histories obtained. Maps of their performance histories, such as those used in the resolution of candidate inquiries, were prepared.
  - Staff members involved in the ongoing evaluation of candidate maps were asked to sort the maps into five piles: "Definitely ought to pass, Probably ought to pass, Who knows?, Probably ought to fail, Definitely ought to fail." Consistency among the judges was evaluated. All of the proposed ROOT rules were applied to each candidate's examination, and the outcomes were compared to the decision made by each of the judges.

Results

Basic Sample. The basic sample consisted of the 3,854 candidates who ran out of time during the first year that the NCLEX-RN was administered using CAT (April 1, 1994 - March 31, 1995). Because it is possible for candidates to repeat the examination three months after failing, it is possible that the same candidate appears in this group more than once. Of the total basic sample, 2,834 candidates were taking the examination for the first time. The total group had an average final ability estimate of -0.4621. The passing standard applied during this time period was -0.4766. The highest final ability estimate achieved by any ROOT was -0.1773, 2.5 SDs (.1128) above the group's mean. This candidate took 136 items and passed. The lowest was -0.8437, 3.4 SDs below, and was achieved by a candidate who completed 108 items.

The maximum number of items taken by any ROOT was 249, the minimum was 82, and the mean was 207. Half of the ROOTs answered more than 215 items, and 25% answered more than 235. Figure 2 shows the distribution of the number of items completed before the candidates ran out of time.

Figure 2. Frequency Distribution of the Basic Sample (n=3854)



Of the basic sample, 56 percent had a final ability estimate above passing, but as a result of the application of the Last 60 rule, only 36 percent passed. The highest final ability estimate of a ROOT who failed was -0.2391, and was achieved by a candidate who completed 129 items. Table 1 shows the passing percentage resulting from the application of each proposed ROOT rule and if only the final ability estimate were considered.

Table 1. ROOT Passing Rate for Each Proposed ROOT Rule

Rule	***Passing Rate***
Last 60	**36%**
Final Ability Estimate	56%
LAPS	49%
IndisPass - wrong	10%
IndisPass - random	18%
Comb: Last 60 or IndisPass - wrong	37%
Comb: Last 60 or IndisPass - random	39%

Hypothetical ROOTs. The random sample of 228 maximum-length examination candidates had an average final ability estimate of -0.4712 (lower than the ROOT basic sample), with a maximum of -0.2695, a minimum of -0.6892 and a standard deviation of 0.10. Candidates with ability estimates higher or lower than these values were far enough from the passing standard that the SEM stopping rule ended their test. Of this group, 51 percent had final ability estimates above the passing standard, and therefore,

51 percent passed. These candidates spent an average of 3 hours and 47 minutes taking the examination. For this group, the slowest candidate completed the examination in 4 hours and 51 minutes, and the fastest candidate finished in 2 hours and 15 minutes.

Each candidate's record was artificially terminated after 60, 100, 150, 200, 210, 220, 230, and 240 items. To each of these hypothetical ROOTs, the following ROOT rules were applied: Last 60, LAPS, IndisPass-wrong, IndisPass-random, Comb: Last 60 & IndisPass-wrong, or Comb: Last 60 & IndisPass-random. Table 2 shows the percent passing following the application of each rule to each hypothetical test length. Table 3 shows the number of disagreements between each of the proposed ROOT rules and the 228 candidates' actual pass/fail outcomes after all 250 items, and Table 4 shows the number of false positive and false negative disagreements.

Table 2. Percent Passing at Different Hypothetical Test Lengths Using Proposed ROOT Rules

	Number of Items									
Rule	**60**	**100**	**150**	**200**	**210**	**220**	**230**	**240**	**249**	**250**
Final Ability Estimate	59%	55%	57%	60%	57%	57%	54%	54%	51%	51%
LAPS	14%	23%	36%	50%	50%	52%	51%	52%	51%	51%
Last 60	11%	21%	25%	32%	32%	33%	33%	34%	36%	38%
IndisPass - Random	0%	0%	0%	1%	2%	16%	27%	42%	49%	51%
IndisPass - Wrong	0%	0%	0%	0%	0%	0%	5%	29%	49%	51%
Last 60 & IndisPass - Random	11%	21%	25%	32%	32%	34%	39%	45%	49%	51%
Last 60 & IndisPass - Wrong	11%	21%	25%	32%	32%	33%	33%	40%	49%	51%

Table 3. Number of Disagreements (out of 228) at Different Hypothetical Test Lengths Between Proposed Rules and Actual Outcomes

Rule	**60**	**100**	**150**	**200**	**210**	**220**	**230**	**240**
Final Ability Estimate	102	91	72	38	30	33	20	13
LAPS	117	102	71	40	36	39	25	14
Last 60	120	106	87	65	61	92	50	42
IndisPass - Random	116	116	116	114	111	76	55	21
IndisPass - Wrong	116	116	116	116	116	155	104	51
Last 60 & IndisPass-R	120	106	87	65	61	54	36	18
Last 60 & IndisPass-W	120	106	87	65	61	80	50	30

Table 4. Number of False Positive and False Negative Disagreements at Different Hypothetical Test Lengths Between Proposed Rules and Actual Outcomes

	60		100		150		200		210		220		230		240		250	
Rule	**F+**	**F-**	**F+**	**F-**	**F+**	**F-**	**F+**	**F-**	**F+**	**F-**	**F+**	**F-**	**F+**	**F-**	**F+**	**F-**	**Fail**	**Pass**
Final Ability Estimate	60	42	50	41	43	29	29	9	22	8	21	8	14	6	10	3	**112**	**116**
LAPS	16	101	19	83	19	52	19	21	17	19	13	11	13	12	8	6		
Last 60	14	106	19	87	14	73	11	54	9	52	6	48	4	46	2	40		
IndisPass - Random	0	116	0	116	0	116	0	114	0	111	0	79	0	55	0	21		
IndisPass - Wrong	0	116	0	116	0	116	0	116	0	116	0	116	0	104	0	51		
Last 60 & IndisPass-R	14	106	19	87	14	73	11	54	9	52	6	45	4	32	2	16		
Last 60 & IndisPass-W	14	106	19	87	14	73	11	54	9	52	6	48	4	46	2	28		

The final two columns of Table 4 show the number of candidates who actually passed and failed at the end of their maximum-length examination. These values provide a point of reference for the number of

false positive and false negative errors (e.g., of the 112 candidates who had final ability estimates below passing after 250 items, 60 would have passed if the final ability estimate after 60 items had been used). Combining the data from Tables 3 and 4 shows that at the minimum possible test length, 60 items, all of the rules make about the same number of mistakes, but many more false negatives (failing people who eventually did pass). At all test lengths, the "final" ability estimate is the best predictor of the actual outcome, but it also has the highest rate of false positives, which are the more serious of the two types of error in a licensure situation. Table 5 shows that the average ability estimate falls as the test progresses, explaining why the use of the final ability estimate's false positives consistently exceed the false negatives (i.e., earlier ability estimates are consistently overpredicting the ability estimate after 250 items).

Table 5. Average Ability Estimate at Each Hypothetical Test Length

	60	100	150	200	210	220	230	240	249	250
Average Final Ability Estimate	-0.4447	-0.4644	-0.4617	-0.4630	-0.4660	-0.4662	-0.4646	-0.4669	-0.4711	-0.4712

Policy-capturing Sample. The sample of 56 ROOTs all had final ability estimates above passing. This is the group for whom the choice of ROOT rules makes the most difference. The average final ability estimate for the group was -0.3912. This group's number of items taken ranged from 147 to 246, with a median of 214 and an average of 208. Table 6 shows their outcomes with the application of each proposed ROOT rule. The Last 60 rule was the rule actually applied to their examination and the 26 in this cell represents the number of these candidates who actually passed this attempt at the NCLEX (46%).

Table 6. Number in Policy-capturing Sample Passing and Failing Under Each Proposed ROOT Rule

Final Theta		Last60		LAPS		IndisPass - w		IndisPass -r		Comb - w		Comb-r	
Pass	*Fail*	*Pass*	*Fail*	*Pass*	*Fail*	*Pass*	*Fail*	*Pass*	*Fail*	*Pass*	*Fail*	*Pass*	*Fail*
56	0	26	30	50	6	5	51	9	47	27	29	29	27

Five judges recorded their perception of the true ability that the 56 candidates, based on their performance maps, were passing, failing, or uncertain. As shown in Table 7, the judges varied both in severity/leniency and in their use of the category representing uncertainty. The judge who would have passed the most candidates (Judge #4) was also the one least likely to be uncertain of a decision. In contrast, judge #1 passed and failed the lowest number of candidates.

Table 7. Judges' Pass/Fail Decisions on the Policy-capturing Sample

	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5
# Pass	28	28	30	49	36
# ?	24	19	15	2	14
# Fail	4	9	11	5	6

If a voting process were used, where the undecided judges abstained from the vote, 45 candidates would have passed. The vote would have been tied on two candidates - one, because every judge cast an undecided vote, and one, where one judge voted pass, another voted fail and the rest were undecided. The judges' pass or fail vote was unanimous on 25 candidates.

The judges' pass/fail vote (coded 1/0) correlated -.05 with the number of items taken and .35 with the final ability estimate. An observed positive correlation between number of items taken and the final ability estimate could be a function of the constraints placed upon the sample: (1) as the test progresses, candidates with higher (and lower) ability estimates are terminated and do not have the opportunity to ROOT, and (2) selecting only ROOTs with passing ability estimates eliminates the opportunity for the low scoring ROOTs to provide symmetry throughout the range of item counts.

The vote agreed with the Last 60 rule in 35 cases. One of the disagreements was a candidate whom the judges had unanimously passed. The Last 60 rule passed one and failed one of the candidates on which the judges had a tied vote. On 10 of the disagreements, the most common response selected by judges was uncertainty.

The agreement of the judges' vote with the proposed ROOT rules is shown in Table 8.

Table 8. Number of Agreements Between Judges' Vote and Outcome of Each Proposed Passing Rule

	Final Ability*	Last 60	LAPS	IndisPass - wrong	IndisPass - random	Comb - wrong	Comb - random
# Agreements	45	35	47	13	17	35	37

* All of the Policy-capturing sample had final ability estimates above passing. Therefore, the number of agreements here is simply the number of candidates that the judges voted to pass.

Only the LAPS rule had a higher rate of agreement with the judges' vote than did the final ability estimate. Among the disagreements, only LAPS (and the final ability estimate) passed any candidates that the judges failed. It passed five - on two of whom the judges were in unanimous agreement that they should not pass.

Discussion

The decision of the judges to fail nine of the 54 Policy-capturing sample, and their lack of agreement on the fate of the others, illustrates the need for a ROOT rule. If only the final ability estimate were used, all of these ROOTs would have passed. However, neither the current Last 60 rule, nor its combinations with the IndisPass rules, approach the final ability's level of agreement with the judges' pass/fail decisions. It and its combinations did, however, avoid passing any candidates that the judges would have failed. This is the overriding concern in a licensure examination program.

Candidates who run out of time with only a few items remaining may benefit from the combination of an IndisPass rule with the Last 60 rule. Only ROOTs who have completed more than 196 questions can qualify as an Indisputable Pass, assuming random responses to all remaining items, and only those who have completed more than 223 can qualify if all remaining items are assumed to be answered incorrectly. The addition of the IndisPass-random to the Last 60 rule allowed an additional three candidates from the Policy-capturing sample to pass, two of whom the judges voted to pass. The other was the candidate on whom three judges were undecided, one voted to pass and one to fail. An additional 110 candidates from the Basic Sample would pass if the IndisPass-random were combined with the Last 60 rule.

A universally applicable rule for making pass/fail decisions for candidates who run out of time on an adaptive examination does not exist. Each of the proposed rules was created to fulfill certain philosophical criteria. In deciding among these rules, the purpose of the examination, the construct it is intended to measure and the cost of incorrect decisions must be incorporated into the theoretical justification for a rule's use. Beyond that, the real world consequences, such as ease of explanation, legal defensibility, and implications for every ROOT, must be considered.

References

Little, R.J.A. and Rubin, D.B. (1987). Statistical analysis with missing data. NY:Wiley.

Bradlow, E.T. and Zeger, L.M. (1995). Identification of Aberrant Response Patterns in the Nurse Licensure Examination. *Manuscript in Preparation.*

Way, W.D. (1994, March). Psychometric Results of the NCLEX™ Beta Test. Paper presented at AERA annual meeting, New Orleans.

Zara, A.R. (1994, March). An overview of the NCLEX/CAT™ Beta Test. Paper presented at AERA annual meeting, New Orleans.