Repeater Patterns on NCLEX™ using CAT versus NCLEX™ using Paper-and-Pencil Testing

Jerry L. Gorham

The Chauncey Group International

Brian D. Bontempo

The National Council of State Boards of Nursing

June 25, 1996

Repeater Patterns on NCLEX™ using CAT versus

NCLEX™ using Paper-and-Pencil Testing

Introduction

The operational implementation of computerized adaptive testing for the National Council of the State Boards of Nursing Licensure Examination (NCLEX-RN™ and NCLEX™-PN) began on April 1, 1994. This implementation was preceded by years of planning, research, and a massive Beta Test effort (Zara, 1992a; 1992b; Way, 1994b; NCLEX/CAT Team, 1994). This transition from traditional paper-and-pencil to CAT for this large-scale, high stakes testing program, has yielded an entirely new range of research questions and has provided hard data for issues of comparability between paper-and-pencil and CAT testing. During the Beta Test period, comparability issues were explored under specific conditions which could isolate the effects of single-day administration compared to traditional two-day administrations, and linearized computerized testing vs. CAT testing (Eignor, et al., 1993; Way, 1994b). Results of the Beta Test effort indicated that the new CAT test was comparable with the traditional paper-and-pencil test. Over the course of almost two years of daily CAT testing, it has become clear that while CAT testing has its own distinct features, these features do not adversely affect the essential psychometric goals of the licensure examination.

The CAT model used for NCLEX is based on an approach described by Weiss and Kingsbury (1984), and Lunz and Bergstrom (1991), among others. The variable-length nature of the model is as follows: for candidates whose theta estimate is close to the pass/fail cutscore, the computer continues to administer additional items. If a candidate's theta estimate is determined to be

sufficiently above or below the pass/fail cutscore, the CAT terminates. Decision rules for termination are based on comparing a confidence interval defined around the candidate's theta estimate with the theta level that defines the pass/fail cutscore.[1] As long as the cutscore theta level is within the confidence interval, an additional item is administered. Once the confidence interval no longer encompasses the pass/fail cut score (provided that at least 60 scored items have been administered), testing is terminated with a pass or fail result. Candidates taking the NCLEX are given a maximum of five hours to complete the test. If a candidate runs out of time before a normal termination is reached, special decision rules are invoked where the candidate's theta estimate for the last 60 items taken are each compared to the pass/fail theta level. If each of the 60 theta estimates is above the pass/fail theta level, a passing decision is returned. However, if one or more of the theta estimates following each of the last 60 items is below the pass/fail theta level, a failing decision is returned.

The focus of this discussion is to document and summarize the results of the past two years of CAT testing for candidates who fail the NCLEX-RN and subsequently repeat the examination. Some of the characteristics of these examinees will be summarized and compared with examinees who failed the NCLEX and repeated the examination under its traditional paper-and-pencil format.[2]

---

[1]This confidence interval is obtained by multiplying the standard error of the candidate's ability estimate by 1.65. The constant 1.65 is used because it results in a one-tailed 95% confidence interval around the candidate's ability estimate.

[2]A more comprehensive account of these issues will appear this fall in a report to the National Council of the State Boards of Nursing - Chauncey Group, Intl. Joint Research Council.

Repeater Population

One relevant research problem pertains to that population of examinees who repeat the NCLEX-RN at least once.[3] In recent years, the repeater population has constituted approximately one-fifth of the overall population, with passing rates at approximately 45%-50% for repeaters, compared to passing rates of 85%-90% for first-time examinees. For the NCLEX-RN, this population is considerable in size, with over 19,000 candidates repeating the CAT one or more times since April 1994. Under the paper-and-pencil exam, there were approximately 44,000 candidates repeating one or more times over a five year period, from July 1989 to February 1994. Although the constitution of the repeater population continues to change, the testing patterns of repeating candidates appear to be fairly consistent.

The conceptual features of a "repeater" population may not be obvious at first glance. This is because generally one speaks in terms of particular administrations (or forms) of an examination rather than sequences of retests across those forms. For example, it is customary to speak of passing rates for an administration of an exam rather than passing rates for all individuals who repeat the exam once. In this sense, the assumptions of test equating and scaling are extremely important, since comparisons between forms would otherwise be meaningless. For repeating examinees, for instance, an ability estimate for a first attempt may be based on two different forms of the exam (which have been equated and scaled). The "test" therefore can be understood as a dynamic but fair psychometric process, rather than a static, event-based administration. In terms of CAT administrations, this feature is even more fundamental since there are virtually as many forms or

[3]Since more data was available for the NCLEX-RN test than for the NCLEX-PN test, this paper will focus on the NCLEX™-RN test only.

administrations as there are examinees

-- each CAT exam is in fact a new "form" or administration. The value of this approach lies in its ability to look across forms and to ensure test stability of a different sort. For licensure and certification testing, a repeater population is necessarily restricted in range as a result of the cutscore which forces the failing candidate to retest. In a strict sense, therefore, the repeater population for a licensure examination represents a restriction of range both in terms of scale and in terms of longitude.

It should also be pointed out that in fact there are as many "repeater" populations as there are attempts at repeating the exam: a population of first-time examinees, first-repeat examinees, second-repeat examinees, and so on. However, inferences based on these subpopulations tend to be unstable simply because the amount of data dwindles quickly across retests.

The focus of this discussion, therefore, will be on detecting general patterns across the CAT data and comparing these patterns to traditional paper-and-pencil repeater patterns.

NCLEX-RN Paper & Pencil

Description of the Data

Repeater data from ten administrations of the traditional NCLEX™-RN paper and pencil examination was sampled beginning with the July 1989 administration and ending with the February 1994 administration. The NCLEX-RN™ examination was chosen over the NCLEX-PN™ examination as a focus for this analysis because more data were available from the RN examination than from the PN examination. Data samples yielded 43,847 persons repeating at least one

examination during this period. Final ability estimates for first-time testers were available for 15,272 of these candidates, and for first repeaters, final ability estimates were available for 21,600 of these candidates. A overall range of one to nine repeats was available for these examinees. Table 1 shows summary statistics of the final ability estimates (final thetas) for these candidates by examination sequence (the term "exam" or "examination" will be used hereafter to refer to the examination sequence for a given repeater candidate, rather than a particular form or administration). The mean theta estimate for these candidates was -0.7418 with a standard deviation of 0.2904. Upon repeating the examination once, these candidates showed an average gain score of 0.2656.

Insert Table 1 here

There were 7,935 candidates taking the examination three times, with a mean theta estimate of -0.6559 for the third testing, and showing an average gain score of 0.0859 over the first testing and loss score of 0.1709 over the second testing. Figure 1 also illustrates a dramatic gain score between the first and second exam, but what appears to be a linear trend of decreasing theta estimates for exams two through five. Part of this increase over the first exam reflects the restricted range for this subgroup since the mean for this group on the first exam is based on failers only. For exams six through ten, theta estimates remain relatively stable, with a mean of -0.7781 and a standard deviation of 0.0081.

Correlations between Testings

Correlations between theta estimates range from 0.577 to 0.785, with a mean correlation of 0.685 between any two testings. The most relevant intercorrelations for repeater data are perhaps correlations between consecutive testings. Figure 2 (P & P only) shows that correlations between consecutive theta estimates for exams one through ten are relatively high and consistent.

Passing Rates

Passing rates for the paper-and-pencil test are presented in Table 2.

Insert Table 2 here

Approximately 52.9% of the examinees repeating the examination once passed, and 36.9% of examinees repeating the exam twice passed. The last column of Table 2 shows the cumulative percent of examinees passing by attempt. Upon repeating the examination once, 52.9% of the total number of repeaters have passed, and upon repeating a second time, 66.8% of the total number of repeaters have passed. On a third repeat, 73.0% of repeaters have passed, but after the third repeat, the cumulative percentage of examinees passing upon each subsequent repeat begins to diminish. By the ninth repeat (exam ten), only 77.2% of the total number of examinees testing have passed. This pattern is illustrated in Figure 1, in which examinees repeating the examination three or fewer times appear to show noticeable improvement after each repeat, but examinees repeating the exam four or more times appear on the whole to reach a ceiling effect at attempt number five (the fourth repeat).

NCLEX-RN Computerized Adaptive Test

Description of the Data

Repeater data from the NCLEX-RN CAT were sampled beginning April 1, 1994 through March 4, 1996. During this period there were 19,119 candidates who repeated the NCLEX-RN at least one time. These figures represent the number of unique candidates repeating rather than the total number of repeats overall. Final ability estimates were available for each of these candidates, since these estimates are computed as part of the CAT examination itself. There were a maximum of seven exams (six repeats) which were available from the data during the period. Final ability estimates for first-time testers in the repeater population yielded an overall mean of -0.8375 and a standard deviation of 0.3545. Upon repeating the examination once, these candidates showed an average gain score of 0.3770 over the first testing and a loss score of 0.1290 over the second exam. Table 3 illustrates the average theta estimates across the first seven attempts.

Insert Table 3 here

There is a dramatic gain score between the first and second exam, but what appears to be a linear trend of decreasing theta estimates for exams three through six, with a mean of -0.7004 and a standard deviation of 0.0073. Figure 3 illustrates these mean thetas across attempts one through seven for CAT. The pattern of theta estimates for CAT is very similar to the pattern for P & P estimates (compare Figures 1 and 3). There is a dramatic increase between attempts one and two, a linear pattern of decreasing theta estimates from attempts two through four, and a relatively constant theta level from attempts five through seven (for CAT) and five through ten (for P & P).

Correlations between Testings

Correlations between theta estimates for the CAT attempts (for N > 25) range from 0.270 to 0.558, with a mean correlation of 0.415 between any two testings. These correlations are lower than corresponding correlations for the paper-and-pencil test, although this is to be expected given the adaptive nature of CAT testing, which produces generally larger standard errors of measurement than traditional paper-and-pencil testing. Figure 2 illustrates correlations between consecutive attempts for paper-and-pencil vs. CAT retests.

Passing Rates

Passing rates for CAT repeaters are presented in Table 4.

Insert Table 4 here

Approximately 52.8% of examinees repeating the exam one time passed. Of those repeating the exam twice, 39.3% passed on the second attempt, and 30.7% passed on a third attempt. The cumulative percentage of examinees passing on the second attempt was 52.8%, the cumulative percentage passing on the third attempt was 61.6%, and the cumulative percentage passing on the fourth attempt was 63.2%. From the cumulative percentages of Table 4, it appears that a ceiling effect similar to the paper-and-pencil passing rates, occurs after the third attempt for CAT compared to a fourth attempt for the paper-and-pencil. A comparison of Figures 4 and 5 illustrate this point. It

would appear that under CAT, the ceiling effect for repeaters occurs more quickly (following the third attempt) than for traditional paper-and-pencil testing. Additional data, however, from CAT testing is needed to confirm or extend this observation.

Figure 6 illustrates a comparison of passing rates for Paper-and-Pencil vs. CAT by number of attempts. It appears that from the perspective of overall passing rates, paper-and-pencil and CAT testing yield very similar results for attempts two through seven. For the second attempt, P & P testing yields a passing rate of 52.9% compared to 52.8% with CAT. For the third attempt, P & P testing yields a passing rate of 36.9% compared to 39.3% for CAT, and for the fourth attempt, P & P testing yields a passing rate of 31.8% compared to 30.7% for CAT. The higher passing rate on the third attempt for CAT (+2.4%) compared to P & P is likely another aspect of the ceiling effects referred to in Figures 3 and 4. By the fifth and sixth attempts, passing rates for P & P and CAT are within 0.9% and 0.4% of one another, respectively. This preliminary data suggests one interesting research question: Does CAT testing yield an efficiency of one fewer retests compared to paper-and-pencil testing, given the apparent ceiling effects for repeaters? Most likely, this cannot be fully answered until more CAT data is available for analysis.

Summary

Perhaps the most important finding that this data suggests is that of comparability of paper-and-pencil and CAT testing: repeater performance across both testing modalities is very similar. For each modality, approximately 53% of all candidates who fail the NCLEX-RN on a first attempt actually pass when repeating the NCLEX-RN one time. On a third attempt, approximately 37%

pass, and on a fourth attempt, approximately 32% of the examinees pass. For subsequent attempts for both modalities it appears that by the fifth attempt, the advantage of retesting is greatly diminished since only 19% of the candidates pass on a fifth attempt. Clearly, more CAT data is needed to assess these comparability issues in greater detail.

Table 1

Final Ability Estimates by Attempt (P & P)

Attempt	N	Mean	Std. Dev.	Minimum	Maximum
1	15272	-0.74184	0.29044	-3.49238	-0.47868
2	21600	-0.48503	0.37767	-3.09734	0.99203
3	7935	-0.65592	0.39416	-4.00000	0.60933
4	5213	-0.73219	0.36397	-2.35873	0.42368
5	3612	-0.77523	0.34245	-2.09172	0.29995
6	2937	-0.77095	0.32740	-4.00000	0.33811
7	1772	-0.78689	0.31210	-1.97837	0.53906
8	1010	-0.77027	0.30377	-1.92361	0.26237
9	465	-0.78625	0.28088	-1.77284	0.43632
10	195	-0.77597	0.25119	-1.45611	-0.20828

Table 2

Passing Rates by Attempt (P & P)

Attempt	Number Testing	Number Passing	Percent Passing	Cumulative Percent Passing
1	43,847	0	0.0%	0.0%
2	43,847	23,214	52.9%	52.9%
3	16,451	6,066	36.9%	66.8%
4	8,517	2,710	31.8%	73.0%
5	4,612	866	18.8%	74.9%
6	2,937	529	18.0%	76.1%
7	1,772	244	13.8%	76.7%
8	1,010	146	14.5%	77.0%
9	465	56	12.0%	77.2%
10	195	22	11.3%	77.2%

Table 3

Final Ability Estimates by Attempt (CAT)

Attempt	N	Mean	Std. Dev.	Minimum	Maximum
1	19,119	-0.83749	0.35447	-3.84032	-0.42325
2	19,119	-0.46045	0.51015	-3.44273	1.09693
3	4,314	-0.58941	0.49508	-2.98939	0.77024
4	996	-0.66724	0.46661	-2.90969	0.66303
5	218	-0.73216	0.42712	-2.28917	0.43162
6	38	-0.73922	0.43998	-1.95727	0.22213
7	6	-0.77420	0.48864	-1.13368	0.16019

Table 4

Passing Rates by Attempt (CAT)

Attempt	Number Testing	Number Passing	Percent Passing	Cumulative Percent Passing
1	19,119	0	0.0%	0.0%
2	19,119	10,089	52.8%	52.8%
3	4,314	1,697	39.3%	61.6%
4	996	306	30.7%	63.2%
5	218	43	19.7%	63.5%
6	38	7	18.4%	63.5%
7	6	1	16.7%	63.5%

Table 5

Mean Theta Estimates by Ethnicity (P & P)

Attempt	White		Black		Asian Other		Asian Ind		Hispanic		Pacific Isl		Native Am	
	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N
1	-0.661	5,127	-0.770	1,580	-0.962	1,693	-1.179	498	-0.828	423	-0.874	129	-0.747	74
2	-0.422	5,318	-0.602	1,734	-0.810	2,019	-0.957	637	-0.689	507	-0.747	162	-0.620	80
3	-0.539	1,215	-0.701	791	-0.887	1,319	-1.000	557	-0.800	241	-0.984	90	-0.680	37
4	-0.601	551	-0.747	563	-0.913	1,004	-0.997	588	-0.836	157	-0.984	69	-0.789	23
5	-0.659	269	-0.757	427	-0.910	802	-0.957	555	-0.810	110	-0.915	46	-0.853	18
6	-0.696	187	-0.793	361	-0.884	647	-0.888	500	-0.830	80	-0.922	40	-0.843	19
7	-0.714	105	-0.773	253	-0.867	449	-0.875	382	-0.809	41	-0.895	16	-0.750	17
8	-0.692	62	-0.761	159	-0.841	271	-0.816	241	-0.804	22	-0.955	7	-0.750	9
9	-0.662	32	-0.759	85	-0.836	135	-0.770	121	-0.867	10	-0.897	5	-0.820	6
10	-0.685	15	-0.756	34	-0.830	54	-0.711	46	-0.860	5	-1.110	2	-0.847	3

Table 6

Mean Theta Estimates by Ethnicity (CAT)

Attempt	White		Black		Asian Other		Asian Ind		Hispanic		Pacific Isl		Native Am	
	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N
1	-0.727	8,887	-0.818	1,641	-0.996	3,038	-1.220	491	-0.842	611	-0.867	186	-0.736	188
2	-0.337	8,887	-0.490	1,641	-0.651	3,038	-0.882	491	-0.478	611	-0.603	186	-0.385	188
3	-0.452	1,802	-0.588	433	-0.766	788	-0.915	204	-0.541	112	-0.705	51	-0.637	53
4	-0.543	411	-0.610	124	-0.826	201	-0.913	64	-0.779	18	-0.327	12	-0.651	14
5	-0.611	101	-0.802	30	-0.844	39	-0.927	12	-0.694	4	-1.056	3	-0.592	7

References

Eignor, D. R., Way, W. D., & Amoss, K. E. (1993, April). Establishing the comparability of the NCLEX using CAT™ with traditional NCLEX examinations. Paper presented at the annual meeting of the NCME, New Orleans.

Kingsbury, G. G. & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. Applied Measurement in Education 2, 359-375.

Lord, F. M. (1980). Application of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lunz, M. E. & Bergstrom, B. A. (1991). Comparability of decisions for computer adaptive and written examinations. Journal of Allied Health 20, 15-23.

NCLEX/CAT™ Team (1994). NCLEX/CAT™ Beta Test Retest Report. Submitted to The National Council of State Boards of Nursing. Princeton, NJ: Educational Testing Service.

Way, W. D. (1993, April). NCLEX/CAT™ Beta Test Simulations Progress Report. Princeton, NJ: Educational Testing Service.

Way, W. D. (1994a, February). NCLEX simulations report for April 1994. ETS Statistical Report.

Way, W. D. (1994b, April). Psychometric results of the NCLEX™ Beta Test. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Weiss, D. J. & Kingsbury, G. G. (1984). Application of computerized testing to educational problems. Journal of Educational Measurement, 21, 361-375.

Zara, A. R. (1992a, April). A comparison of computer adaptive and paper-and-pencil versions of the National Registered Nurse Licensure Examination. Paper presented at the annual meeting of the AERA, San Francisco.

Zara, A. R. (1992b, April). An investigation of computerized adaptive testing for demographically- diverse candidates on the National Registered Nurse Licensure Examination. Paper presented at the annual meeting of the NCME, San Francisco.