



Improving the Accuracy and Interrater Reliability of the American Society of Anesthesiologists Physical Status Classification System

Thomas Pallaria¹, Nicholas Barone^{2*}, Boris Porshnev³ and Maureen McCartney-Anderson⁴

¹Program Director and Assistant Professor, Rutgers School of Nursing Anesthesia Program, USA

²Department of Nurse Anesthetist, Rutgers University, USA

³Doctor of Nursing Practice in Anesthesia, Rutgers University, USA

⁴Assistant Professor in the Nurse Anesthesia Program, Rutgers University, USA

*Corresponding author: Dr. Nicholas Barone, DNP, APN/CRNA, Department of Nurse Anesthetist, Rutgers University, 65 Bergen Street, Newark, New Jersey, 07107, USA, Tel: 718-644-1128; Email: N.Barone@Rutgers.edu

Received Date: July 12, 2019; Published Date: July 27, 2019

Abstract

Purpose: To determine if a more objective ASA-Physical Status (PS) classification system would improve the accuracy and interrater reliability of anesthesia providers in using this tool.

Design: Questionnaire-based pretest-posttest study.

Methods: Study participants assigned PS scores to five realistic patient case scenarios before and after the study intervention. Statistical analysis included a repeated measures ANOVA followed by pairwise comparisons between each group, a paired t-test to measure the overall effect of the study intervention, and a Fleiss' Kappa calculation to measure interrater reliability.

Findings: The average pretest score for all anesthesia providers was 72.50% (SD, 16.650) with an average posttest score of 95.25% (SD, 9.137). The difference between pretest scores and posttest scores as a whole was statistically significant ($t = 14.77$; $df = 79$; $p = .000$). Before the study intervention, there was poor to fair interrater reliability between anesthesia providers ($k = .301$, $p < 0.001$) and after the implementation of the authors updated and revised PS classification tool, there was excellent interrater reliability ($k = .911$, $p < 0.001$).

Conclusions: The increased objectivity of the revised PS classification system improved the overall accuracy and interrater reliability among anesthesia providers.

Keywords: ASA-PS; Physical status; Classification system; Interrater reliability

Abbreviations: PS: Physical Status; ASA: American Society of Anesthesiologists; SRNAs: Student Registered Nurse Anesthetists

Introduction

The preoperative period is a crucial time for providers to assess the patient's physical status, their overall health and ability to endure a specific anesthetic. This assessment period has important implications and is used to guide the anesthetic plan, resource allocation, billing practices, and communication between anesthesia providers and surgical staff. In order to collect and compare statistical data based on patient physical status, Meyer Saklad [1] published the first physical status (PS) classification system which was edited and updated in 1963 by Robert D. Dripps [2]. This classification system was embraced by the American Society of Anesthesiologists (ASA) and is still used today by anesthesia providers around the world in the perioperative period. The PS classification system was originally proposed as a method for the collection and tabulation of statistical data in anesthesia to record the overall health status of a patient prior to surgery. The original version of the classification system contained case examples for each category but were subsequently removed in the updated 1963 version, igniting criticism for its subjectivity and poor interrater reliability [3-7]. It has become apparent through multiple studies that the definition-only format of the PS classification system provides inadequate information for anesthesia providers to issue accurate, consistent, and reliable PS scores for the patients they encounter. Thus, two providers assessing the same patient may assign different PS scores. The lack of interrater reliability would not be a problem if this scale was not so ingrained into anesthetic practice during the perioperative period [8]. The authors propose that the reintroduction of specific case examples to the PS classification system will improve its objectivity and interrater reliability.

Background

The importance of this classification system can be seen in its evolution into a myriad of roles such as a stratification tool of perioperative risk, a method to determine resource allocation, a component of the billing process and reimbursement, institutional and federal policy making, a clinical criterion for student registered nurse anesthetists (SRNAs) and resident anesthesiologists for graduation, and a performance evaluation metric as part of audits performed by various institutions [4,9,10]. However, it has become apparent through multiple studies that the definition-only format of the classification system provides inadequate information for anesthesia providers to issue accurate, consistent, and reliable PS classifications for the patients they encounter. The fact that more non-anesthesia providers are using this

classification system for the several purposes mentioned above is one of the stated reasons why the ASA published case examples for each ASA-PS category in 2014 [11]. The ASA stated the need to aid clinicians and others in the determination of the ASA PS" and to provide "some context for the classic ASA PS definitions [3]. These case examples provided the user with objective criteria in assigning PS scores to patients and can improve interrater reliability and consistency in PS classification.

There are several situations in which non-anesthesia trained personal are using the PS classification system. Certain institutional policies allow emergency department physicians to provide sedation for PS I and II patients but require an anesthesia provider present for PS III and IV patients. In United States military and federal hospitals, nurse anesthetists can independently provide anesthesia for PS I and II patients but need 'supervision' or 'collaboration' with a physician prior to providing an anesthetic for PS III, IV, and V patients.⁴ The PS classification system is routinely used to correlate anesthesia and patient outcomes such as morbidity and mortality. Nationally, the PS score is used in risk adjustment models such as the National Surgical Quality Improvement Program in comparing surgical outcomes amongst different hospitals [12,13]. Patients that are judged to be of a higher PS class are perceived to require more resources and vigilance to ensure a positive medical outcome [10]. The PS classification system is also a key component of billing practices for anesthesia groups such that providers can incur addition base units for sicker patients. Correctly assigning PS scores to patients has the potential to cut costs for patients and improve resource allocation for anesthesia groups given that the literature has consistently shown that the greatest inaccuracy occurs in differentiating PS II and PS III patients [14].

Since the 2014 update to the ASA-PS by the ASA, there has only been one study published demonstrating its potential impact on accuracy and reliability [15]. This study found that less than 50% of participants were even aware that the ASA published specific examples of each physical status category. The authors concluded that the addition of specific examples to each PS classification improved the accuracy and interrater reliability among anesthesia providers in appropriately assigning patients into specific PS groups [15]. There appears to be a general lack of innovation and evolution of the PS classification scale even though there have been numerous studies that demonstrate the suboptimal accuracy and interrater reliability among providers. Not only is this practice change feasible, it is eloquent in its simplicity and potential impact.

Materials and Methods

This study took place after approval from the Rutgers University Institutional Review Board (IRB) and letters of cooperation from the involved anesthesia groups. This pretest-posttest questionnaire-based study poses the clinical question: Does the addition of specific case examples in the form of a cognitive aid improve the accuracy and interrater reliability of the PS Classification System for SRNAs, CRNAs, anesthesia residents, and anesthesiologists?

This pretest-posttest questionnaire-based study poses the clinical question: Does the addition of specific case examples in the form of a cognitive aid improve the accuracy and interrater reliability of the PS Classification System for SRNAs, CRNAs, anesthesia residents, and anesthesiologists? This study sought not only to further demonstrate the lack of interrater reliability of the PS classification system, but to implement the conclusions of the literature base surrounding this issue by creating and

disseminating a cognitive aid that aimed to improve the objectivity, accuracy, and interrater reliability between four groups of anesthesia providers mentioned above. The authors proposed that the issue lies not in the lack of research demonstrating the poor reliability of this assessment tool, but in the lack of research utilization in that the research was not being implemented into practice. Thus, the authors created a physical, readily available resource, the "badge buddy", which contains specific every day examples of the several PS classifications (see Figure 1). The authors "badge buddy" further revises and updates the ASA's 2014 PS classification system. Specifically, the authors extensively researched the literature surrounding PS classification and included more examples that were not included in the ASA's 2014 version. Only examples that were evidenced-based such as sleep apnea were included in the authors revised classification system. Anesthesia providers could utilize this updated "badge buddy" as a cognitive aid in order to improve accuracy and interrater reliability.


		
ASA-PS	Definitions	Examples/Comments
I	Normal Healthy patient	No physiologic disturbances, non-smoker, good exercise tolerance.
II	Patient with mild systemic disease	No functional limitations. + smoker, mild OSA, social drinker, pregnancy, obesity (BMI 30-40), well controlled DM/HTN, mild lung disease.
III	Patient with severe systemic disease	Functional limitations, 1 or more moderate to severe diseases. Poorly controlled DM/HTN, severe OSA, COPD, > 40 BMI, alcohol abuse, hepatitis, pacemaker, ejection fraction < 40%, ESRD with regular dialysis, history (> 3 month) of MI, CVA, TIA, or CAD/stents.
IV	Patient with severe systemic disease that is a constant threat to life	Recent (< 3 months) history of MI, CVA, TIA, or CAD/Stents, ongoing cardiac ischemia, severe valve dysfunction, ejection fraction < 25%, sepsis, DIC, ARDS, ESRD not undergoing regular dialysis
V	Moribund patient not expected to survive without operation	Ruptured thoracic/abdominal aortic aneurysm, multi-system trauma, intracranial bleed with mass effect, ischemic bowel, multiple organ/system dysfunction.
VI	Patient declared brain-dead who presents for organ donation	

Figure 1: The ASA PS Classification System was revised and updated in order to increase its objectivity and interrater reliability.

In this questionnaire-based pretest-posttest study, participants were asked to complete two consecutive surveys consisting of five hypothetical case scenarios created by the researchers. The interval between the pretest and the posttest was about two hours. The authors created five realistic case scenarios using an

extensive review of the literature of similar study methodologies, in addition to detailed conversations and interactions with several experienced anesthesia providers that can be regarded as experts in the discipline of anesthesiology. These scenarios were designed to highlight the most common deficiencies of the current PS

classification system that contribute to its subjectivity and lack of interrater reliability. The surveys were then administered to the experts in order to satisfy a minimum degree of validity. The post-test survey used the same case scenarios as the pre-test survey which was intended to allow each study participant to serve as his or her own control.

After completion of the pretest, the anesthesia providers participated in an interactive presentation that was designed to increase their knowledge of the PS classification system. They were then given a cognitive aid that contained the author's revised PS classification system in the form of a "badge buddy" that was used as an objective reference to improve accuracy and interrater reliability. Participants were then asked to complete the posttest with this new information. The researchers predetermined the appropriate ASA classification for each of the five case scenarios and performed statistical analysis regarding the accuracy and interrater reliability of participant responses after the completion of both surveys. Participant responses were analyzed to determine if the administration of the cognitive aid improved accuracy in assignment of ASA class. The researchers determined if there is any significant difference with regards to accuracy and interrater reliability among the four groups of anesthesia providers: SRNAs, CRNAs, anesthesia residents, and anesthesia attendings. In order to obtain a more diverse sample population, the researchers recruited study participants at three different locations: Rutgers University and two major teaching hospitals in the New Jersey area.

The authors of this study used convenience sampling to recruit participants with an educational background in anesthesia; specifically, attending anesthesiologists, CRNAs, SRNAs, and anesthesia residents. All persons that directly deliver anesthesia care to patients were eligible regardless of years of training or experience. The inclusion criteria consisted of providers that have an anesthesia background and deliver anesthesia on a regular basis. The exclusion criteria consisted of individuals who did not have an anesthesia background and did not perform anesthesia.

All statistical analysis was performed using SPSS version 25 software. The mean number of correct answers for the pre-test was compared to the mean number of correct answers for the post-test for each group of anesthesia providers using a repeated measures ANOVA using both within-subject's factors (difference of pre-test and post-test scores of all anesthesia providers across time) and between-subject's factors (differences across provider type and experience levels). Anesthesia provider type and

experience levels were the two main variables that were analyzed to determine if these factors had any impact on test scores before and after the study intervention.

The interrater reliability assessed the degree of agreement between two or more providers [16]. The percentage of agreement between providers was assessed by calculating the number of agreements in observations divided by the total number of observations. The second measure of interrater reliability used was Fleiss' Kappa which ranges from 0 to 1.0, where 1.0 represents the strongest interrater reliability. In general, kappa values from 0.21-0.40 represent fair interrater reliability, whereas 0.4 to 0.59 represents a moderate interrater reliability, and values from 0.6 to 0.79 represent good interrater reliability. Kappa values above 0.80 represent excellent interrater reliability [16].

Results

In total, there were 80 anesthesia providers that participated in this study. Thirty-five SRNAs were recruited from Rutgers University, 25 anesthesia providers were recruited from one of the major teaching hospitals in New Jersey, and 20 providers from the other teaching hospital. SRNAs and CRNAs made up about 75% of the sample size. The rest of the sample size consisted of 13 anesthesiologists and 6 anesthesia residents (see Table 1 for the samples frequency distributions by variable type).

Provider Type	1	SRNA	35
	2	CRNA	26
	3	MD	13
	4	Resident	6
Site	1	Barnabas	20
	2	NBI	25
	3	Rutgers	35
Years of Experience	1	0-2 Years	51
	2	3-5 Years	10
	3	6-10 Years	13
	4	> 10 years	6

Table 1: Frequency by Variable Type.

The average pretest and posttest scores by provider type can be seen in Table 2 and Figure 2. The average pretest score for all anesthesia providers was 72.50 (SD, 16.650) with an average posttest score of 95.25% (SD, 9.137). The difference between pretest scores and posttest scores as a whole was statistically significant ($t = 14.77$; $df = 79$; $p = .000$). As seen in Table 2, all anesthesia provider groups statistically significantly improved after the study intervention. Even though the SRNA group had the lowest

average pretest scores, this group made greatest improvement in accuracy, demonstrating an increase of 27.43 points (SD, 13.793). By provider type, the anesthesiologist (MD) group had the highest pretest and posttest scores with average pretest score of 86.15% (SD, 17.097), and a posttest score of 98.46% (SD, 5.547). In comparing test scores between provider group, there was

a statistically significant difference in pretest scores, with the anesthesiologists (MD) scoring the highest, but no significant difference between provider group after the study intervention in posttest scores. After adjusting for experience levels of the provider groups, there was no significant difference.

Provider Type		Pretest Score	Posttest Score	P Values
SRNA	Mean	67.43	94.86	p = < .05
	N	35	35	
	Std. Deviation	15.405	10.109	
CRNA	Mean	72.31	93.85	p = < .05
	N	26	26	
	Std. Deviation	16.077	9.414	
MD	Mean	86.15	98.46	p = < .05
	N	13	13	
	Std. Deviation	17.097	5.547	
RESIDENT	Mean	73.33	96.67	p = < .05
	N	6	6	
	Std. Deviation	10.328	8.165	
TOTAL	Mean	72.50	95.25	p = < .05
	N	80	80	
	Std. Deviation	16.650	9.137	

Table 2: Pretest and Posttest Scores by Provider Type.

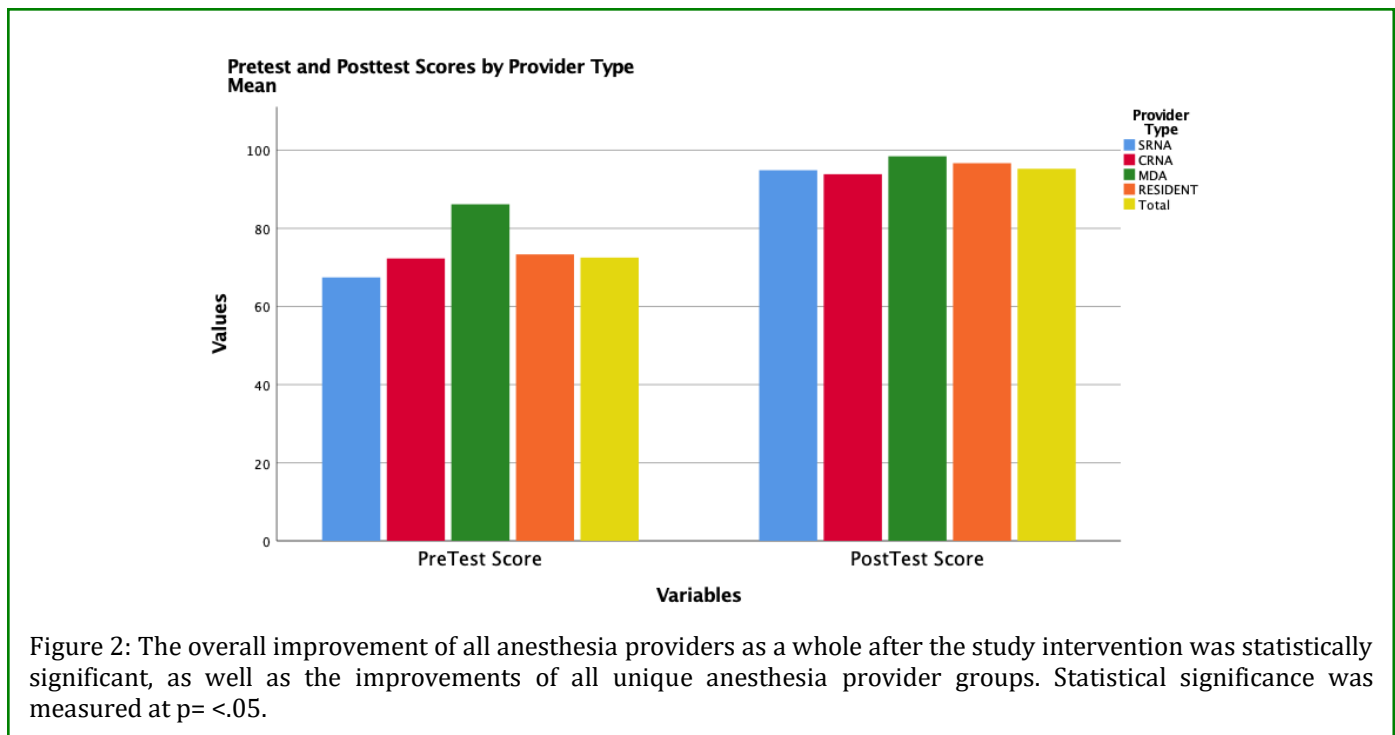


Figure 2: The overall improvement of all anesthesia providers as a whole after the study intervention was statistically significant, as well as the improvements of all unique anesthesia provider groups. Statistical significance was measured at $p < .05$.

In addition to the analysis of test scores by provider type, the researchers also analyzed scores by experience levels. As was expected, the providers with the least experience, the SRNAs, scored the lowest pretest scores with an average score of 66.67% (SD, 15.405). The providers with greater than 10 years' experience had the highest test

scores, with an average pretest score of 96.67% (SD, 8.165) and an average posttest score of 100% (SD, .000). In fact, the most experienced group was the only group to not reach statistical significance with comparison to pretest scores due to the fact that both scores were near perfect (see Table 3 and Figure 3).

Years of Experience		Pretest Score	Posttest Score	P Values
0-2 years	Mean	66.67	93.73	p = < .05
	N	51	51	
	Std. Deviation	14.787	10.190	
3-5 years	Mean	78.00	96.00	p = < .05
	N	10	10	
	Std. Deviation	14.757	8.433	
6-10 years	Mean	80.00	98.46	p = < .05
	N	13	13	
	Std. Deviation	14.142	5.547	
> 10 years	Mean	96.67	100.00	p = > .05
	N	6	6	
	Std. Deviation	8.165	.000	
Total	Mean	72.50	95.25	p = < .05
	N	80	80	

Table 3: Pretest and Posttest Scores by Experience Levels.

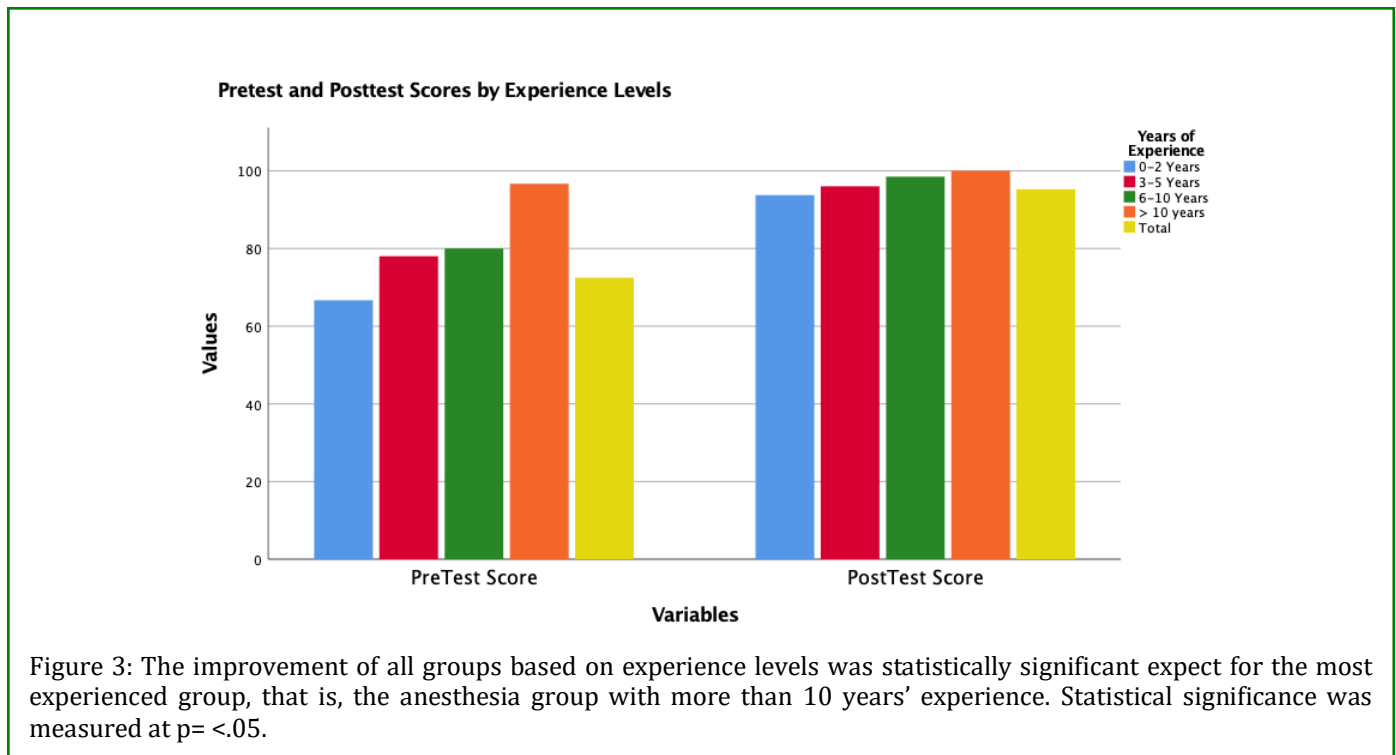


Figure 3: The improvement of all groups based on experience levels was statistically significant expect for the most experienced group, that is, the anesthesia group with more than 10 years' experience. Statistical significance was measured at $p < .05$.

The improvement in accuracy can be seen by the drastic increase in posttest scores. In addition to measuring accuracy, the authors measured interrater reliability. As

seen in Tables 6 and 7, only 15% correctly assigned all five case scenarios with a score of 100%. This is in stark contrast to the 77% of participants that scored 100% on

the posttest. Therefore, we can conclude that there was a high degree of interrater reliability after the study intervention. The authors measured Kappa values to determine the exact degree of interrater reliability. Fleiss' Kappa was run before and after the study intervention to determine the effectiveness of the intervention and the

improvement in interrater reliability. As seen in Table 4 and 5, there was poor to fair interrater reliability between anesthesia providers before the study intervention ($k=.301$, $p < 0.001$) and excellent interrater reliability after the study intervention ($k=.911$, $p < 0.001$).

Kappa		Asymptotic Standard Error	Z	P Value
Overall	.301	.073	7.981	.000

Table 4: Interrater Reliability Before Study Intervention: Fleiss Kappa.

Kappa		Asymptotic Standard Error	Z	P Value
Overall	.911	.053	6.463	.000

Table 5: Interrater Reliability After Study Intervention: Fleiss' Kappa.

		Frequency	Percent	Valid Percent
Valid		40	6	7.5
		60	30	37.5
		80	32	40.0
		100	12	15.0
		Total	80	100.0

Table 6: Pretest Score Distributions.

		Frequency	Percent	Valid Percent
Valid		60	1	1.3
		80	17	21.3
		100	62	77.5
		Total	80	100.0

Table 7: Posttest Score Distributions

Providers tended to inaccurately upgrade the patients PS score on the pre-test and then appropriately downgrade their assessment on the post-test after the educational session and the use of the cognitive aid. This observation is consistent with the literature that states providers have difficulty in providing accurate PS scores to patients with multiple 'controlled' systemic diseases. There were no significant differences between hospital sites in terms of accuracy and interrater reliability. The researchers chose to conduct this study at two different clinical sites in order to assess if there were any cultural differences in assigning ASA scores. These results probably stem from the fact that providers belong to the same anesthesia group and often rotate between hospitals, establishing similar cultures in both hospitals.

Providers with more clinical experience had higher scores on the pre-test than providers with less experience. The anesthesiologists in the group all tended to have greater experience levels than the other providers. After adjusting for years of experience, there was no significant

difference between anesthesia provider groups. The variable of experience level was statistically significant in that it predicted higher accuracy in using the PS classification system. SRNAs demonstrated the greatest improvement in accuracy after the study intervention. This can be explained by the increased receptivity of young anesthesia providers, having less innate experience to draw knowledge from. The anesthesiologists tended to score higher on the pre-test and posttest therefore demonstrating the least improvement after the study intervention (improvement in average posttest score of 12.31 points (SD, 15.359)). Interestingly, a majority of anesthesiologists and anesthesia residents stated they do in fact use the PS as a risk prognostication tool. Additionally, almost 100% of providers stated that the PS classification system needs to be updated.

Discussion of Findings

The authors' hypothesis was confirmed: When provided with objective examples for each ASA-PS classification,

accuracy and interrater reliability improved in each group of anesthesia providers. This is only the second study of its kind to be done that specifically evaluates whether a more objective ASA-PS classification system can increase accuracy and reliability among different anesthesia providers. The results of this study are in alignment with previous studies indicating the need for an updated classification system. [4, 6-10,15].

Results of this study indicated that with the ability to reference objective examples on a cognitive aid, provider accuracy and interrater reliability significantly improved when assigning PS scores to hypothetical case scenarios. These results should translate into real world PS scoring as care was taken when creating these examples to maintain the fidelity of what a provider would encounter in clinical practice. These findings support evidence from previous literature which suggested that when providers are made aware of evidence-based examples of the ASA-PS scale, their accuracy and consistency amongst their colleagues improved [9,15]. Perhaps a future study can investigate whether use of a physical cognitive aid such as the "badge buddy" in this investigation yields more significant results in provider accuracy and interrater reliability than with an educational intervention alone.

A limitation in the methodology of this study is noted in the convenience sampling used to recruit study participants. This type of sampling may not depict an accurate representation of the intended population of the institution or clinical site. For example, 63% of the sample had less than 2 years of experience, with only 25% having greater than 5 years of experience. Providers with less clinical experience using the PS scoring system (i.e. SRNAs, Residents) performed less effectively on the survey than providers with more experience. This may have led to a more dramatic improvement between pre and post-tests amongst the entire sample as a whole. Since the anesthesiologist group tended to have more experience levels than the other providers, this may have influenced the pre-test results in favor of this provider type. Also, more years of experience, or medical education, may not be the only possible factors in the high pre-test scores of the anesthesiologists. MD's in team settings do most or all pre-op documentation and assign PS scores to each patient in writing. Thus they have more practice in assignment.

Conclusion

Recent literature has shown that the ASA classification system has poor interrater reliability between anesthesia providers resulting in inconsistent application and variances in classification. Since the subjectivity and poor

interrater reliability could have several implications, this pretest-posttest study sought to measure the effect of introducing a more objective PS classification system with the benefit of a cognitive aid. The average pretest score for all anesthesia providers was 72.50% (SD, 16.650) with an average posttest score of 95.25% (SD, 9.137), demonstrating a statistically significant increase in the accuracy of anesthesia providers using the PS classification system with the authors cognitive aid ($t = 14.77$; $df = 79$; $p = .000$). There was poor to fair interrater reliability between anesthesia providers before the study intervention ($k=.301$, $p < 0.001$) and excellent interrater reliability after the study intervention ($k=.911$, $p < 0.001$). The results of this study again confirm the subjective nature of the PS classification system and demonstrate that a revised PS classification system with more objectivity would improve both the accuracy and interrater reliability among anesthesia providers in using this tool. The more objective scoring system is more likely to benefit anesthesia providers with less experience. Due to the many of uses of this classification system, it is imperative that a more object tool be created.

References

1. Saklad M (1941) Grading patients for surgical procedures. *Anesthesiology* 2(3): 281-284.
2. Dripps RD (1963) New classification of physical status. *Anesthesiology* 24(3): 111.
3. Abouleish AE, Leib ML, Cohen NH (2015) ASA Provides Examples to Each ASA Physical Status Class. *ASA Newsletter* 79(6): 38-49.
4. Aronson WL, McAuliffe MS, Miller K (2003) Variability in the American Society of Anesthesiologists Physical Status Classification Scale. *AANA J* 71(4): 265-274.
5. Fitz-Henry J (2011) The ASA classification and peri-operative risk. *Ann R Coll Surg Engl* 93(3): 185-187.
6. Haynes SR, Lawler PG (1995) An assessment of the consistency of ASA physical status classification allocation. *Anaesthesia* 50(3): 195-199.
7. Riley R, Holman C, Fletcher D (2014) Inter-Rater Reliability of the ASA Physical Status Classification in a Sample of Anesthetists in Western Australia. *Anaesth Intensive Care* 42(5): 614-618.
8. Cuvillon P, Nouvellon E, Marret E, Albaladejo P, Fortier LP, et al. (2011) American Society of Anesthesiologists' Physical Status system: a

- multicentre Francophone study to analyse reasons for classification disagreement. *Eur J Anaesthesiol* 28(10): 742-747.
9. Sankar A, Johnson SR, Beattie WS, Tait G, Wijesundera DN (2014) Reliability of the American Society of Anesthesiologists physical status scale in clinical practice. *Br J Anaesth* 113(3): 424-432.
 10. Daabiss M (2011) American Society of Anaesthesiologists physical status classification. *Indian J Anaesth* 55(2): 111-115.
 11. (2014) ASA Physical Status Classification System. American Society of Anesthesiologists.
 12. Helkin A, Jain SV, Gruessner A, Fleming M, Kohman L, et al. (2017) Impact of ASA score misclassification on NSQIP predicted mortality: a retrospective analysis. *Perioper Med* 6: 23.
 13. Kuza CM, Hatzakis G, Nahmias JT (2017) The Assignment of American Society of Anesthesiologists Physical Status Classification for Adult Polytrauma Patients: Results From a Survey and Future Considerations. *Anesth Analg* 125(6): 1960-1966.
 14. Vogt AW, Henson LC (1997) Unindicated preoperative testing: ASA physical status and financial implications. *J Clin Anesth* 9(6): 437-441.
 15. Hurwitz EE, Simon M, Vinta SR, Zehm CF, Shabot SM, et al. (2017) Adding Examples to the ASA-Physical Status Classification Improves Correct Assignment to Patients. *Anesthesiology* 126(4): 614-622.
 16. Hallgren KA (2012) Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutor Quant Methods Psychol* 8(1): 23-34.