NURSE EDUCATOR
Volume 26, Number 6, pp 264–267

http://www.nursingcenter.com

Professional Issues

Moving a National Licensure Examination to Computer

Ellen Julian, PhD
Anne Wendt, PhD, RN
Denny Way, PhD
Anthony Zara, PhD

After almost 10 years of research and planning, the National Council of State Boards of Nursing (National Council) moved its nursing licensure examinations from the traditional paper-and-pencil (PP) format to computerized-adaptive testing (CAT) in April 1994. The National Council made this move to provide nursing candidates with more frequent testing opportunities, a shorter and more comfortable testing experience, quicker turn-around of results, and to relieve the state Boards of Nursing of the burden of test administration. The decision was based on the accumulation of evidence in the literature of CAT and paper-and-pencil comparability and increased CAT efficiency, and evidence gathered from its own research. This retrospective report summarizes the National Council's 10 years of research into the comparability of PP and CAT formats for the NCLEX® examinations, conducted on over 11 000 NCLEX candidates and should provide background information for other professional organizations interested in computer-based testing.

There are two separate NCLEX examinations, the National Council Licensure Examination for Registered Nurses and (NCLEX-RN®) examination and the National Council Licensure Examination for Practical/Vocational Nurses (NCLEX-PN®) examination. These two examinations have different test plans, item pools, passing standards and, to a large extent, candidate pools (some states allow RN candidates who have not fully completed their educational program to take the PN examination).[1] Historically, each was administered in paper-and-pencil format (twice a year, to a total of over 120 000 RN candidates and over 60 000 PN candidates per year). Passing rates hovered around 90% for first-time takers educated in the United States (the "reference group" on whom item difficulties are estimated).[2]

The NCLEX-RN examination in the paper-and-pencil (PP) format was a 2-day examination consisting of 372 questions—300 operational items used in scoring and 72 pretest items being field tested. All operational items had known difficulties, having been calibrated and equated to a common scale using the Rasch model.[3] The 1-day NCLEX-PN examination had 240 items, 204 of which were operational, and 36 of which were pretest. The passing standards, set using a modified-Angoff technique and reevaluated every 3 years, were maintained as logit ability levels on the Rasch-calibrated "bank scale," and each candidate's equated ability estimate was compared to that standard to determine pass or fail status.

On the PP examinations, efforts were made to better target the examination difficulties to the pass/fail standard. Historically, the average item difficulty has been lower than the passing standard, as has been the average difficulty of items in the pool. In addition, it was hard to predict and write items with difficulties near the passing standard which resulted in items with a great variety of difficulties in the item pools. These factors, combined with the high-stakes nature of the examinations and the resultant need for large item pools, made the option of building a peaked test implausible.[4] In 1986, the decision-making body of the National Council voted to pursue CAT as a more efficient method for assessing the competence of nurses to practice safely and effectively in entry-level positions.

In part because of the shortage of items with difficulties in the region of the passing standard, the goal of CAT was envisioned as estimating a candidate's ability as accurately and efficiently as possible. A candidate's examination would end when the probability of the candidate's ability being on the other side of the passing standard (ie, of the final pass/fail decision being different if more items were administered) fell below a specified level.

Author affiliations: Medical College Admission Test, Association of American Medical Colleges, Washington, DC (Dr Julian); National Council of State Boards of Nursing, Chicago, Ill (Dr Wendt); Educational Testing Services (Dr Way); Professional Licensing and Certification, NCS Pearson Company, Eden Prairie, Minn (Dr Zara).

Corresponding author: Anne Wendt, PhD, RN, National Council of State Boards of Nursing, 676 N. St. Clair, Suite 550, Chicago IL 60611 (awendt@ncsbn.org).

CAT software was developed that uses Item Response Theory (IRT) item difficulties and a maximum-likelihood ability estimation method to calculate candidates' abilities. (Note: Bayesian estimation is used until the candidate has answered at least one item correctly and one incorrectly).[5] Items are selected that both maximize the information provided and cover any underrepresented areas of the test plan. The minimum test length of 60 items was deemed to be the minimum number that could adequately cover all test plan areas. The maximum test length was set at the point where the precision provided would be approximately equivalent to that provided by the PP examination (250 items for RNs and 180 for PNs). In addition, 15 pretest items (25 for PNs, because that item pool was shallower), selected at random, would be administered, interspersed among the first 60 operational items. Each item had to be answered before progressing to the next. Once an answer option was highlighted and confirmed, and the next item selected and presented, the candidate could not return to the previous item.

Testing continued until (1) the maximum number of items was reached, (2) the 4 hours allowed had expired, or (3) the current ability estimate was more than three standard errors of measurement (SEMs) from the passing standard and the minimum number of items had been seen. The third of these "stopping rules" required that the decision being made be with 99% confidence before ending the examination. Analyses revealed that ending the examination at the point when decisions could be made with 95% confidence (1.65 SEMs) resulted in only 1% different decisions.

CAT Field Tests

From 1990 to 1992, the National Council conducted two field tests of the NCLEX-RN examination using CAT. Two separate administrations were needed because the characteristics of candidates (particularly the proportion of first-time takers and foreign-educated candidates) presenting to take the NCLEX-RN examination vary across the year. In 1992, a field test of the NCLEX-PN examination was conducted using the same design.

Eight licensing jurisdictions' Boards of Nursing and educational programs recruited volunteers for the field tests. To qualify to take part in the study each volunteer had to meet their jurisdictions' requirements. The volunteers took the NCLEX paper-and-pencil examination at a regularly scheduled administration, either followed or preceded (within 2 weeks) by the NCLEX-CAT examination. CAT items were selected from pools of over 2000 possible items. The research questions addressed were: (1) the equivalence of the pass/fail decisions made by the CAT and paper-and-pencil methodologies, (2) the impact of CAT on protected demographic groups, and (3) any impact of computer experience on the CAT-PP differences. The volunteers were told repeatedly that only the PP examination counted for the licensure decision. Somewhat lower scores on CAT were anticipated as a result.

Over 960 RN-candidates participated in the RN field tests and a similar number (912) in the PN field test. Concordance between CAT and paper-and-pencil pass/fail decisions was high (81% and 82% in the RN field tests, and 87% in the PN field test), as was the corrected (for attenuation) correlation between Rasch ability estimates from the two methodologies (RN = .83 and .93, and PN = .87). For CAT pass/fail decisions made with greater than 95% confidence (where testing ended before the maximum number of items was administered), the two methodologies agreed on 97–99% of the decisions.

Overall, candidates scored slightly lower on CAT (both average abilities and passing rates were slightly lower, although the passing rate for English-second language PNs increased). The CAT ability estimates were lower than PP for all demographic groups, but the difference was less for foreign-educated and English second-language PN candidates. No effect of computer experience on CAT performance was found. Candidates reported feeling more comfortable taking the test on computer, and thought the items were somewhat easier to understand, and much easier to read.[6,7]

Field Test Results

The Field Tests demonstrated that NCLEX candidates generally received the same pass/fail outcome from PP and CAT. The high correspondence in pass/fail decisions for candidates who received the shorter CAT examinations (because of their distance from the passing standard) was expected. Similarly, those candidates who were very close to the passing standard (±2 SEMs) would be expected to have lower decision consistency across any two versions of the examination.

Even though the slightly lower average ability estimate, and the correspondingly lower passing rate on CAT was anticipated in this study, it was still a matter of concern. Candidates' differential motivation on two versions of an examination, when one "counts" and the other does not, had to be eliminated. In other words, CAT had to "count," for a conclusive comparison of candidates' performances on the two examinations to be made. Thus a CAT Beta Test was needed.

Beta Test Design

The design for the CAT Beta Test included a nationwide, random assignment to equivalent groups design, with four groups defined for RN candidates, and three for PN candidates.

One group of approximately 2000 RN volunteers took the PP examination at their normally scheduled July administration, and another 2000 RN candidates took CAT at approximately the same time. The same numbers of PN candidate-volunteers took a special administration of a PP NCLEX-PN examination or CAT in July, 1993.

To offer insight into reasons for any observed CAT-paper-and-pencil differences, two additional, smaller groups of RN candidates and one of PN candidates (approximately 500 each) took the NCLEX examinations under modified conditions. For RNs, the modified conditions were a 1-day version of the paper-and-pencil examination (PP1day), so that possible effects of a 5-hour testing session might be identified, and a computerized-linear test (CLT), where all candidates were administered the same items and allowed to review and change their

answers, so that possible effects of the adaptive nature of CAT and the lack of opportunity for answer review might be investigated. Because the NCLEX-PN examination is normally a 1-day examination, only the CLT additional condition was administered.

As with the field tests, the dual focuses of the Beta Test were on the comparability of pass/fail decisions by the different testing methodologies, with the added element of consequences for CAT performance, and of any disparate impact of CAT on the protected demographic groups. A higher than representative proportion of repeat test-takers, ESL candidates and candidates from Hispanic, African American, and Filipino ethnic groups was sampled to enable comparisons with sufficient statistical power. Recruitment of volunteers was assisted by the offer of a free examination (the normal PP examination fee was $40), and of a free CAT retake (at an earlier time than the next paper-and-pencil would be available) if they failed the Beta Test, no matter which condition they took the examination.

Beta Test Administration

Almost twice as many candidates applied for the Beta Test as were selected, more than needed for every group except Hispanics. To be eligible for participation in the Beta Test, the candidates must have registered for the normal PP examination and the state in which they desired licensure must have declared them eligible (having completed all of the state's requirements for taking the NCLEX examination).

Overall, 6377 RN candidates and 4428 PN candidates were randomly assigned to Beta Test conditions, and after random selection were stratified by state and demographic group. These numbers included an anticipated loss to attrition, with an original target of 5000 RN and 4500 PN candidates. The number of PN candidates selected was lower than anticipated due to unexpected patterns of eligibility (eg, graduation dates not synchronized with the Beta Test). The proportions of candidates from protected demographic groups was adjusted to ensure the desired power for the hypothesis tests (or at least to maximize it, in the case of Hispanics).

Over 100 test administration sites delivered the CATs over a 3-week period. The PP was administered as normally scheduled. Beta Test volunteers who were assigned to the PP group took their NCLEX-RN examination with all of their non-volunteering peers. The CLT and PP1day conditions were administered on those same days.

The item pool available for the Beta Test consisted of 1863 RN items and 1618 PN items, generally representative of the test-plan areas and all difficulty levels.

Beta Test Results

In all, 5902 RN candidates and 3317 PN candidates participated in the Beta Test. Their distribution across conditions is shown in Table 1.

A small number of candidates (22 RN and 25 PN) were tested in other than their assigned condition, and they were not included in the comparisons of conditions. The average ability level and passing rates for each comparison are shown in Table 2.

The primary contrasts of interest were between the CAT and PP groups. The other groups existed only to be used if needed for investigation into the source of differences in the CAT and PP groups. No significant difference was found between the CAT and PP average competence estimates, for either the RN or PN groups! In addition, no differences were found between CAT and PP average competence estimates for any of the critical subgroups (minority and ESL candidates)! While no statistically significant differences in average or standard deviation of the ability distributions were found, the general shapes of the two distributions did differ. The CAT ability distribution was more peaked than the PP distribution, and apparently bimodal, with one peak at about the average (above the passing standard), and the other just below the passing standard.

RN and PN passing rates for CAT and PP were compared, for the total group and each of the critical subgroups. Of these comparison tests, only one was significant—African-Americans PN candidates had a higher passing rate on CAT than PP. In addition, passing rate analyses were conducted on the combined RN and PN groups, to increase the power of the statistical tests for some of the smaller critical subgroups. *None* of the passing-rate differences were greater than 4% and *none* of the statistical comparisons were significant.

For both RN and PN examinations, just over one-half of the candidates took the minimum number of questions, and approximately 15% took the maximum number. The median time spent on the examination was approximately 1.6 hours (slightly less for PNs, slightly more for RNs). Approximately 7% of the RN candidates, and less than 1% of the PN candidates, ran out of time before satisfying the standard-error stopping rule or completing the maximum number of questions.[8,9]

Conclusion

In preparing for the advent of CAT, the National Council performed two basic types of comparability studies: test-retest, where each candidate took both CAT and paper-and-pencil exam-

Table 1. Candidates Participating in Beta Test

Examination	CAT	PP	CLT	PP1day	Total
NCLEX-RN	2,452	2,562	432	456	5,903
NCPLEX	1,566	1,440	311	NA	3,317

CAT, Group Testing Via Computerized Adaptive Testing; PP, Group Testing Via Trational Paper-and-Pencil Test (two Days; CLT, Group Testing Via Computer Linear test; PP1day, Group Testing Using Traditional Paper-and-Pencil Test (one Day).

Table 2. Average Ability and Passing Rates for Each Testing Condition

Examination		CAT	PP	CLT	PP1day
RN	Avg. Ability Est (SD)	−0.22 (0.63)	−0.20 (0.57)	−0.19 (0.54)	−0.19 (0.54)
	% Passing	68.6	68.7	70.4	70.4
	N	2,430	2,562	432	456
PN	Avg. Ability Est (SD)	−0.26 (0.62)	−0.23 (0.62)	−0.17 (0.60)	
	% Passing	74.9	72.7	78.1	
	N	1,555	1,426	311	

CAT, Group Testing Via Computerized Adaptive Testing; PP, Group Testing Via Traditional Paper-and-Pencil (Two Days); CLT, Group Testing Via Computer Linear Test; PP1day, Group Testing Using Traditional Paper-and Pencil Test (One Day)

inations, and randomized-groups, where candidates were assigned to either experimental or control groups. The test-retest studies focused on whether the *same* candidates would pass both examination formats, whereas the randomized-groups addressed the question of whether the same *number* of candidates would pass each format.

A total of over 11 000 RN- and PN-candidates participated in these research studies. No evidence was found that CAT disadvantaged any of the candidate groups. When both CAT and paper-and-pencil counted towards licensure, the slight difference in passing rates found in the test-retest study evaporated. Computer experience was not found to have any effect, either through analysis of performance data or in the candidates' survey responses. In short, *all* of the evidence pointed towards CAT being a viable format for administering the NCLEX examination.

References

1. National Council of State Boards of Nursing. *Profiles of Member Boards*. Chicago, 1996.
2. National Council of State Boards of Nursing. *Licensure and Examination Statistics*. Chicago, 1996.
3. Wright BD, Stone M. *Best Test Design*. Chicago: MESA Press, 1969.
4. Birnbaum A. Some latent trait models and their use in inferring an examinee's ability. In FM Lord, MR Novick, *Statistical theories of mental test scores*. Reading MA: Addison-Wesley, 1968.
5. Thissen D, Mislevy RJ. Testing algorithms. In Wainer H (Ed.), *Computerized Adaptive Testing: A Primer*. Hillsdale, NJ: Erlbaum, 1990.
6. Zara AR. *An overview of the NCLEX/CAT® beta tes*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, 1994.
7. Zara A R. *A comparison of computerized adaptive and paper-and-pencil versions of the national registered nurse licensure examination*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1992.
8. Way WD. *Psychometric results of the NCLEX® Beta Test*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, 1994.
9. Eignor DR, Way WD, Amoss KE. *Establishing the comparability of the NCLEX using CAT® with traditional NCLEX examinations*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, 1994.

News Notes & Tips

Continued from page 261

school of nursing. Brown and Kohlenberg[3] addressed the development and use of an assessment instrument. They developed a three-tiered scale that is used to provide a documented record of evaluation of the applicant's credentials and interview impressions. This type of documentation is important for two reasons. One is the litigious environment in regard to equal opportunity employment. The second is the availability of data to evaluate the process after the dean is hired.

Evaluate. The final step in the search for a dean should be to evaluate the process after the dean has been hired. Several questions should be addressed in this evaluation. The first question needs to evaluate if the new dean met the qualifications stated at the beginning of the search. If the qualifications were not met, the second question that needs to be determined is at what step in the process a misdirection occurred. The third question to be evaluated is if the process used by the institution was effective or if changes could be recommended to improve the process for seeking a dean.[1]

Searching for a leader in a school of nursing is expensive, time consuming, and immensely important. Finding ways to increase the likelihood of a successful placement will improve spending, increase time for teaching, research, and service activities, and increase the likelihood of successful placement. Being able to predict or assess current trends, accurately advertise position qualifications, and appropriately select applicants for the position of dean can help everyone in this process.

References

1. Nichols EF, Bower DA, Collier J, Gray VR. Searching for a dean. *Nurse Educator*. 1989;14(5):9-13.
2. Bidwell AS. Searching for the dean: contemporary and traditional considerations. *Nurse Educator*. 1998;23:46-50
3. Brown HN, Kohlenberg EM. Searching for a dean: getting the qualities you want. *J Nurs Ed*. 1994;33:93-94.
4. Coop LA. When a dean steps down (Editorial). *J Prof Nurs*. 1995;11:197-198.
5. Sherrod RA. Chairing a dean's search committee: notes for the novice. *Nurse Educator*. 1996;21(3):5,10.

Continued on page 270