This retrospective study examined whether practice in computer-based tests (CBTs) in student registered nurse anesthetists improved scores on the National Certification Examination (NCE). A group of students with extensive practice in CBTs was compared with a group of students whose tests were administered primarily on paper. Groups were matched on age, grade point average (GPA), and gender (n = 73 each). Higher GPA was associated with higher NCE scores overall. Adjusted for GPA (with analysis of covariance), the mean NCE proficiency score (φ) in the CBT group was 2.68 (95% confidence interval, 2.54-2.82), which was higher than the mean score in the paper-based group of 2.36 (95% confidence interval, 2.22-2.50), with an effect size of 0.52. When subgroups were examined, CBT practice improved NCE scores only in those students with graduate GPA less than or equal to 3.50, with an effect size of 1.1. It was concluded that, controlling for GPA, student registered nurse anesthetists at a university with greater exposure to CBTs had higher scores on the NCE than a comparison group with less practice in CBTs. This difference was significant only in students with GPA of 3.50 or less, consistent with a beneficial effect of practice in CBTs.

Keywords: Certification of nurse anesthetists, computer-based testing, education of nurse anesthetists, practice effect, test mode effect.

Successful completion of the National Certification Examination for Nurse Anesthetists (NCE) is a vital outcome of graduate education in nurse anesthesia, since passing is necessary for entry into practice. The NCE changed from pencil and paper to a computer adaptive test form in 1996.1 Students at that time took paper-based tests exclusively. Even today, almost 40% of students taking the NCE report that their nurse anesthesia educational programs did not include any computer-based testing.2 For many, their exposure to computer-based tests (CBTs) was limited to the National Council of State Boards of Nursing Licensing Examination for Registered Nurses and the Graduate Record Examination. Whether these students would benefit from practice in taking examinations administered on computer and, if so, what type of preparation is best have not been studied.

Whether computer-based and paper-based modes of test administration produce equivalent scores has been studied extensively in many populations over the past 15 years as testing has migrated to the CBT format.3-6 While test results from paper-based tests or CBTs may be equivalent in the aggregate, one testing mode or the other may produce a higher score for an individual student. This is called the test mode effect. Many factors have been studied as possible contributors to this effect: race, gender, ethnicity, general computer familiarity, test anxiety, computer anxiety, prior exposure to computer testing, type of test (power vs speeded), subject matter, prior ability, and interface (eg, font, legibility, line length, scrolling, and item review).7-19

Although no studies on the test mode effect in student registered nurse anesthetists could be identified, there is some evidence that subgroups do not perform equally well on the NCE, which is a possible indication that the test mode effect is present. For example, students in their 40s have an 85% pass rate, compared with a 94% pass rate in those less than 30 years of age.20 Secondary analysis of these data shows there is a low probability that this association between age and passing rate arose solely because of chance (P < .005 by χ²).

No studies were identified examining the value of practice in CBTs for nurse anesthesia graduate students. Few studies could be identified that examined the benefits of practice on CBTs as an instructional strategy in any type of students, and these studies are flawed or incomplete. One, an uncontrolled study of a technology-rich classroom environment in high-risk children, failed to demonstrate gains compared with normative results on state-mandated tests, though at-risk children in a technology-rich environment performed better than teachers expected.21 Another, in undergraduates, demonstrated the learning value of both in-class and computer-based quizzes for performance on later in-class tests in a controlled, but not randomized, design.22 However, the authors did not evaluate whether practice in a CBT environment increased scores on later computer-administered tests.

The purpose of this study was to determine if prac-
tice in computer-based testing improved scores on the Certification Examination for Nurse Anesthetists, by comparing students with extensive experience in computer-based testing, to students who had little or no experience. The potential importance of the findings is plain to current students, who want to have the best chance to pass the NCE on their first attempt. This study may also be important for faculty in nurse anesthesia, in other graduate specialities, or for those teaching undergraduates. First-attempt pass rates on licensure and certification examinations are tracked by state boards of nursing, accreditors, and potential applicants. Better preparation of graduates to pass these examinations on the first attempt may enhance compliance with accreditors and regulators, as well as enhance the public perception of educational program quality. This, in turn, may result in enhanced alumni relations and a larger applicant pool.

**Materials and Methods**

Records of 205 graduates (1998-2009) from a single university were examined. These existing records, kept for general administrative purposes, included transcripts, demographic data, and reports from the Council on Certification of Nurse Anesthetists. The records were entered into a database without names, to reduce any potential bias. Permission to access these records for the purpose of this study was granted by the local institutional review board. This study was a case-control design in that it was retrospective, employed a comparison group, and lacked random selection of participants or random allocation to groups. 

Beginning in fall 2005, all hourly and final course examinations in a nurse anesthesia program were delivered on computer by eZ.exam (Questionmark Corp, Norwalk, Connecticut). These examinations were high stakes because they determined course grades and academic progressions. Faculty saw the potential benefits of adopting CBT as (1) economic (saving paper, duplicating costs, and faculty time in correcting examinations); (2) student satisfaction (quicker score feedback); (3) academic integrity (preserved or enhanced test security); and (4) instructional (quicker feedback on wrong answers, more practice in CBT delivery). The primary desired outcome was to improve students’ scores on the NCE, perhaps by diminishing test anxiety.

The eZ.exam testing software was not meant to mimic the NCE in all respects, although they share certain similarities. Tests were proctored in a university computer laboratory, and all students took examinations at the same time. The tests were composed of 40 to 100 items, in 2 to 4 sections organized by topic. Students were given 2 hours to complete these examinations but often completed them in less time. The NCE is a 3-hour examination with a maximum of 180 questions. In both eZ.exam and the NCE, questions are presented on screen 1 at a time. While students may not review or change any previous answers on the NCE, eZ.exam was set up to allow limited review in each section of the test. Once the section was complete and submitted, review or changing answers was not permitted. Unlike the NCE, in which items are assembled adaptively so that each examinee receives a unique test form assembled from pooled items, all students taking a test on eZ.exam received the same test items, presented to each student in random order. Like the NCE, most eZ.exam questions were multiple choice, with a few alternative format items (eg, short answer, calculation, or multiple-response items). Students receive score feedback immediately on completion in both the NCE and eZ.exam. In eZ.exam, they also received feedback on items answered incorrectly, before they left the examination room. This was much quicker than the typical wait to review wrong answers (which for paper-based tests was at least a week, at the next class session).

Approximately 2,000 graduates take the NCE each year. The sample for the current study (N = 150) was chosen from a single university. The CBT (experimental) group graduated between 2007 and 2009 (n = 75). They had more extensive experience with high-stakes CBTs, having taken 23 midterm and final course examinations in this format. A comparison group of students (paper group) was chosen, who graduated from 1998 to 2006, with most (57 of 75) graduating between 2002 and 2006. The paper group had little or no CBT experience (defined as 0 to 3 tests done on computer during the program). The comparison group was chosen to have an equal number of participants as the CBT group, and was matched to them on age, gender, and graduate grade point average (GPA), to control extraneous sources of variability, especially academic ability. The matching process was carried out by the author, who was blinded to dependent variable scores during the selection of the comparison group.

The computer administering the NCE calculates a proficiency score ($\phi$) for each NCE examinee using the Item Response Theory. Proficiency scores are not published. They ranged from approximately -0.5 to +4.0 in this sample. The level of proficiency required to pass the NCE is determined by the Council on Certification based on a modified Angoff method. The proficiency scores are converted to a scale. Scaled scores are equated, so that equal scores represent a constant level of ability from year to year, regardless of the difficulty of the test form. The passing score for the NCE is set at 450. Before 2008, NCE score reports were trimmed, so that the highest reported score was 600, and the lowest reported score was 300. In 2008, this was changed so that actual scores were reported. To avoid artifact related to the scaling and equating process, plus the trimming of reported scores, $\phi$ scores were used for the NCE score.
The standard deviation (SD) of scaled NCE scores nationally is 47.9 units. With a sample size of 75 in the CBT and paper groups, the study had adequate power for a t test to detect a difference between groups as small as 22 units. This corresponds to an effect size of 0.46, considered a medium effect size. Because the scaled passing score on the NCE is 450, and the maximum reported score is 600, this study had adequate power to detect differences between groups as small as 15% of the range of passing scores. Any smaller improvements in score due to CBT practice were not considered as practically important.

Descriptive statistics were calculated for each group. Interval or ratio data were compared between groups using unpaired 2-tailed Student t tests. This included GPA, age at graduation, Graduate Record Examination scores (verbal, quantitative, and analytical writing), examination scores from core courses (ANE [Anesthesia] 5490 Basic Principles of Anesthesia, and ANE 5600 Pharmacology I), and Self-Evaluation Examination scores. Either $\chi^2$ or Fisher exact test was used to compare groups on gender and the number of students passing the NCE on the first attempt.

All NCE scores were analyzed with analysis of covariance (ANCOVA). The NCE scores were compared between groups (CBT or paper), with GPA as the covariate. The GPA was chosen to level any remaining differences in academic ability between groups, after matching was completed. Of all the available variables measuring academic ability, it had the highest correlation with NCE score ($r = .518$, $P < .01$). In addition, including GPA as a covariate increased the explanatory power of the model ($R^2$ increased from .022 to .315). The data generally satisfied the assumptions of ANCOVA, including homogeneity of regression. Although the data distributions and variance of the groups were not equal (Figure 1), ANCOVA maintained acceptable robustness and power, especially since the sample size was more than 12.

To determine if the benefit of practice was more pronounced in any subgroups, an analysis of variance (ANOVA) was performed with post hoc blocking. Students were classified into 3 GPA blocks based on graduate GPA, as follows: less than or equal to 3.50, 3.51 to 3.75, and 3.76 to 4.00. The NCE score was analyzed by group (CBT or paper), GPA block, and their interaction. If the interaction term was significant, planned contrasts were performed to determine whether practice in CBTs pro-

Figure 1. Distribution of NCE Proficiency (phi) Scores

a There are fewer low scores in the CBT group; the paper testing group was more negatively skewed (Kolmogorov-Smirnov test, $P = .04$). Variance in the 2 groups was not equal (Levene test, $P < .05$).

Abbreviations: NCE, National Certification Examination; CBT, computer-based test.
duced higher NCE scores in the low, medium, or high GPA categories.34

Statistical analysis software (SPSS version 18, SPSS Inc, Chicago, Illinois) was used as an aid to statistical computation. A 2-tailed $P$ value < .05 was accepted as evidence of statistical significance throughout.

**Results**

There were 146 NCE scores analyzed, because 4 φ scores were unavailable. Table 1 shows that the paper and CBT groups were comparable. They differed only on scores achieved on the first examination in ANE 5490 Basic Principles of Anesthesia. The groups were well matched on GPA, age at graduation, and gender. There were 54 women (of 73 subjects) in the paper group, and 53 women of 73 in the CBT group, a difference that was not statistically significant. The groups had equivalent numbers who did not pass the NCE on the first attempt: 3 in the paper group and 4 in the CBT group, a nonsignificant difference.

The mean NCE proficiency score was higher in the CBT group (2.68 ± 0.44 [adjusted mean ± SD]) than in the paper group (2.36 ± 0.91) ($P = .002$). The mean difference was 0.32. The range of φ scores in the sample was 3.41 units. The effect size for CBT vs paper groups overall was 0.52.

A blocked analysis of variance showed that NCE scores generally increased with higher GPA ($P < .001$), but the Group × GPA block interaction was significant ($P < .01$). This indicated that more exposure to computer-based testing did not have the same effect in all GPA categories. The NCE scores were similar in each GPA category, except in students with GPA of 3.50 or less. In those students with GPA at or below 3.50, NCE scores were significantly higher in the CBT group than the paper testing group ($P = .02$), as shown in Table 2 and Figure 2. The effect size for CBTs vs paper tests in students with GPA of 3.50 or less was 1.1. The effect size for this group of lower-ability students, in scaled score terms, would represent an approximately 50-point gain in score on the NCE.

**Discussion**

As one might expect, student registered nurse anesthetists with higher GPA achieved higher scores on the NCE. However, students in this sample with more extensive experience in computer-based testing achieved higher scores on the NCE than students with less experience in CBTs. This benefit of practice with CBTs was most pronounced in students with the lowest GPA (see Figures 1 and 2, and Table 2). The beneficial effect of practice in computer-based testing on NCE scores has, to the author’s knowledge, not been demonstrated previously for student registered nurse anesthetists.

The passing rate did not differ between groups, since they were matched on GPA, the variable most likely to be associated with scores on the NCE. However, higher NCE scores, particularly as seen in this study in the lower-GPA students, would increase the likelihood of passing the NCE.

More practice in CBTs may have decreased test or computer anxiety, particularly in students with lower academic ability. Students with higher ability may be indifferent to test mode. They may have sufficient resources and academic self-confidence to do well even in a test mode they do not prefer or that is unfamiliar. In high-stakes tests, students may perceive CBTs more negatively, elevating anxiety and perhaps decreasing scores.35,36 Test anxiety is inversely proportional to academic ability, higher in both women and Hispanic students, and causes lower performance on academic tests.37,38 Computer anxiety may decrease the capacity of working memory, thus functioning as extraneous cognitive load.11 Students with high ability may be able to compensate for this ex-

<table>
<thead>
<tr>
<th>Variable</th>
<th>Paper</th>
<th>CBT</th>
<th>$P$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade point average (0-4)</td>
<td>3.64 ± 0.18</td>
<td>3.59 ± 0.20</td>
<td>.11</td>
</tr>
<tr>
<td>Age at graduation (y)</td>
<td>33.1 ± 5.54</td>
<td>32.3 ± 5.17</td>
<td>.33</td>
</tr>
<tr>
<td>GRE verbal score</td>
<td>435.0 ± 65.6</td>
<td>450.0 ± 84.5</td>
<td>.24</td>
</tr>
<tr>
<td>GRE quantitative score</td>
<td>504.0 ± 93.9</td>
<td>519.0 ± 116</td>
<td>.38</td>
</tr>
<tr>
<td>GRE analytical writing score</td>
<td>3.83 ± 0.69</td>
<td>4.02 ± 0.59</td>
<td>.28</td>
</tr>
<tr>
<td>ANE 5490 exam 1 (%)</td>
<td>92.3 ± 5.33</td>
<td>88.5 ± 5.04</td>
<td>&lt;.001$^b$</td>
</tr>
<tr>
<td>ANE 5600 exam 1 (%)</td>
<td>90.7 ± 6.34</td>
<td>90.0 ± 6.78</td>
<td>.69</td>
</tr>
<tr>
<td>SEE score</td>
<td>396.0 ± 34.5</td>
<td>391.0 ± 33.7</td>
<td>.37</td>
</tr>
</tbody>
</table>

**Table 1. Comparison of CBT and Paper Groups**

$^a$ Values are mean ± SD. Probability determined by $t$ test.

$^b$ $P < .05$.

Abbreviations: CBT, computer-based test group; GRE, Graduate Record Examination; ANE, anesthesia; exam, examination; SEE, Self-Evaluation Examination.
traneous cognitive load by increased effort; lower ability students may not be able to do so.\textsuperscript{7,38,39}

The results suggest several strategies for improvement in testing for educational programs. Reduction of test anxiety is important. The provision of immediate correct-answer feedback (for items answered incorrectly) after tests given by eZ.exam may have provided a learning benefit. While tests are given primarily for purposes of assessment, they may also improve learning.\textsuperscript{22,40-42} This testing effect may improve scores on later tests, by providing practice in retrieval from long-term memory, thereby strengthening memory traces more than simple rehearsal practice, particularly if feedback is provided immediately, and if tests are spaced over time.\textsuperscript{40}

It is likely that there are test-taking skills unique to the CBT environment. Testing during the educational program should take place on a computer, to promote automaticity in manipulating the keyboard and to decrease computer anxiety. The tests should be spaced apart in time. Feedback should be provided immediately.

The results of the current study have suggested changes in the computer-based testing strategies employed in the nurse anesthesia program from which the sample was chosen. Orientation to computer-based testing has been improved (including study skills and test anxiety reduction strategies), and a board review course with a comprehensive examination has been instituted in the senior year. Testing now includes alternative format questions. Such items as drag-and-drop, calculation, and multiple-response are now included as part of the NCE,\textsuperscript{43} so including them in the comprehensive examination offered in the senior year should enhance preparation for the NCE.

<table>
<thead>
<tr>
<th>GPA Category</th>
<th>Paper Mean ± SD (n)</th>
<th>CBT Mean ± SD (n)</th>
<th>Mean difference (CBT – paper)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤3.50</td>
<td>1.65 ± 1.06 (n = 16)</td>
<td>2.38 ± 0.41 (n = 26)</td>
<td>0.73</td>
<td>.02\textsuperscript{a}</td>
</tr>
<tr>
<td>3.51-3.75</td>
<td>2.33 ± 0.79 (n = 33)</td>
<td>2.61 ± 0.34 (n = 27)</td>
<td>0.28</td>
<td>.07</td>
</tr>
<tr>
<td>3.76-4.00</td>
<td>3.04 ± 0.43 (n = 24)</td>
<td>2.96 ± 0.39 (n = 20)</td>
<td>−0.08</td>
<td>.52</td>
</tr>
</tbody>
</table>

\textsuperscript{a} P < .05 between paper and CBT groups at each GPA level by planned contrast within analysis of variance.

Abbreviations: NCE, National Certification Examination; GPA, grade point average; CBT, computer-based test group.

\textbf{Figure 2. NCE Proficiency Scores in CBT and Paper Groups at 3 Categories of GPA}\textsuperscript{a}

\textsuperscript{a} Data are mean ± SD, with SD represented by height of error bar. Asterisk indicates \( P < .05 \) between paper and CBT groups at each GPA level by planned contrast within analysis of variance.

Abbreviations: NCE, National Certification Examination; GPA, grade point average; CBT, computer-based test group.
Further research might compare students from multiple university sites. This would allow inferences to be drawn with greater confidence about the entire national population of student registered nurse anesthetists. The data could also be reexamined for means of measuring practice effects, which were not measured in the current study. For example, if computer-based testing practice is beneficial, and test item difficulty is relatively constant, then students might spend less time per item on tests given later in the educational program, compared with earlier tests.

There are alternative explanations for the results. Without random selection and allocation to groups, it cannot be assured that academic ability was comparable between them. Retrospective designs suffer from several threats to internal validity: maturation, selection, and history.44 Because students were sampled from different time periods, the curriculum may have changed in subtle ways that produced higher NCE scores for more recent students. Finally, the difficulty of the NCE test form may have been higher in the past, making it more likely that recent scores would be higher. The likelihood that either change actually occurred is small, however, since the national average scaled score for first-time test-takers of the NCE has decreased from 560 in 2002 to just above 500 in 2009.20

Data were obtained from students at a single university. The extent to which these individuals are typical of student registered nurse anesthetists generally is unknown. This study was conceived after the CBT program was implemented, so it is retrospective and evaluative in nature, and cannot establish causation. To do so would have required randomly allocating some students to CBTs and some to paper testing. Not only would this have been complex, it also would have created an appearance of inequitable treatment, which was unacceptable to the faculty. Therefore, generalizing the results beyond these individuals should be done cautiously.

The current study describes the benefits of practice in CBTs only for those who successfully completed the educational program. If there are new test-taking skills unique to the CBT environment, it is possible that failure to adapt to CBTs may have been a factor contributing to academic attrition.

In summary, student registered nurse anesthetists with higher GPA achieved higher scores on the certification examination. Students in this sample with more extensive experience in computer-based testing achieved higher scores on the NCE than did students with less experience in CBT. This benefit of practice with CBTs was most pronounced in students with the lowest GPA (≤3.50).

REFERENCES

1. Zaglaniczny KL. The transition of the National Certification Examination from paper and pencil to computer adaptive testing. AANA J. 1996;64(1):9-14.


AUTHOR

Michael P. Dosch, CRNA, PhD, is associate professor and chair of nurse anesthesia at the University of Detroit Mercy, Detroit, Michigan. Email: doschmi@udmercy.edu.

ACKNOWLEDGMENTS

This research was supported by a grant from the AANA Foundation. The author wishes to acknowledge Karen Plaus, CRNA, PhD, FAAN, and Timothy Muckle, PhD, of the National Board of Certification & Recertification for Nurse Anesthetists (NBCRNA) for their assistance. Finally, the author wishes to thank the members of the Wayne State University dissertation committee: Ingrid Guerra-Lopez, PhD (chair), Prudentia Worth, CRNA, PhD, Timothy Spannaus, PhD, and James Moseley, EdD.

DISCLAIMER

The author is a member of the National Certification Examination Committee of the NBCRNA. This research was conducted as a dissertation for the author’s doctoral degree at Wayne State University, Detroit, Michigan. The opinions expressed herein are the views of the author and do not reflect the policy or views of the University of Detroit Mercy, Wayne State University, or the NBCRNA.