

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Copyright is retained by the first or sole author, who grants right of first publication to *Practical Assessment, Research & Evaluation*. Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. PARE has the right to authorize third party reproduction of this article in print, electronic and database forms.

Volume 29 Number 7, March 2024

ISSN 1531-7714

Transforming Assessments of Clinician Knowledge: A Randomized Controlled Trial Comparing Traditional Standardized and Longitudinal Assessment Modalities¹

Shahid A. Choudhry, Timothy J. Muckle, Christopher J. Gill, Rajat Chadha, Magnus Urosev, Matt Ferris,
John C. Preston

National Board of Certification and Recertification for Nurse Anesthetists¹

The National Board of Certification and Recertification for Nurse Anesthetists (NBCRNA) conducted a one-year research study comparing performance on the traditional continued professional certification assessment, administered at a test center or online with remote proctoring, to a longitudinal assessment that required answering quarterly questions online on demand. A randomized controlled trial of 1,000 certified registered nurse anesthetists (500 randomly assigned to the traditional assessment group and longitudinal assessment group) aimed to 1) compare assessment performance between groups, 2) compare perceptions and user experience between groups; and 3) describe participant feedback about usability of the longitudinal assessment platform. The mean scaled score for the traditional assessment group exceeded that of the longitudinal assessment group when scoring the first responses; however, upon scoring the longitudinal assessment group's most recent responses on repeat questions previously answered incorrectly, the mean scaled score was higher than the traditional assessment group. Both groups were satisfied with their experience, with slightly higher feedback ratings for the longitudinal assessment group who also found the platform easy to use and navigate. Overall results suggest the longitudinal assessment is a feasible, acceptable, and usable format to assess specialized knowledge for continued healthcare professional certification.

Keywords: Longitudinal Assessment, Traditional Assessment, Continued Professional Certification, Continued Competency, Adult Learning, APRN Credentialing, Natural Language Processing

¹ The authors would like to express our sincere gratitude to the CRNA participants of this study. Their willingness to share their time and experiences with us was essential to the study's success. We would also like to acknowledge the invaluable contributions of the Evaluation and Research Advisory Committee: Longitudinal Assessment Subcommittee. Their dedication and service to the profession laid the foundation for this study, and we are grateful for their support and service to the profession. The members, in alphabetical order, are as follows: Myron Arnaud, DNP, CRNA; Grady Barnhill, MEd; Deniz Dishman, PhD, DNAP, CRNA, NSPM-C; Sarah Giron, PhD, CRNA, FAANA; Chuck Griffis, PhD, CRNA, FAANA; Susan P. McMullan, PhD, CRNA, FAANA, FAAN; Timothy A. Newcomer, PhD, CRNA; Jared W. Riel, MA, ICE-CCP; Dennis Spence, PhD, CRNA, FAAN; Andi N. Rice, DNP, CRNA; Robyn C. Ward, PhD, CRNA, FAANA. The authors would also like to sincerely thank Internet Testing Systems for setting up the longitudinal assessment platform, which facilitated data collection for this study.

Introduction

Background and objectives

Over the past decade, a paradigm shift has been observed from a traditional model of cross-sectional, point-in-time, continuing certification examinations that reoccur at set intervals to a prospective, longitudinal assessment format over a continuum with a more frequent cadence. Diplomates have not perceived single-point-in-time examination as an optimal way to promote or assess learning over time and have found the comprehensive cognitive assessment not conducive to lifelong learning (Culley et al., 2013; Sun et al., 2016). Since the reported use of longitudinal assessment in continuing certification amongst healthcare professions (Reid et al., 2018, Spence et al., 2021) and the publication of the Continuing Board Certification: Vision for the Future Commission report (ABMS, 2019), 17 medical specialty boards have implemented and moved to longitudinal assessments for maintenance of certification (ABMS, 2023), together with other continued board certification exams (NBRC, 2023; NCCPA, 2023).

Longitudinal assessment is based upon adult learning principles and defined as an alternative method of assessment that involves continuous measurement and attainment of knowledge over an extended period that seeks to promote learning and retention (Price et al., 2018; Giron et al., 2021; ABMS, 2022). Longitudinal assessment as a method shows significant application broadly in knowledge assessment and acquisition among experienced healthcare professionals and likely plays a role in assuring public safety by bringing knowledge deficits to the fore (Reid et al., 2018; Price et al., 2018; Griffis et al., 2022). Given the acceleration in growth of clinical knowledge and medical research, it is essential that healthcare professionals are provided tools to identify knowledge gaps (Fry et al., 2023).

Longitudinal assessment has been identified as a strategy to improve continuing professional certification for healthcare professionals by providing ongoing evaluation of knowledge, skills, and competence. The potential benefits of longitudinal assessment have also been identified, including enhancing patient safety, increasing accountability, and accentuating the importance of ongoing professional

development and lifelong learning. It has been suggested that longitudinal assessments can play a crucial role in promoting these principles in the healthcare field (Giron et al., 2021). Additionally, a scoping review (Ward et al., 2023) examined the use of longitudinal assessment and found it possessed desirable attributes and could be a valuable tool to assess lifelong learning and competence in healthcare professionals and certifying organizations, offering benefits such as addressing knowledge gaps and improving exam performance.

Furthermore, a concept analysis highlighted the potential benefits of implementing longitudinal assessment for Certified Registered Nurse Anesthetists (CRNAs), including improved clinical reasoning, enhancing patient care, and promoting lifelong learning (Griffis et al., 2022). This analysis also identified the antecedents and attributes of longitudinal assessment, emphasizing the significance of a standardized assessment tool that includes:

- Utilizing principles of psychology to enhance the testing effect of learning (frequent, repetitive testing)
- Spaced learning (exposure to materials interspersed with other activities)
- Interleaving subject matter (simultaneously presenting several different learning topics)
- Providing instant/immediate feedback
- Learning through a repetitive experience
- Offering a convenient learning platform
- Being self-directed

Rather than a singular comprehensive examination, longitudinal assessment delivers shorter, periodic assessments, with immediate feedback and rationales that show significant promise in helping CRNAs increase their understanding and reinforce their knowledge in nurse anesthesia practice.

All CRNAs have been required to take the Continued Professional Certification Assessment (CPCA), which is a traditional single-point-in-time assessment required to maintain their certification as part of the Continued Professional Certification (CPC) Program. The CPCA is traditionally taken every eight years. The CPCA is not administered as a pass/fail

assessment, but a performance-standard assessment designed to evaluate current anesthesia knowledge of CRNAs in the four core domains of nurse anesthesia practice (NBCRNA, 2021) and identify potential areas where additional education may be needed.

Research study specific aims

NBCRNA sought to better understand longitudinal assessment as an alternate format to the traditional CPCA to assess and maintain CRNA knowledge over time. Such assessments would need to meet the CPC Program requirements and be evidence-based, comparable, feasible, usable, and acceptable. Within the context of this article the CPCA will be referred to as the “traditional assessment” format. The longitudinal assessment version of the CPCA, or CPC-LA, will be referred to as the “longitudinal assessment” format.

In keeping with the above stated purpose, the aims of this mixed-methods research study were to:

- 1) Compare pass rates and mean scaled scores on the assessment among CRNA participants who took the traditional assessment in its current form versus those who took the assessment in a longitudinal assessment format, and to determine if participants who responded incorrectly on their first question attempt in the longitudinal assessment group improved their performance upon repeat administration of those questions.
- 2) Discern any differences in perceptions and attitudes using an agreement Likert-scale among CRNA participants who completed the traditional assessment versus longitudinal assessment on several statements, including overall satisfaction with their assessment experience and promotion of lifelong learning.
- 3) Describe longitudinal assessment participants’ experience and engagement through data triangulation by analyzing quantitative data collected from surveys and the platform, as well as from focus group discussions, to better understand participants’ frequency of answering questions, timing of completing quarterly assessments, and overall usability.

Methods

Study design, recruitment and eligibility

This mixed-methods research study was a prospective randomized controlled trial, using a parallel design, with two arms and an allocation ratio of 1:1. A sample from the eligible population was chosen randomly in the control group, with an equal-sized group chosen to receive the intervention at random from the same eligible population to reduce selection bias. To recruit a participant sample into the study, a call for volunteers with a link to complete a baseline survey was sent out using SurveyMonkey early in 2022 to approximately 44,000 CRNAs in NBCRNA’s database eligible to take the traditional assessment. This baseline survey was used to determine study eligibility and collect demographic information. To be eligible, research participants were required to be active CRNAs practicing nurse anesthesia who did not plan to retire before the study ended. Additionally, participants were only eligible if they had not previously taken the traditional assessment and if they consented to be randomly assigned to either the traditional assessment group (control group) or the longitudinal assessment group (intervention group). They also had to sign a participation agreement that outlined the study procedures and requirements and complete post-assessment surveys to provide their feedback on their experience. The study was approved by the Western Copernicus Group Institutional Review Board (WCG® IRB) and determined to be exempt research.

Stratified sampling, randomization, and power analysis

A stratified random sampling was performed to identify 1,000 CRNAs from the call for volunteers and randomly assign them to one of two groups: 500 CRNAs in the traditional assessment group and 500 CRNAs in the longitudinal assessment group. Demographic characteristics of the respondents’ age distribution, gender, and years of practice experience collected from the baseline survey were used to match demographic data with the eligible CRNA population to ensure representativeness of the sample.

A power analysis indicated that a minimum of 320 participants was required to ensure that the results from the research study would achieve the desired confidence level. The analysis used a 1:1 allocation ratio with a decision to oversample to account for study attrition and those lost to follow-up. 500 participants were selected in each group to achieve an

alpha significance level of 0.05, statistical power of 0.80, and predicted effect size of 0.30 based on a mean difference of 5 points ($SD= 12.6$).

Study duration and requirements

The duration of the research study was one year, which started April 4, 2022, and ended on March 31, 2023. Participants in the traditional assessment group were required to schedule and take a three-hour, 150-question assessment in person at a Pearson VUE test center or online using a live remote proctor prior to the conclusion of the study.

Participants in the longitudinal assessment group were required to answer a total of up to 135 questions on demand over the span of four quarters within the following timeframes:

- Quarter 1 – April 4 to July 4, 2022 (91 days)
- Quarter 2 – July 5 to October 2, 2022 (89 days)
- Quarter 3 – October 3, 2022, to January 2, 2023 (91 days)
- Quarter 4 – January 3 to March 31, 2023 (87 days)

The longitudinal assessment group was required to answer 30 questions for the first quarter and then up to 35 questions for each subsequent quarter, which included up to five repeat administrations of questions (i.e., up to 11% of the total) previously answered incorrectly, left unanswered, or forfeited.

Questions for both the traditional and longitudinal assessment were developed using the same process and were drawn from the same item bank based on the CPCA content outline (NBCRNA, 2021) used to evaluate current anesthesia knowledge of CRNAs in the four core domains of nurse anesthesia practice. All questions were pre-calibrated to the Rasch measurement (logit) scale based on prior exam administrations. While both groups received different sets of questions from the same item bank, these scores were equated to a common Rasch scale.

The longitudinal assessment group participants were required to log into a web-based platform hosted by Internet Testing Systems (ITS) to complete the quarterly assessments. Although the ITS longitudinal

assessment platform could be accessed from a mobile/computing device with a stable internet connection, it was recommended to use a desktop or laptop computer for optimal performance during the research study. Longitudinal assessment participants were not able to skip questions or advance to another quarter's questions until the present quarter had concluded.

Longitudinal assessment participants were allotted 60 seconds to answer each question. A timer was visible showing time remaining once the question was launched. After each response was selected and submitted, participants were asked to rate their level of confidence in their response and the relevance of the question to their clinical practice.

After longitudinal assessment participants submitted their confidence and relevance ratings, the platform displayed the answer choice selected and whether it was correct or incorrect, along with the rationale and references. Longitudinal assessment participants were also given the option to mark the question as a favorite and submit question feedback. Previously answered questions and their associated responses were not visible to the longitudinal assessment participants due to item security considerations. Additionally, a digital watermark behind each question was used as an added security measure to prevent screen captures along with a warning message that would appear if a screenshot was detected.

Rationales and references related to all questions answered could be viewed and accessed by using the review option in the platform's navigation menu at any time. The rationales and references in the review page were organized by the four core domains and displayed according to the traditional assessment content outline (NBCRNA, 2021). The ITS longitudinal assessment platform included a dashboard that provided real-time assessment performance and progress information. The dashboard also provided normative/comparative performance data, which displayed how well participants' peers had done on similar questions.

Participants of this study were given a price reduction for the exam fee as an incentive and satisfied the assessment component of the CPC Program if they completed all requirements of the one-year study.

Outcome measures

Performance

The primary outcome measured in the study was assessment performance in the traditional assessment group compared to the cumulative performance of the longitudinal assessment group after completing four quarterly assessments. Performance on the two assessments was measured on the same logit scale using the Rasch Model. Since only the longitudinal assessment group were readministered up to five questions answered incorrectly previously or left unanswered starting the second quarter, the performance scores for longitudinal assessment participants were calculated using two methods: 1) analyzing only the initial attempt of each question; and 2) analyzing responses only on the most recent attempt on any questions administered or repeated when answered incorrectly.

Perceptions and attitudes

Perceptions among CRNA participants who completed the traditional and longitudinal assessment were measured and collected over the study duration at different time points. Participants in the traditional assessment group were prompted to take a post-exit survey immediately after completing the assessment. Traditional assessment participants were asked to rate several statements using a four-point agreement Likert scale, including overall satisfaction with the testing experience, ability to identify knowledge gaps, promotion of lifelong learning, assessment of core knowledge related to safe practice of nurse anesthesia, level of difficulty being assessed according to practice experience, staying current in nurse anesthesia, and helping to provide better care to patients. Additionally, participants were asked to indicate if they used any preparation materials and the number of hours spent per week studying/preparing for the assessment. An open-ended prompt was also provided at the end to leave any final comments.

Participants in the longitudinal assessment group were required to take quarterly post-assessment feedback surveys sent from SurveyMonkey after answering the questions in each quarter. Longitudinal assessment participants were asked to rate several statements using a four-point agreement Likert scale related to the user experience of the ITS longitudinal assessment platform, including ease of use, navigation,

tracking performance, helpfulness of the information visible on the platform's dashboard, appropriateness and frequencies of notification and reminders, convenience of answering questions, usefulness of rationales and references provided with answers after every question, and clarity of questions and appropriateness of ability level. Participants were also asked if they encountered any technical problems with the platform or had needed to contact customer support for technical assistance. Final comments on overall experience and suggestions for any changes to the longitudinal assessment platform were collected at the end of each quarterly survey.

Longitudinal assessment participants also completed a final survey sent from SurveyMonkey after their last quarterly assessment. Longitudinal assessment participants were asked to rate the same statements as participants in the traditional assessment group rated in their post-exit survey using a four-point agreement Likert scale, in addition to statements on the feasibility and acceptability of the longitudinal assessment format, most desirable configuration and important features, as well as whether they would recommend the longitudinal assessment format to a colleague.

Longitudinal Assessment Focus Groups. At the end of the longitudinal assessment final survey, participants were asked to participate in an optional focus group. Focus group participants were recruited based on responses to a question in the final survey that asked if they would be interested in participating and their availability. Focus groups were conducted using Zoom videoconferencing (Archibald et al., 2020) and were facilitated by an external moderator. The purpose of the focus groups was to better understand the experiences of participants in the longitudinal assessment group during the study. Participants were asked to offer their perceptions of the longitudinal assessment format's utility and value, the attributes that facilitated their participation and engagement, any barriers that may have limited or hindered their experience, and were asked to react to potential enhancements in a future configuration of the longitudinal assessment.

Participant Engagement and Usability. Participant behavior and engagement was measured using the data available from the ITS longitudinal assessment platform, including when and how

frequently over the study duration quarterly assessments were completed. Data on the longitudinal assessment platform's usability was collected in the final survey using the Systems Usability Scale (SUS, n.d.). The SUS is a 10-item questionnaire using a five-point agreement Likert scale, to measure overall usability using a composite score from 0 to 100. A score of 68 or more is considered above average and indicates that the platform is generally easy to use and navigate, and that users are likely to find it useful (SUS, n.d.).

Data analysis

Performance

Performance-based measures of the study for the traditional and longitudinal assessment groups were compared using the following statistical tests and analyzed using IBM SPSS Statistics Version 26 (IBM Corp, 2023): (1) a two-proportion z-test for the percentage of participants in both groups meeting the performance standard and (2) an independent samples t-test for comparing the performance of participants in the two groups. A paired t-test was also conducted to compare the mean scaled scores based on the first and most recent response on repeat questions within the longitudinal assessment group. Estimates of effect size to show the magnitude of the difference in performance were derived using Cohen's D (Lakens, 2013). Scores for both groups were psychometrically equated to a common scale using item response theory (IRT) methodology using the Rasch Model, then subjected to linear transformation to the scale of traditional assessment scores ranging from 300-900, with 450 representing the established traditional assessment performance standard.

Perceptions and attitudes

For the perception-based measures of the study, quantitative data was collected from surveys to better understand participants' perceptions using a Likert scale. Participants rated their level of agreement on several statements about their experience and attitudes using a four-point agreement Likert scale. The scale values ranged from 1 (Strongly Disagree) to 4 (Strongly Agree).

Confidence and relevance data were collected from the longitudinal assessment participants after they

answered each question in the ITS longitudinal assessment platform. A four-point confidence scale was used to measure the respondents' level of confidence in their responses. The scale values ranged from 1 (Not at all confident) to 4 (Highly confident). A four-point relevance scale was used to measure the respondents' perception of the relevance of the questions to their practice. The scale values ranged from 1 (Not at all relevant) to 4 (Highly relevant).

A Kruskal-Wallis nonparametric test was performed to compare the ratings between the traditional and longitudinal assessment groups. The test statistic for the Kruskal-Wallis (KW) test is the H-statistic, which is calculated by first finding the sum of the ranks for each group. A p-value less than 0.05 was used to reject the null hypothesis if there was a statistically significant difference between the medians of the groups.

Qualitative data collected from open-ended comments from the quarterly and final surveys were analyzed using Natural Language Processing with the RoBERTa-base sentiment model (Cardiff NLP, 2023). The RoBERTa-based CardiffNLP model was used to classify the comments into three distinct categories: Positive, Neutral, or Negative. These comments were processed into bigrams (two-word sequences that are most frequent) and trigrams (three-word sequences that are most frequent) to identify words and phrases associated with each sentiment category. Trigrams allow readers to understand not just what the subject matter is, but also where the sentiment is specifically directed. Analysis of bigrams and trigrams identifies the underlying sources of positive, neutral, and negative sentiments, facilitating extraction of emerging themes from the examinee comments. Qualitative data collected from the focus groups used thematic analysis to identify themes based on participants' reflections in the longitudinal assessment group.

CONSORT Map

A Consolidated Standards of Reporting Trials (CONSORT) map (Figure 1) was used to illustrate enrollment of research participants, their allocation to each arm, disposition status, and how they were analyzed in the randomized controlled study (Schulz et al., 2010).

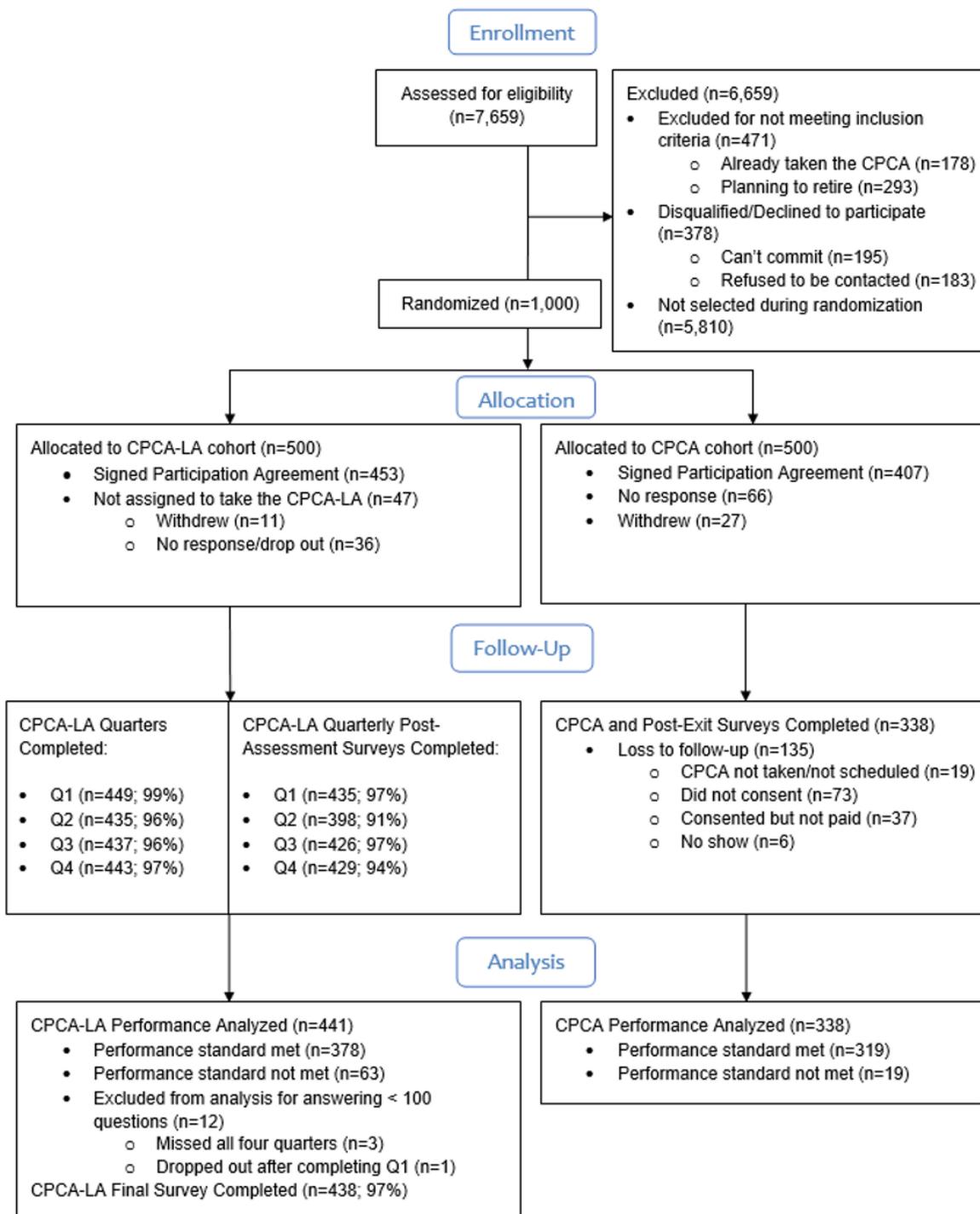
Results

Participant CONSORT Map

Figure 1 represents the research participants' group allocation and progression through the study. Out of the 43,722 active CRNAs eligible to take the traditional

assessment, 7,659 responded to the call for volunteer's baseline survey. One thousand CRNAs were selected and randomly assigned to one of two conditions: 500 CRNAs in the current traditional assessment group and 500 CRNAs in the longitudinal assessment group, matched 1:1 based on gender, age, and years of practice representative of the entire population.

Figure 1. LA Research Study CONSORT Map



Of the 500 CRNAs randomly selected in the longitudinal assessment group, 453 signed the participation agreement and were enrolled to access the ITS longitudinal assessment platform. Out of the 453 CRNAs in the longitudinal assessment group, 47 participants were no longer participating by the end of the study (36 were lost to follow-up and 11 withdrew from the study). Of the 500 CRNAs randomly selected in the traditional assessment group, 407 participants signed the participation agreement. Out of the 407 CRNAs in the traditional assessment group, 93 participants were no longer participating by the end of the study (66 were lost to follow-up and 27 withdrew from the study). Some reasons cited for withdrawing from the study were not being able to commit/participate during the study timeframe or other life events (e.g., military deployment, childbirth, marriage).

In the longitudinal assessment group, the average completion rate of the quarterly assessments was approximately 97%. The average completion rate of the post-assessment quarterly surveys was approximately 95%. The average completion rate for the final survey was 97%. Of the 453 longitudinal assessment participants, 12 were excluded from the analysis for answering fewer than 100 test questions, which was the minimum threshold for meaningful participation. The final analysis included 441 longitudinal assessment participants and 338 participants from the traditional assessment group.

Participant demographics and geographic locations

Table 1 shows the demographic characteristics of the research study participants by each group. In the longitudinal assessment group, there were 61% (n=270) females and 39% (n=171) males. In the traditional assessment group, there were 61% (n=207) females and 39% (n=131) males. The average number of years of practice was 15.3 years ($SD=10.0$) and 15.2 years ($SD=9.5$) in the longitudinal assessment and traditional assessment groups, respectively. The majority (31.5%) of longitudinal assessment participants ($M=47.6$, $SD=10.5$) were in the 40 to 49 age group, which was comparable to the percentage (31.7%) of traditional assessment participants ($M=47.8$, $SD=10.3$) in the same age group. The top five geographic locations of longitudinal assessment participants represented in the study were

Pennsylvania, Texas, Florida, Ohio, and Tennessee. The top five geographic locations of traditional assessment participants represented in the study were Pennsylvania, Texas, Ohio, Florida, and Illinois. These geographic locations are consistent with the known distribution of CRNA population.

The two groups in the study were comparable in terms of their overall demographic composition. Gender, the average age, and years of practice for the traditional assessment group were analogous to the longitudinal assessment group. There was also representation geographically across the US.

Performance comparison between the longitudinal assessment and traditional assessment participants

From the longitudinal assessment group, 378 out of 441 met the performance standard, while 63 did not meet the performance standard. From the traditional assessment group, 319 out of 338 met the performance standard, whereas 19 participants did not meet the performance standard (Figure 1).

When the results were first analyzed scoring the initial attempt from the longitudinal assessment group, 85.7% of the longitudinal assessment participants met the performance standard compared to 94.4% of the traditional assessment participants (Table 1). This difference in percentages of participants meeting the performance standard was significantly different ($X^2(1)=14.35$, $p<.001$). The mean score for the participants in the longitudinal assessment group ($M=562.5$, $SD=109.4$) was significantly lower ($t(763.4)=-5.1$, $p<.001$) than the mean of scaled scores for the traditional assessment group ($M=600.2$, $SD=95.8$). The effect size or magnitude of this difference between the means of the two groups (Lakens, 2013) was small (Cohen's $D=0.36$).

When comparing the percentage of participants that met the performance standard in the longitudinal assessment group after scoring the most recent responses on repeated questions that were previously answered incorrectly (91.8%), the finding was not significantly different ($X^2(1)=1.52$, $p=.218$) from the percentage of participants meeting the performance standard in the traditional assessment group (94.4%). However, the mean scaled score for the longitudinal assessment group ($M=649.4$, $SD=139.7$) after scoring the most recent response of repeated questions was significantly higher ($t(768.0)=5.8$, $p<.001$) than the

Table 1. LA Research Study Participant Demographics, Performance and Perceptions

Variable/statements	CPC-LA group		CPCA group
Females % (n)	61% (270)		61% (207)
Males % (n)	39% (171)		39% (131)
Mean Age (Standard Deviation)	47.6 (<i>SD</i> =10.5)		47.8 (<i>SD</i> =10.3)
• 30-39 % (n)	19.6% (115)		28.1% (95)
• 40-49 % (n)	31.5% (150)		31.7% (107)
• 50-59 % (n)	24.8% (96)		23.1% (78)
• 60-69 % (n)	20.5% (68)		14.5% (49)
• 70-79 % (n)	3.4% (10)		2.7% (9)
Average Years of Practice (Standard Deviation)	15.3 (<i>SD</i> =10.0)		15.2 (<i>SD</i> =9.5)
	Initial Attempt	Most Recent Attempt	
Percent Meeting Performance Standard Two-proportion z test comparison with the CPCA group (95% CI of difference from the CPCA group)	85.7% (378/441) $X^2(1) = 14.35, p<.001$ (4.3% to 13.0%)	91.8% (405/441) $X^2(1) = 1.52, p=.218$ (-1.3% to 6.3%)	94.4% (319/338)
Mean Scaled Score t-test comparison with the CPCA group (95% CI of difference from the CPCA group)	562.5 (<i>SD</i> =109.4) $t(763.36) = -5.12, p<.001$ (-52.2 to -23.2)	649.4 (<i>SD</i> =139.7) $t(768.02) = 5.82, p<.001$ (32.6 to 65.8)	600.2 (<i>SD</i> =95.8)
CPCA Post-Exit Survey & CPC-LA Post-Quarterly Assessment Average Ratings*			
1. Overall, I was satisfied with my testing experience.	3.4		3.4
2. The [CPCA/LA] helps me identify my knowledge gaps.	3.2		2.9
3. The [CPCA/LA] promotes lifelong learning.	3.1		2.6
4. The [CPCA/LA] accurately reflected core knowledge related to the safe practice of anesthesia.	3.1		2.9
5. The level of difficulty of the [CPCA/LA] was well-matched to my practice experience.	3.0		2.8
6. The [CPCA/LA] helps me stay current in nurse anesthesia.	3.0		2.5
7. The [CPCA/LA] helps me provide better care to my patients by helping me maintain my core nurse anesthesia knowledge.	2.8		2.5
8. On average, how many hours did you spend PER WEEK studying for the exam that you took today during the LAST 30 days?	% (n)		% (n)
A. 0	64.7% (286)		37.1% (124)
B. 1-4	31.4% (139)		41.9% (140)
C. 5-8	2.7% (12)		13.8% (46)

D. 9-12	1.1% (5)	2.7% (9)
E. 13-16	0% (0)	1.5% (5)
F. 17-20	0% (0)	3% (10)
G. >20	0% (0)	0% (0)
9. Did you use any of the following in preparation for this assessment? (Select all that apply)		
A. Professional review course	7.0% (31)	15.6% (71)
B. Practice exams available at the NBCRNA website	7.9% (35)	24.7% (112)
C. Continuing education resources available at the AANA website	11.8% (52)	17% (77)
D. Other	6.6% (29)	18.9% (86)
E. None of the above	55.4% (245)	23.8% (108)
F. Core Modules	30.1% (133)	NA
10. Most important features that should be available in the LA platform:		
A. Answering clinical scenario/case-based questions.	69.0% (305)	NA
B. Repeating items answered incorrectly based on the confidence and relevance ratings.	58.6% (259)	NA
C. Accessing the LA platform from a mobile app.	55.7% (246)	NA
D. Answering article-based questions via a "View Article" button.	30.1% (133)	NA
E. Showing/Hiding the timer displayed when answering questions.	29.0% (128)	NA
F. Displaying the time spent reviewing the correct answer and rationales in the dashboard.	13.1% (58)	NA
G. Other (please specify)	9.7% (43)	NA
H. None of the above	2.9% (13)	NA
I. Having the questions read out loud.	0.9% (4)	NA
ITS LA Platform Quarterly Post-Assessment Survey Average Ratings*		
11. Log-in process was easy	3.8	NA
12. Platform was easy to navigate without too much effort	3.8	NA
13. Easy to track performance using the dashboard	3.7	NA
14. Information displayed on the dashboard was helpful	3.7	NA
15. Frequency of notifications sent out were appropriate	3.7	NA
16. Process for answering the questions was convenient	3.6	NA
17. Rationales provided were useful for learning	3.6	NA
18. Information displayed on the Review page was helpful	3.5	NA
19. References with the answers to the questions were useful	3.4	NA
20. Questions were clearly written at the appropriate ability level	3.3	NA
ITS LA Platform Final Usability Survey Average Ratings*		

21. Completing 30-35 questions per quarter was feasible for my schedule.	3.6	NA
22. I would take the CPC-LA format again.	3.5	NA
23. Participating in the CPC-LA reduced my anxiety about maintaining my CRNA certification.	3.2	NA
24. The CPC-LA measured the intended knowledge of what is required by CRNAs.	3.1	NA
25. Participating in the CPC-LA increased my overall confidence in taking assessments.	3.0	NA
26. Participating in the CPC-LA increased my knowledge base in anesthesia.	2.9	NA
27. The CPC-LA helped change how I practice nurse anesthesia.	2.3	NA
ITS LA Final Survey System Usability Score (SUS) Ratings**		
28. I would use the CPC-LA platform more frequently.	1.9	NA
29. I found the CPC-LA platform unnecessarily complex.	4.1	NA
30. I thought the CPC-LA platform was easy to use.	2.4	NA
31. I think that I would need the support of a technical person to be able to use the CPC-LA platform again.	4.4	NA
32. I found the various functions in the CPC-LA platform to be well integrated.	2.1	NA
33. I thought there was too much inconsistency in the CPC-LA platform.	4.0	NA
34. I would imagine that most people would learn to use the CPC-LA platform very quickly.	2.3	NA
35. I found the CPC-LA platform very cumbersome to use.	4.3	NA
36. I felt very confident using the CPC-LA platform.	2.2	NA
37. I needed to learn a lot of things before I could get going with the CPC-LA platform.	4.3	NA
Total SUS Score	$(31.9 * 2.5) = 79.8$	NA

CI = Confidence Interval

CPCA = Continued Professional Certification Assessment

CPC-LA = Continued Professional Certification-Longitudinal Assessment

LA = Longitudinal Assessment

CPCA/LA = Same item asked in Continued Professional Certification Assessment post-assessment survey or in the Continued Professional Certification-Longitudinal Assessment final survey.

NA = Not applicable or data not available/collected

*Four-point agreement Likert scale: 1 = Strongly Disagree, 2 = Somewhat Disagree, 3 = Somewhat Agree, 4 = Strongly Agree

**0 = Strongly Disagree, 1 = Somewhat Disagree, 2 = Neither Disagree nor Agree, 3 = Somewhat Agree, 4 = Strongly Agree

mean for the traditional assessment group ($M=600.2$, $SD=95.8$). The effect size of this difference was small (Cohen's $D=0.40$).

Additionally, results from the paired t-test when comparing the mean scaled scores within the longitudinal assessment group showed that the mean scaled score when scoring the most recent response ($M=649.4$, $SD=139.7$) was significantly higher compared to the scaled scores based on the first response ($M=562.5$, $SD=109.4$), ($t(440)=44.4$, $p<.001$). Therefore, the longitudinal assessment group's performance significantly improved when considering the most recent response of a readministered question for scoring. The effect size of this difference was large (Cohen's $D=2.1$).

Perception comparison between the longitudinal assessment and traditional assessment participants

Table 1 also displays the average rating results from participants in both groups on several statements collected from the post-exit survey, longitudinal assessment quarterly surveys, and the longitudinal assessment final survey based on a four-point agreement Likert scale. Overall, both groups were satisfied with their testing experience regardless of the format in which they took the assessment, with an average rating of 3.4 out of a four-point scale. Participants in the longitudinal assessment group rated most other items slightly higher than participants in the traditional assessment group. Overall, participants were most satisfied with the longitudinal assessment in terms of its ability to help them identify knowledge gaps (3.2), promote lifelong learning (3.1), and accurately reflect core knowledge related to the safe practice of anesthesia (3.1). The lowest endorsed statement was "longitudinal assessment helps me provide better care to my patients by helping me maintain my core nurse anesthesia knowledge" (2.8). According to the KW test statistics, none of the differences observed were statistically significant.

Of participants in the traditional assessment group, 41% spent one to four hours per week studying for the exam according to the post-exit survey responses, while most participants in the longitudinal assessment group spent zero hours per week studying according to their responses in the final survey (Table 1). The top three most important features that should be available

in a longitudinal assessment platform as rated by respondents were answering clinical scenario/case-based questions (69.0% [$n=305$]), repeating items answered incorrectly based on the confidence and relevance ratings (58.6% [$n=259$]), and accessing the longitudinal assessment platform from a mobile application (55.7% [$n=246$]) (Table 1). A small number of respondents specified other features that they thought were important for a future iteration of a longitudinal assessment platform, such as the ability to customize the platform by showing/hiding the timer, the ability to track progress of time spent reviewing the rationales, and the ability to have the questions read aloud. Other options included more time to answer questions, partial credit for multiple-select questions or having no multiple-select questions at all, access to practice/warm-up questions within the longitudinal assessment platform, and the ability to look up definitions.

Longitudinal assessment participants in the final survey were asked to indicate which type of assessment format they would prefer in the future. Among the traditional assessment group, 53.9% of participants ($n=179$) said they would prefer a longitudinal assessment format, while only 46.1% of participants ($n=153$) said they would prefer the current traditional assessment. Among the longitudinal assessment group, 77.6% ($n=343$) responded that they would prefer the longitudinal assessment format, as opposed to only 5.2% ($n=23$) who favored the current traditional assessment. 17.2% ($n=66$) preferred other modalities or intervals such as taking the traditional assessment once every four years ($n=34$), every two years ($n=12$), or every other year ($n=5$).

Longitudinal assessment participants were asked to rank-order potential longitudinal assessment policy considerations on a three-point scale, with 1 being the MOST desirable and 3 being the LEAST desirable. The most desirable longitudinal assessment policy consideration was dropping the lowest quarter's scores, which was selected by 50.7% ($n=224$) of respondents, followed by 28.1% ($n=124$) of respondents who selected skipping a set number of questions while taking the longitudinal assessment without being penalized, and 21.3% ($n=94$) who selected electing time off by not being required to answer questions for a quarter.

Participant engagement and feedback about the longitudinal assessment platform

Longitudinal assessment Quarterly Post-Assessment Survey Results. On average, longitudinal assessment participants rated the ITS platform 4.3 out of five stars in the quarterly post-assessment surveys and final survey. The overall average ratings across the four quarters for several specific statements were above three on a four-point agreement Likert scale (Table 1). Based on the results presented in Table 1, longitudinal assessment participants generally found the platform easy to use and navigate. The average rating across all statements was 3.6, which indicates that users were generally satisfied with the platform. The highest-rated questions were “Log-in process was easy” (3.8) and “Platform was easy to navigate without too much effort” (3.8). The lowest-rated questions were “References with the answers to the questions were useful” (3.4) and “Questions were clearly written at the appropriate ability level” (3.3), which were still rated above a three out of a four-point Likert scale. The average rating for each question generally increased from Q1 to Q4, which suggests that users' satisfaction with the platform increased over time, whereas there was a slight decrease in some of the average ratings for statements 16-20 (Table 1).

Longitudinal Assessment Platform Final Usability Survey Ratings. Overall, participants found the longitudinal assessment to be a feasible and useful way to maintain their CRNA certification (Table 1). The average rating for all statements was 3.1, which indicates that participants generally endorsed overall agreement with the platform. The highest-rated questions were “Completing 30-35 questions per quarter was feasible for my schedule” (3.6) and “I would take the longitudinal assessment format again” (3.5). The lowest-rated questions were “Participating in the longitudinal assessment increased my knowledge base in anesthesia” (2.9) and “The longitudinal assessment helped change how I practice nurse anesthesia” (2.3). Additionally, 95% of participants indicated that they would recommend the platform to their colleagues.

Longitudinal Assessment Platform Usability Score. The final survey included the Systems Usability Scale (SUS) to measure overall usability based on a score from 0 to 100, with a score of 68 or more considered above

average. The ITS longitudinal assessment platform's SUS score was approximately 80 (Table 1).

Longitudinal Assessment Platform Engagement.

User Interaction. The completion of questions by weeks remaining in the quarter and the duration spent answering questions for the longitudinal assessment group was analyzed. The number of participants completing the quarters remained steady throughout the year, with a slight increase in the number of participants completing Q4 in the last week before the quarter closed. Additionally, most longitudinal assessment participants completed the quarterly assessments in less than an hour over the course of the study, answering their questions in batches or in a single sitting, rather than starting and returning or completing them over multiple sessions.

Confidence and Relevance Ratings. The results from the post-response ratings, provided after answering each question, indicated that participants generally found the questions to be reasonably relevant and were confident in their answers. Longitudinal assessment participants' mean ratings on the relevance of items to practice and confidence in their responses were 2.8 and 2.6, respectively, on a four-point scale (Table 2).

The average relevance ratings from Q1 to Q4 were stable at 2.8 or 2.9 and the average confidence ratings were steady at 2.6.

The results presented in Table 2 indicate that participants who answered the question correctly in the longitudinal assessment study rated relevance and confidence slightly higher on average than participants who did not answer the question correctly. The mean relevance rating for participants who answered the question correctly was 2.9, while the mean relevance rating for participants who did not answer the question correctly was 2.6. The mean confidence rating for participants who answered the question correctly was 2.7, while the mean confidence rating for participants who did not answer the question correctly was 2.2.

Additionally, Table 2 shows that the longitudinal assessment participants who met the performance standard found the questions to be slightly more relevant and were more confident in their answers than those who did not meet the performance standard. The mean relevance rating for those who met the performance standard was 2.9, and the mean confidence rating was 2.6. For those who did not

Table 2. CPC-LA Confidence/Relevance Ratings and Time Spent (Seconds)

Average Ratings*** and Time		N	Mean	SD
Relevance ratings	Quarter 1	13,068	2.9	0.9
	Quarter 2	14,872	2.8	0.9
	Quarter 3	14,998	2.8	0.9
	Quarter 4	15,188	2.8	0.9
	Total	58,126	2.8	0.9
Confidence ratings	Quarter 1	13,069	2.6	0.9
	Quarter 2	14,872	2.6	0.9
	Quarter 3	14,998	2.6	0.9
	Quarter 4	15,188	2.6	0.9
	Total	58,127	2.6	0.9
Relevance ratings	Answered question incorrectly	16,518	2.6	0.9
	Answered question correctly	41,608	2.9	0.9
	Total	58,126	2.8	0.9
Confidence ratings	Answered question incorrectly	16,519	2.2	0.9
	Answered question correctly	41,608	2.7	0.9
	Total	58,127	2.6	0.9
Relevance ratings	Did not meet performance standard	7,914	2.4	0.9
	Met performance standard	50,212	2.9	0.9
	Total	58,126	2.8	0.9
Confidence ratings	Did not meet performance standard	7,914	2.2	0.9
	Met performance standard	50,213	2.6	0.9
	Total	58,127	2.6	0.9
Relevance ratings	Did not time-out	56,673	2.8	0.9
	Timed-out	1,453	2.6	0.9
	Total	58,126	2.8	0.9
Confidence ratings	Did not time-out	56,674	2.6	0.9
	Timed-out	1,453	1.9	0.9
	Total	58,127	2.6	0.9
Time Answering Questions (Seconds)	Quarter 1	13,230	25.8	15.0
	Quarter 2	15,303	25.0	15.1
	Quarter 3	15,341	24.9	15.1
	Quarter 4	15,375	24.8	14.7
	Total	59,249	25.1	15.0
Time Reviewing Rationales (Seconds)	Quarter 1	13,230	27.3	102.0
	Quarter 2	15,303	35.4	1,499.3
	Quarter 3	15,341	28.3	555.9
	Quarter 4	15,375	17.1	142.1
	Total	59,249	27.0	817.4

Time Answering Questions (Seconds)	Did not meet performance standard	8,505	23.6	14.6
	Met performance standard	50,744	25.3	15.0
	Total	59,249	25.1	15.0
Time Reviewing Rationales (Seconds)	Did not meet performance standard	8,505	22.1	226.3
	Met performance standard	50,744	27.9	878.4
	Total	59,249	27.0	817.4

***Confidence scale: 1 = Not at all confident, 2 = Somewhat confident, 3 = Confident, 4 = Highly confident
 Relevance scale 1 = Not at all relevant, 2 = Somewhat relevant, 3 = Relevant, and 4 Highly relevant

meet the performance standard, the mean relevance rating was 2.4, and the mean confidence rating was 2.2. The difference in relevance and confidence ratings between those who met the performance standard and those who did not was consistent across all four quarters.

Table 2 also depicts the average time spent answering questions, which remained relatively consistent across all four quarters, ranging from 25.8 seconds per question in Q1 to 24.8 seconds in Q4. The average time spent answering questions was 25.3 seconds for those who met the performance standard, compared to 23.6 seconds for those who did not meet the standard. The average time spent reviewing rationales increased from 27.3 seconds in Q1 to 35.4 seconds in Q2. This increase was followed by a decrease in Q3 and Q4.

Longitudinal Assessment Open-ended Comments.

A total of 4,611 comments were analyzed using PyTorch, Cpython 3.10, and RoBERTa (CardiffNLP, 2023). Out of the analyzed comments, 668 were categorized as positive, 3,475 as neutral, and 468 as negative. The most common bigrams for positive, neutral, and negative sentiments were “immediate feedback” (n=31), “answer questions” (n=131), and “answer questions” (n=48), respectively. To provide additional context to the bigrams, trigrams were analyzed, with the most common expressions being “liked immediate feedback” (n=6) for positive, “time answer questions” (n=52) for neutral, and “multiple answers questions” (n=19) for negative sentiments. While the bigram “answer questions” is somewhat ambiguous, the trigrams clarify the focus. For instance, “time answer questions” suggests that the context is about the time involved in answering, whereas “multiple answers questions” indicates that the issue

lies in having to select multiple answers/responses for a question.

Longitudinal Assessment Focus Group Emerging Themes. Longitudinal assessment participants who expressed interest in participating in a focus group were recruited. A total of 25 CRNAs accepted the invitation to participate. Out of the 25 recruited participants, 21 had successfully met the performance standard and four had not met the performance standard. The number of longitudinal assessment participants who attended each session was eight for the first focus group session, six for the second session, and six for the third session, for a total of 20 participants, of which 70% (n=14) were females and 30% (n=6) were males. The average years of practice among the longitudinal assessment focus group participants was sixteen. The majority age group was 50-59, (n=8), followed by 31-39 (n=5), 40-49 (n=4), and 60 or older (n=3).

Focus group participants were asked to offer their perceptions of the longitudinal assessment format’s utility and value, as they were probed across a range of considerations. Along with moderated discussions about longitudinal assessment elements that fostered their participation and engagement, and the longitudinal assessment elements that may have limited or hindered their experience, the focus group participants were also asked to react to potential enhancements in a future configuration of the longitudinal assessment. The following is a summary of the themes that emerged:

Positive Experiences.

- **Convenience:** Participants appreciated being able to take the longitudinal assessment on demand and on their own schedule, at their own pace, and in a location of their choosing.

- **Immediate feedback:** Participants appreciated receiving immediate feedback on their answers, which they felt could help them improve their learning and retention.
- **Rationales:** Participants appreciated the rationales provided for both correct and incorrect answers, which they felt could help them learn from their mistakes.
- **Non-punitive:** Participants appreciated that the longitudinal assessment was non-punitive, meaning that they would not lose their certification if they did not pass.
- **Appropriate and applicable questions:** Participants felt that most of the questions were appropriate and applicable to practice.

Negative experiences.

- **1-minute time limit per question:** Some participants had strong negative reactions to the 60-second time limit per question, particularly when the question was a multiple response item, which for some were more complex than other questions.
- **Need to confirm/click “Proceed” to see next question:** Some participants found it frustrating to have to confirm their acceptance of the time limit and question type before every item.
- **Lack of awareness of review functions and accessibility:** A substantial number of participants were unaware that the ITS longitudinal assessment platform provides certain review functions, such as the ability to review content areas and access rationales after initially answering questions. Additionally, a substantial number of these participants had been unaware that they could access the longitudinal assessment in multiple sessions during a quarter, rather than all at once in a single session.
- **Questions on topics where participants are no longer active:** Some participants expressed concern about being tested on topics in clinical specialty areas in which they no longer actively practice, such as someone who has specialized in pediatric anesthesia since their initial

certification but still must answer questions about geriatric anesthesia practice in the longitudinal assessment.

- **Length and number of questions:** Some participants expressed concern that any future longitudinal assessment configuration should not include too many questions or be too time-consuming.

Overall Preferences.

- Despite the areas where they would like to see improvement, participants expressed high praise for many aspects of the longitudinal assessment. When asked which approach they would prefer as the future method for continuing certification, they universally said they would prefer the longitudinal assessment to the current traditional assessment.

Discussion

Performance

The reported performance scores for the longitudinal assessment group were first analyzed based on only initial question attempts and used to determine achievement of the performance standard. When calculating the scores this way, the study found a statistically significant difference in performance between the two groups. The traditional assessment group attained the established performance standard at a higher rate (94.4%) than the longitudinal assessment group (85.7%). The performance difference observed in the mean scaled scores was also significant, favoring the traditional assessment group, with the effect size observed as small to moderate.

However, upon follow-up analysis scoring the most recent responses on repeat question attempts, longitudinal assessment participants were revealed to have improved their performance when scoring the readministered questions, after incorrectly answering them on the first attempt. This finding suggests that study participants were able to learn from previous errors and apply that knowledge in subsequent questions. This result provides evidence for the learning value that longitudinal assessment provides and is consistent with related research (Brown & McDaniel, 2014; Dion et al., 2022; Turner et al., 2019;

Favier, van der Vleuten, & Ramaekers, 2017; Schuwirth, & van der Vleuten, 2012).

The moderately lower performance observed among the longitudinal assessment group when calculating the scores only using the initial question attempt could be explained by the lack of studying or preparation for the exam. In contrast to a formal, traditionally standardized, secure, point-in-time examination for which candidates may spend hours in preparation, longitudinal assessment formats are a more informal method that might be expected to foster a fundamentally different mindset in the participant. First, longitudinal assessment formats allow examinees to take the test on demand in a relaxed, comfortable setting, such as their office or home. This can be a major benefit for candidates who have difficulty traveling to a test center or who prefer to take the test in a familiar environment. Second, longitudinal assessment formats can be more flexible in terms of time because of the intermittent nature of a periodic assessment, and enable answering questions in batches or in a single sitting. Moreover, because longitudinal assessments are designed to be formative, performance is in part fulfilled by participation, so that studying for the longitudinal assessment becomes less of a focus than for a singular, high-stakes event involving months of build-up. This is evidenced in part by the fact that most longitudinal assessment participants reported spending zero hours studying for the assessment. This may have contributed to their lower scores under the initial score calculations.

Despite reduced amounts of focused study in the longitudinal assessment group, the formative elements of periodic, repeated assessments with interleaving topics appear to have had a similar effect to that of focused study. This is demonstrated by the similar pass rates between the traditional exam versus longitudinal assessment groups after counting the most recent responses on repeated questions. This confirms fulfillment of the longitudinal assessments' intended purpose of both measuring and enhancing knowledge and learning over time.

While the results of the study provide evidence substantiating the learning value proposition of longitudinal assessment, it is advisable to be aware of some of the potential drawbacks. First, longitudinal assessment formats require more time commitment than conventional assessments, simply because they

comprise more total questions spaced out over time in comparison to taking a traditional assessment in one sitting. Additional cognitive loads and test fatigue may also be involved with review and processing of question rationales and normative performance data. In addition, longitudinal assessment formats can be more challenging from a time-management perspective. Test-takers may need to be more disciplined to complete the test on time, especially given the pause-return-and-resume capability of longitudinal assessment, and to avoid distractions if the questions are accessed via mobile devices. According to user behavior data from longitudinal assessment participants, 28% accessed the platform from their mobile devices, which suggests that they could have answered questions in distracting environments and/or multi-tasking. Further research is warranted to better understand how user interaction and engagement impact performance, and the degree of learning by healthcare professionals using longitudinal assessment.

Perceptions

The survey results suggest that CRNAs are in favor of a more continuous assessment format over the more traditional singular traditional assessment administration, which is supported by the higher agreement rating with the statement that longitudinal assessment promotes lifelong learning. The authors believe that a longitudinal assessment format would allow CRNAs to demonstrate and develop their knowledge and skills over time, rather than providing only a singular opportunity to demonstrate proficiency every eight years.

Survey results also imply that users were generally satisfied with the longitudinal assessment format and found it to be a feasible way to maintain their CRNA certification because a longitudinal assessment format would be more flexible and accommodating than conventional continued certification testing for CRNAs with busy clinical schedules. The platform was easy to use and navigate, and users found the information displayed on the dashboard to be helpful. Additionally, participants in the longitudinal assessment group preferred this format over the current traditional assessment format.

However, longitudinal assessment participants on average expressed comparatively less agreement with the impact of this format in improving their knowledge

base or changing how they practice nurse anesthesia. The latter is perhaps not surprising since the traditional assessment and longitudinal assessment are both designed to assess core knowledge of the field, as opposed to knowledge of recent developments or topical (emergent) knowledge. That said, one of the elements of a longitudinal assessment modality is the ability to incorporate questions on emerging topics more readily than on conventional assessments (Rottman et al., 2023).

The final longitudinal assessment survey results also suggest that participants are concerned about their overall performance and show a preference for dropping their lowest quarter's scores. Interestingly, upon probing deeper on this topic during the focus groups, we learned that participants believe dropping the lowest quarter has no benefit because of the time invested in answering questions, and that it may in fact negate the intent of lifelong learning.

Longitudinal Assessment Platform Engagement

Most of longitudinal assessment participants answered their quarterly questions in a single session, rather than completing them over multiple sessions. According to the focus group findings, some of this behavior might be explained by participants not having known that they could take the assessment more periodically during multiple sessions instead of waiting until the quarter end date and taking it in one session. However, since survey results indicated that most participants completed their assessments in less than one hour, it is also likely that it was simply more convenient for many to answer all 30-35 questions at one time.

Participants generally found the questions to be reasonably relevant and were confident in their answers. Participants who timed out (i.e., no answer was submitted within the 60-second time limit), in the longitudinal assessment group, had lower relevance and confidence ratings than participants who did not time out overall. Offering a 60-second time limit to answer each question appeared to be adequate, despite it being a point of contention left by longitudinal assessment participants in the open-ended prompt. Of all question responses, 96.4% were submitted within the 60-second time limit. The average time spent answering questions was 25.3 seconds for those who met the performance standard, compared to 23.6 seconds for those who did not meet the standard. This

suggests that some who met the standard spent slightly longer answering questions, whereas some who did not meet the standard may have rushed through their answers or may have not read the questions carefully enough. It is important to note the need to balance satisfaction with the 60-second time limit per question, with the need to incorporate measures to maximize item security and to promote a realistic measure of inherent knowledge, rather than knowledge obtained using external sources. The 60-second time limit per question helps support both goals.

Limitations

There were various limitations to this study. One study limitation identified was the use of a self-selected sample, which could mean participants may not have been entirely characteristic of the general CRNA population. Self-selection may represent an early adopter phenomenon seen across industries (Kaminski, 2011). CRNAs who were more motivated to participate in the research study may have been promoters or early adopters of the longitudinal assessment format, and as such were more likely to volunteer. Additionally, some CRNAs may have only expressed interest due to the financial inducement of the reduced assessment application fee. Conversely, those who did not volunteer, and were harder to reach or recruit, may have offered a distinct perspective as detractors of the longitudinal assessment format.

While the study involved self-selection inherent to voluntary participation, multiple methods were implemented to mitigate the potential impact on generalizability, which included stratified random sampling, employing a random assignment to either assessment condition, and exceeding the minimum sample size in each group as indicated by the power analysis. It is surmised that the higher attrition and lower participation rates observed in the traditional group reflects loss of interest due to the nuisance of having to schedule time off and travel to take the traditional assessment, as well as the lack of flexibility and novelty when compared with the longitudinal version.

Furthermore, the study was conducted over a condensed time period of only one year. This may not be enough time to see the full long-term effects of the longitudinal assessment format on CRNA knowledge and may not necessarily be indicative of performance over a longer period. Participants in the longitudinal

assessment group did not need to take the traditional assessment format subsequently, and vice versa, and therefore when asked to indicate their preferred assessment modality did not have a frame of reference to make the comparison. Participants were only asked to take one assessment format to avoid response bias (particularly memory effects) and attrition bias.

Conclusions

This mixed-methods research study, which may represent the first known randomized controlled study comparing a continuous, longitudinal assessment with a traditional, conventional high-stakes assessment in healthcare professional certification, possesses both quantitative and qualitative data that was collected and triangulated. The study represents a methodical investigation into the comparability, feasibility, and usability of a longitudinal assessment in place of a traditional assessment. Findings indicate that when analyzing the most recent responses on repeat questions for the longitudinal assessment group, the proportion of candidates meeting the performance standard was not significantly different statistically from the traditional assessment group. However, the longitudinal assessment group showed a higher mean scaled score than the traditional assessment group. Additionally, the longitudinal assessment group demonstrated improvement in performance when scoring the most recent response on repeated questions previously answered incorrectly compared to scoring the initial response.

In general, both groups indicated satisfaction with their assessment experience regardless of the format in which they took the assessment. Participants in the longitudinal assessment group rated most other items slightly higher than participants in the traditional assessment group. Participants were generally satisfied with the ITS longitudinal assessment platform, finding it easy to use and navigate. Furthermore, the majority of the longitudinal assessment program requirements and elements identified by Giron et al. (2021) were used to inform the design of the ITS longitudinal assessment platform for the research study and are further being considered to determine the implications for development and maintenance of an enduring longitudinal assessment program.

Longitudinal assessment is a continuous assessment format that allows healthcare professionals to demonstrate their knowledge and skills over a

continuum, rather than a cross-sectional assessment at a single-point-in-time. This may prove beneficial for healthcare professionals with busy patient care schedules, as it allows them to take the assessment on demand at their own pace and time, and enables more flexibility since they do not have to wait or take time off from providing crucial patient care to schedule the assessment at a test center or remotely with an online proctor. Because longitudinal assessment is based on adult learning principles designed to promote learning, the promising results from this study show longitudinal assessment as a valuable tool to reinforce healthcare knowledge.

Overall, the results of this research study support that the longitudinal assessment is a feasible, usable, and acceptable method to maintain healthcare professional certification, as well as to promote lifelong learning. Further research and secondary analyses should be undertaken to explore the factors affecting performance, engagement, and learning optimization for longitudinal assessment participants. Additionally, further research inquiries could be undertaken to determine effective assessment security strategies (e.g. forensic data analysis to safeguard intellectual property), cost-benefits and effectiveness, and the long-term impact of longitudinal assessment on knowledge trajectory and healthcare practice.

References

- American Board of Medical Specialties (ABMS). (2019). Continuing Board Certification: Vision for the Future Commission Final Report. Retrieved from https://www.abms.org/wp-content/uploads/2020/11/commission_final_report_20190212.pdf
- American Board of Medical Specialties (ABMS). (2022). Conceptual foundations for designing continuing certification assessments for physicians. Retrieved from <https://www.abms.org/wp-content/uploads/2022/07/conceptual-foundations-continuing-certification-assessments-for-physicians.pdf>
- American Board of Medical Specialties (ABMS). (2023). All ABMS Member Boards Now Offer Formative Assessments. ABMS Newsroom. <https://www.abms.org/newsroom/all-abms->

[member-boards-now-offer-formative-assessments/](#)

- Archibald, M. M., Ambagtsheer, R. C., Casey, M. G., & Lawless, M. (2020). Using Zoom videoconferencing for qualitative data collection: Perceptions and experiences of researchers and participants. *Qualitative Research in Psychology*, 17(3), 354-371. <https://doi.org/10.1080/14780887.2019.1697959>
- Brown, P. C., Roediger, H. L. III, & McDaniel, M. A. (2014). *Make it stick: The science of successful learning*. Belknap Press of Harvard University Press.
- CardiffNLP. (2023). Twitter-RoBERTa-base for Sentiment Analysis. Retrieved from <https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment?text=I+like+you.+I+love+you>.
- Culley, D. J., Sun, H., Harman, A. E., & Warner, D. O. (2013). Perceived value of board certification and the maintenance of certification in anesthesiology program (MOCA®). *Journal of Clinical Anesthesia*, 25(1), 12–19. <https://doi.org/10.1016/j.jclinane.2012.09.001>
- Dion, V., St-Onge, C., Bartman, I., Touchie, C., & Pugh, D. (2022). Written-based progress testing: A scoping review. *Academic Medicine: Journal of the Association of American Medical Colleges*, 97(5), 747–757. <https://doi.org/10.1097/ACM.00000000000004507>
- Favier, R. P., van der Vleuten, C. P. M., & Ramaekers, S. P. J. (2017). Applicability of progress testing in veterinary medical education. *Journal of Veterinary Medical Education*, 44(2), 351–357. <https://doi.org/10.3138/jvme.0116-008R>
- Fry, E. T., Kuvin, J., & Sibley, J. (2023). Maintenance of Competence in Cardiovascular Practice: It's Time for More Learning, Less Testing. *Journal of the American College of Cardiology*, 81(9), 924-927. <https://doi.org/10.1016/j.jacc.2023.01.009>
- Giron, S. E., Dishman, D., McMullan, S. P., Riel, J., Newcomer, T., Spence, D., & Choudhry, S. A. (2021). Longitudinal assessment: A strategy to improve continuing professional certification. *Journal of Professional Nursing*, 37(6), 1140–1148. <https://doi.org/10.1016/j.profnurs.2021.09.002>
- Griffis, C. A., Dishman, D., Giron, S. E., Ward, R. C., & McMullan, S. P. (2022). Concept analysis of longitudinal assessment for professional continued certification. *Nursing Forum*, 57(2), 311–317. <https://doi.org/10.1111/nuf.12678>
- IBM Corp. (2019). *IBM SPSS Statistics for Windows, Version 26.0*. Armonk, NY: IBM Corp.
- Kaminski, J. (2011). Diffusion of innovation theory. *Canadian Journal of Nursing Informatics*, 6(2), 1-6. Retrieved from <https://cjni.net/journal/?p=1444>
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4, 863.
- National Board for Respiratory Care (NBRC). (2023). *Credential Maintenance Program (CMP)*. <https://www.nbrc.org/credentialed-practitioners/#credential-maintenance>
- National Commission on Certification of Physician Assistants (NCCPA). (2023). *NCCPA Announces Permanent Alternative to PANRE, PANRE-LA*. <https://www.nccpa.net/news/panre-la/>
- National Board of Certification & Recertification for Nurse Anesthetists (NBCRNA). (2021). *CPC Assessment Content Outline*. Retrieved from https://www.nbcrna.com/docs/default-source/continued-certification/cpc-toolkit/cpc-assessment-content-outline.pdf?sfvrsn=8d1c23ca_40
- Price, D., Biernacki, H., & Nora, M. (2018). Can maintenance of certification work? Associations of MOC and improvements in physicians' knowledge and practice. *Academic Medicine*, 93(12), 1872-1881. <https://doi.org/10.1097/ACM.0000000000002338>
- Price, D. W., Swanson, D. B., Irons, M. B., & Hawkins, R. E. (2018). Longitudinal assessment s in continuing specialty certification and lifelong learning. *Medical Teacher*, 40(9), 917–919. <https://doi.org/10.1080/0142159X.2018.1471202>

- Reid, R., Duffy, E., Cohen, C., & Friedberg, M. (2018). Identification of alternative physician assistant recertification models. RAND Corp. https://www.rand.org/pubs/research_reports/R2455.html
- Rottman, B. M., Caddick, Z. A., Nokes-Malach, T. J., & Fraundorf, S. H. (2023). Cognitive perspectives on maintaining physicians' medical expertise: I. Reimagining maintenance of certification to promote lifelong learning. *Cognitive Research: Principles and Implications*, 8(1), 1-15. <https://doi.org/10.1186/s41235-023-00496-9>
- Schulz, K. F., Altman, D. G., & Moher, D. (for the CONSORT Group). (2010). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMJ*, 340, c332. doi:10.1136/bmj.c332
- Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2012). The use of progress testing. *Perspectives on Medical Education*, 1(1), 24–30. <https://doi.org/10.1007/s40037-012-0007-2>
- Spence, D., Ward, R., Wooden, S., et al. (2019). Use of resources and method of proctoring during the NBCRNA Continued Professional Certification Assessment: Analysis of outcomes. *Journal of Nursing Regulation*, 10(3), 37–46. [https://doi.org/10.1016/S2155-8256\(19\)30147-4](https://doi.org/10.1016/S2155-8256(19)30147-4)
- Spence, D., Wicks, T., Wojnakowski, M., & Plaus, K. (2021). Benchmarking study on continuing certification in health care: Program variables, commonalities and trends. *Journal of Nursing Regulation*, 12(2), 34–40. [https://doi.org/10.1016/S2155-8256\(21\)00054-5](https://doi.org/10.1016/S2155-8256(21)00054-5)
- Sun, H., Zhou, Y., Culley, D. J., Lien, C. A., Harman, A. E., & Warner, D. O. (2016). Association between participation in an intensive longitudinal assessment program and performance on a cognitive examination in the maintenance of certification in anesthesiology program®. *Anesthesiology*, 125(5), 1046–1055. <https://doi.org/10.1097/ALN.0000000000001301>
- Turner, A. L., Olmsted, M., Smith, A. C., Dounoucos, V., Bradford, A., Leslie, L. K., (2019). Pediatrician perspectives on learning and practice change in the MOCA-Peds 2017 pilot. *Pediatrics*, 144(6). <https://doi.org/10.1542/peds.2019-2305>
- Usability.gov. (n.d.). System Usability Scale (SUS). Retrieved March 8, 2023, from <https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html>
- Ward, R., Baker, K., Spence, D., Lenard, C., Sapp, A., & Choudhry, S. (2023). Longitudinal assessment to evaluate continued certification and lifelong learning in healthcare professionals: A scoping review. *Evaluation & the Health Professions*. [Advance online publication]. <https://doi.org/10.1177/01632787231164381>

Citation:

Choudhry, S. A., Muckle, T. J., Gill, C. J., Chadha, R., Urosev, M., Ferris, M., & Preston, J. C. (2024). Transforming assessments of clinician knowledge: A randomized controlled trial comparing traditional standardized and longitudinal assessment modalities. *Practical Assessment, Research, & Evaluation*, 29(7). Available online: <https://doi.org/10.7275/pare.2028>

Corresponding Author:

Shahid A. Choudhry

National Board of Certification and Recertification for Nurse Anesthetists

Email: schoudhry [at] nbcrna.com

ⁱ All authors are employed by the National Board of Certification and Recertification for Nurse Anesthetists where Shahid A. Choudhry, PhD is the Director of Research and Evaluation; Timothy J. Muckle, PhD, ICE-CCP is the Chief Assessment Officer; Christopher J. Gill, PhD, MBA, CRNA, ACNPC-AG, FACHE is the Chief Credentialing Officer; Rajat Chadha, PhD is the Director of Psychometrics; Magnus Urosev, MEd is the Data Scientist; Matt Ferris, MA, MBA, CAE, ELS is the Senior Director of Testing; and John C. Preston, DNSc, CRNA, APRN, FAANA, FNAP, FAAN is the Chief Executive Officer.