Austin Journal of Anesthesia and Analgesia

Open Access

Austin Publishing Group

Review Article

Cognitive Performance Assessment Using Simulation, Among Anesthesiology Residents – A Review

Sidi A*

Department of Anesthesiology, University of Florida College of Medicine, USA

***Corresponding author:** Avner Sidi, Department of Anesthesiology, University of Florida College of Medicine, USA

Received: January 14, 2020; **Accepted:** February 18, 2020; **Published:** February 25, 2020

Abstract

Our goal was to develop and validate a realistic assessment technique (using simulation), that will allow to assess Cognitive and Technical skills and performance of Anesthesiology residents. To achieve all these integrated goals we built a 3 phase plan.

The first phase in our investigation was to assess the construct validity (= progression of scores within progressing levels of training) of the simulation-based OSCE (=Objective Structured Clinical Examination) summative assessment tools developed and established by the Israeli Board Examination, and its potential generalizability to an American training program for formative or summative assessment. The exam related to 1st phase of our investigation was administered 66 times to 50 different residents. In the second phase we evaluated the deficiencies in cognitive performance according to error rates and performance grades within and across different clinical domains, and between PGY (= post graduate year of training) levels – in 47 residents tested 80 times. In the last phase, where our primary aim was to detect changes in "higher-order" deficiencies, comparing 2 successive academic years -- 35 PGY-3 & -4 residents were tested 50 times.

The pass rate in the 1st phase was significantly higher for PGY3 and PGY4 residents compared to PGY2 residents in the OR; this rate was also significantly higher for PGY4 residents compared to PGY2 residents when all three clinical domains were combined (11/22=0.50 *vs*. 2/23=0.09). The cognitive success rate by PGY4 residents in the 2nd phase was 0.5 - 0.68, and significantly lower than the non-cognitive success rate for Resuscitation and Trauma. In the 3rd phase we found a change in mean error rates across years. In all 3 clinical domains, the cognitive success rate was higher (range, 0.74–1.00) than the previous year's value (range, 0.39–0.87). The reduction in error rates is primarily due to decreases in non-technical errors, predominantly in resuscitation & trauma.

Conclusions: In its 1st phase, our study demonstrated the "generalizability", sharing of scenarios. In the 2nd phase our main findings revealed that PGY-3 & -4 residents' error rates were higher for the cognitive items as compared to the non-cognitive ones in each domain tested. In the final phase we demonstrated that not only simulation is effective at identifying these errors, but also that simulation may be a valuable way to teach and combat these errors.

Keywords: Assessment; Cognitive; Skills; Learning; Simulation; Anesthesiology; Residency

Introduction

The definition of performance in anesthesia varies dramatically – from vague (vigilance, data interpretation, plan formulation, and implementation) [1] to very technical, organized, and detailed (gathering information for preoperative evaluation, equipment pre-use preparation, intra-operative checks, postoperative management, airway assessment) [2,3]. Some investigators evaluate performance in anesthesia by separating basic knowledge (gathering information) or the technical (initiating and working with protocols, reviewing checklists) from the cognitive and behavioral or affective (decision-making and team interaction) aspects [4,5]. This separation is based on strong analogies to performance during management of critical events in aviation, another complex and dynamic domain [5]. Most educators in anesthesia today believe it is important to measure two separate aspects of skilled performance in managing crisis situations: implementing appropriate technical actions (technical performance) and manifesting appropriate crisis solving and management of anesthesia non-technical behaviors.

The definition of Anesthesia Non-Technical Skills (ANTS), [6-10] includes: (a) task management (planning, prioritizing, keeping standards, using resources); (b) team work (coordinating, exchanging information, using authority, assessing capabilities, supporting); (c) situation awareness (interpreting information, recognizing, anticipating; (d) decision making (identifying & selecting options, re-evaluating). Conversely, technical skills refers to everything that is not ANTS: basic & technical knowledge (gathering information,

Citation: Sidi A. Cognitive Performance Assessment Using Simulation, Among Anesthesiology Residents – A Review. Austin J Anesthesia and Analgesia. 2020; 8(1): 1085.

preparation of drugs and equipment, initiating and working with protocols and checklists) [11,3], [12-14] and psychomotor skills (perception, guided response) [15]. The ANTS concept was developed and evaluated in a project between the University of Aberdeen Industrial Psychology Research Center and the Scottish Clinical Simulation Center. A team of anesthetists and psychologists was assembled and designed the anesthetists' non-technical skills system using methods of task analysis similar to the one used for pilots [7,16]. The ANTS include the main nontechnical skills (cognitive and affective) associated with good anesthetic practice [11,3,17].

Models that integrates lower-level knowledge and lower-level skills-based learning with a higher-level skills (of attitude, skills, behavior and culture of patient safety) – were developed for the simulated [13,18] and non-simulated[19 environment. One of the early models integrates four progressive capabilities: understanding (knows), application (knows how), integration (shows how) and practice (does) [19]. Knowledge is at the base of this framework and action/doing is at the top. Basic anesthesia knowledge is also a predictive academic variable for anesthesia resident clinical higher-level performance and is measured by using different tests during the first year of training [11].

The anatomical locations in the human brain for upper-level and lower level knowledge / learning are different, with the use of different neuro-transmitters: Cognitive learning and memory (motivation, decision-making) is based in the basal ganglia contrasting with the known role of the medial temporal lobe in declarative memory [20]. Nontechnical skills can be divided into two subgroups: (1) cognitive or mental skills (decision-making, planning, strategy, risk assessment, situation awareness); and (2) social or interpersonal affective skills (teamwork, communication, leadership). Both are necessary for safe and effective performance in the operating room, [21] and represent 2 of 3 legs in the skills triangle (with the psychomotor skills being the third leg), was already presented in previous publication in this journal and other publications [15,18,22].

Competency assessment of non-technical (= cognitive and affective) and technical (= psychomotor) skills [15,22], is extremely hard be accomplished using only traditional examinations [11,23-25]. Most clinical competence assessments use either performance-based methods (e.g., objective structured clinical examinations aka OSCEs) or tests that assess the "technical rationality" part of clinical reasoning (e.g., multiple-choice questions). These fail to capture the uncertainty of some clinical scenarios that will be encountered. Problem-solving in the operating room requires a mixture of knowledge and experience [24].

Current evaluation methods (including simulation-based) typically measure basic knowledge and performance, rather than competency, in the complex tasks of acute care [2]. This is why it is important to develop more efficacious methods to measure acute care clinical performance. Simulation could be used to measure advanced cognitive diagnostic and therapeutic management skills and the ability to integrate knowledge, clinical judgment, communication, and teamwork into the simulated practice setting.

Our goal was to develop and validate a realistic assessment technique (using simulation –environment & methodology), that will allow us to assess skills and performance of Anesthesiology residents, differentiate between Cognitive and Technical performance, enable us to detect deficiencies, and identify longitudinal changes in cognitive skills – meaning that cognitive performance deficiencies can be improved over time. In order to achieve all these integrated goals we built a 3 step / phase plan.

Our first phase was to assess the "construct validity" (= progression of scores within different progressing levels of training) of the simulation-based OSCE summative assessment tools developed originally for non-American Examination [26,27] and their potential generalizability to an American training program for formative or summative assessment. American Anesthesiology residents across all post-graduate years (PGY 2-4) in one institution (of 80 in the residency program) were examined. This validation could not be performed in the Israeli Board setup which tested only graduating residents equivalent to American PGY4 residents. The other aim was to demonstrate the "generalizability", sharing the scenarios developed for the non-American examination with an American academic environment for formative (teaching) and summative (testing) assessment [28].

The second phase of our investigation was to evaluate the deficiencies in cognitive performance according to error rates and performance grades within and across different clinical domains, and between PGY levels. Based on our previous preliminary work 47, we hypothesized that we would uncover some deficiencies in knowledge and skills, and that there would be fewer higher-order cognitive deficiencies in graduating compared to starting PGY residents [18].

In the last phase, our primary aim was to evaluate cognitive performance, and detect "higher-order" deficiencies according to error rates and performance grades within three different clinical domains (OR, trauma, and cardiac resuscitation) and between PGY levels, comparing 2 successive academic years. Our main objective was to demonstrate that simulation can effectively serve as assessment of cognitive skills and can help detect "higher-order" deficiencies, which are not as well identified through more traditional assessment tools. We hypothesized that simulation can identify longitudinal changes in cognitive skills – meaning that cognitive performance deficiencies should improve over time. We expected to see improvement in some deficiencies in knowledge and skills and hypothesized that there would also be fewer higher-order cognitive deficiencies for residents in the subsequent academic year from a learning effect. This learning effect is known as "construct validity" or progression of scores over time within progressing levels of training [28,1]. We expected that progression in scores will also be evident for the whole group of graduating residents evaluated in other fields and different scenarios.

Methods

In order to achieve the above mentioned 3 goals, we built a 3 phases plan (see also detailed description in the Introduction). The first phase of our investigation we used summative assessment tools developed by the Israeli Board Examination [26.27.29]. In the 2nd and 3rd phases of our investigation we evaluated the deficiencies in cognitive performance according to error rates and performance grades within and across different clinical domains, and between PGY levels Following Institutional Review Board (IRB) approval, all study phases were conducted at the University of Florida anesthesiology residency program.

Scenarios

In Phase 1: Two similar but not identical scenarios (to counter scenario content leakage and enhance content security) were used in each of three clinical domains: resuscitation, trauma, and operating room crisis management - in a simulated environment [26.27.29]. These scenarios were originally developed by the Israeli Board of Anesthesiology Examination Committee [26,27], [30,31]. Faculty members from the Department of Anaesthesiology in the University of Florida, assisted by educational and simulation experts, translated the scenarios with maximal adherence to the original script [26,27], scenario protocol, language, and assessment tools. No change was made in scoring, assessment, pass/fail determinations, orientation of residents, or the examination process itself.

In phase 2,3: We used a previously described scenario approach: (first stage: basic knowledge; second stage: exploring advanced cognition by discussion /debriefing)[18,26-29,32,33] [see Figure 67 and our previous publications. [28,33]].

Two similar but not identical scenarios were used in each of three clinical domains (cardiac resuscitation, trauma management, intra-operative crisis management), in a simulated environment. These six scenarios were originally developed and used by the Israeli Board of Anaesthesiology Examination Committee [26-29],[32,33]. Faculty members in the University of Florida, Department of Anesthesiology, assisted by educational experts, translated and adapted the material and methods.

Participants

In Phase 1: Fifty Anesthesiology residents in Post-Graduate Years (PGYs) 2-4 were evaluated. The examination was administered 66 times to 50 different residents. All residents were recruited by the chief residents, and had previously participated in an orientation and sessions with the Human Patient Simulator (CAE Healthcare, Sarasota, FL). We evaluated all PGY groups within a 3 months window, in each phase of the study. Each consented resident received oral instruction and printed materials explaining the study objectives (of evaluating teaching or learning errors), and assurance that results were confidential and had no impact on their residency program evaluations. All residents had prior orientation to the high-fidelity Human Patient Simulator as a part of their curriculum. Practice and assessment of clinical skills in a simulator environment was not novel to the participants.

In Phase 2: 47 PGY2-4 residents participated 80 times.

In Phase 3: 35 PGY-3 and -4 residents (of 50 in the residency program) were tested 50 times during two subsequent years, and 18 of those 35 residents were evaluated in both years (first as PGY3 then as PGY4). Thus, we studied 35 residents (18 of these 35 residents were the same = participated in both phase 2 and 3) across 2 successive years as they graduated to the next level 1 year later. Eighteen examinees were evaluated in the same domain and in identical or similar scenarios during 2 consecutive years (2011–2012 and 2012–2013).

Assessment tools

In all phases; Full description of the assessment model and Tools in all 3 phases of our investigation, our study protocol, assessors, and the scoring system -- appear in Appendix 1 [see also Figure [67], and our previous and other publications. [18,33],[26-29],[32,34,35-37]. This model integrates four progressive capabilities: understanding (knows), application (knows how), integration (shows how) and practice (does) [19]. This checklist scores performance using the item-based Angoff method [35,36] (see Appendix 1).

Feedback

In all phases, the residents completed questionnaires on realism of each scenario, including the perceived relevance of the scenario(s) and the residents' satisfaction from their performance in the simulation.

Calculations (Appendix 2)

For every item in each scenario, the following parameters were calculated as previously described [26-28], and compared between PGY groups, calculating Group (PGY) Error rate; Item performance grade; and Individual (Resident) Success Rate;

Statistics

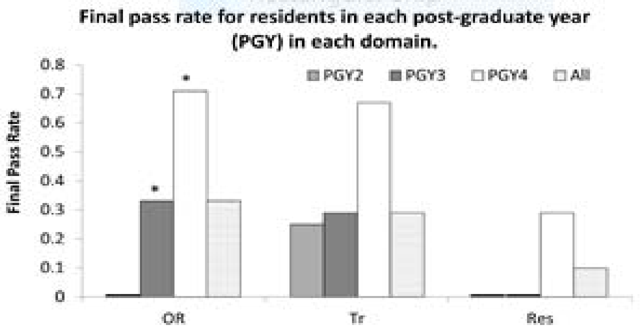
Each checklist and script included identification of the resident PGY level. Results were analysed using a SAS9.2 statistical software package. Checklist results were manually entered into an Excel (Microsoft, Redmond, WA) spreadsheet. A non-inferiority test (for proportion correct scores) was conducted between each pair of scenarios in each field to test equivalence between scenarios, assuming an allowable difference of ≤30% in performance or difficulty grades, while checking power for range of difference [38]. The non-inferiority test was performed in order to determine that the two scenarios in each of the clinical areas (within the same type or field) were not inferior to each other; A subsequent equivalence test (for proportion correct scores) was conducted between each pair of scenarios to evaluate similarity between them [31]. Equivalence was accepted with 80% certainty if the ratio (log difference) of the grades was within 20% for the pair. The log difference was used because the grades distribute log-normally. The 80% and 20% thresholds were used because these are accepted rates of equivalency tests [39]. Variables are presented as mean ± SD. Differences were considered significant when $p < 0.05$.

The feedback questionnaires that analyze realism of the scenario, perceived relevance, and satisfaction from own performance in the simulation scored on a scale from 1-5; 5 being the highest) – were scored by the examinees. Correlations between resident satisfaction with their own performance in the simulation and both the total proportion-correct scores and the general-scores, were calculated.

Variables were compared between groups by a random mixed-effect ANOVA model. We calculated means for PGY and field as random variables and the scenario as the fixed variable.

The error and pass rates for scenarios were compared using a 2-prop-z-test.

In phase 2,3: Individual success rates are presented as mean ± SD and grouped error rates are presented as ratios of errors for each scenario within a clinical domain for each PGY level. t- and Kruskal-Wallis tests were used to determine if individual success rates were significantly different between two scenarios within each field. An equivalence test was conducted between scenarios in



Domain		OR	Tr	Res
Final Pass Rate	PGY2	0	0.25	0
	PGY3	0.33*	0.29	0
	PGY4	0.71*	0.67	0.29
	All	0.33	0.29	0.10

Figure 1: Final pass rate residents in each post-graduate year (PGY) in each domain.

Abbreviations: OR: Operating Room; Tr: Trauma; Res: Resuscitation

*p < 0.05 compared to PGY2

†p < 0.05 compared to PGY3

each domain to test for equivalence 54 between the two scenarios in each domain. Group error rates for nontechnical and technical items were compared for each scenario within each PGY by using a 2-proportional z-test. Scenarios within each domain and PGY level were similarly compared for error rates.

Linear mixed models were used to compare individual success and error rates between PGY groups. PGY level, domain, and scenario were considered fixed effects and identification of the resident was considered a random effect in order to adjust for correlations among observations from the same subject. The Kenward-Roger method was used to calculate the denominator degrees of freedom due to the unbalanced study design. The Tukey-Kramer method was used to adjust for multiple comparisons. For all analyses, alpha was designated as 0.05. Data were analyzed using SAS 9.3 (SAS, Cary, NC).

Cognitive Errors Analysis – All items tested in each scenario script were evaluated, concentrating on the grouped error rates of >0.7 by the graduating PGY4 group during the first (non-cognitive) stage and the second (cognitive) stage. We then related the deficiencies we observed to a list from a recent publication that identified important cognitive errors in anesthesiology practice [37].

Results

Phase 1

The ANOVA analysis of the different PGY levels was significant, and analysis revealed what drove those differences:All scenarios were compared in the difficulty level (performance grade), and were not different amongst different PGY and clinical domains tested. A non-inferiority test [38] and a subsequent equivalence test [39] (for proportion correct scores) demonstrated the similarity between the 2 scenarios for the OR and Resuscitation. The corresponding P values to the equivalence tests are for Resuscitation = 0.0976 (equivalent at 10% level); Trauma = 0.2712 (not equivalent at 10% level); OR = 0.0619 (equivalent at 10% level); Overall = 0.005 (equivalent at 10% level). Thus, in the case of Trauma the equivalence cannot be said with the same 80% certainty, and scenario 1 has higher grades than scenario 2.

There are no significant differences in the performance grades (calculation of scenario difficulty) within any scenario pair in a domain. The error rate was lower for PGY4 residents compared to PGY2 residents in each domain, and scenario – except in scenario OR #1 and Trauma #2, where the error rate was relatively high for all PGYs. When scenario #1 and #2 in each clinical domain was considered as one unit, the error rate was significantly lower in each domain for PGY4 residents.

The critical items error rate was significantly lower for PGY4 residents compared to PGY3 residents in the OR domain; this rate was also significantly lower for PGY4 residents compared to PGY2 residents in the resuscitation domain.

The final pass rate was significantly higher for PGY3 and PGY4 residents compared to PGY2 residents in the OR (Figure 1) [28]; this rate was also significantly higher for PGY4 residents compared to PGY2 residents when all three clinical domains were combined (11/22=0.50 *vs*. 2/23=0.09).

Phase 2

There were no significant differences in error rates between the pair (scenario 1 *vs*. 2) in Resuscitation or OR domains for all items (p= 0.14 or p=0.44, respectively), within any of the PGY `levels, whereas the difference in error rates was significant for Trauma scenario 1 *vs*. 2 (p=0.001), for PGY2 *vs*. PGY4 residents.

When we investigated differences in error rates for non-cognitive or cognitive items, a significant difference within pair of scenarios (error rates in scenario 1 *vs*. 2) was found only in the OR amongst PGY3 residents tested for cognitive performance and in Trauma for PGY2 and PGY4 residents for non-cognitive performance.

The cognitive success rate by PGY4 residents was 0.5 - 0.68, and significantly lower than the non-cognitive success rate for Resuscitation and Trauma.

Difference in overall (all items) success rates between the two

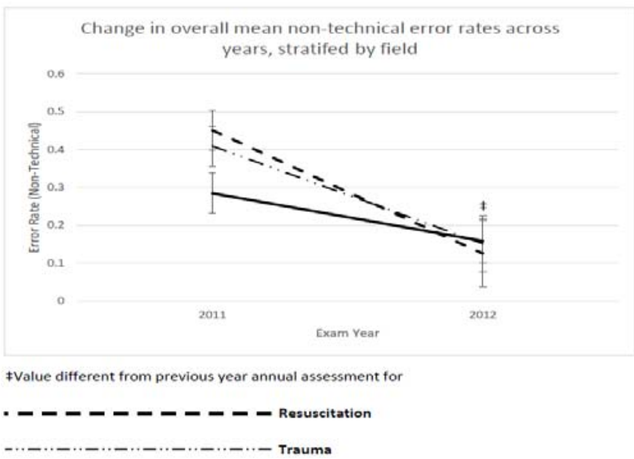


Figure 2: ‡Value significantly different (P < 0.05 and P < 0.003, respectively, for trauma and resuscitation) from the previous annual assessment.
Error rates in trauma and resuscitation were higher in 2011 compared to OR. Both of these error rates show greater decreases, so the error rates across all three fields are similar in 2012.

Table 1: Cognitive Errors found in the Operating-Room (O.R.), Trauma and Resuscitation Scenarios according to reference[18], compared with the errors catalogue ranking according to reference [37].

Catalogue Ranking Importance 37	Error Frequency Ranking in OR Trauma Resusitation18	Cognitive Error Type	Cognitive Error Definition
1	1	Anchoring	Focusing on one issue at the expense of understanding the whole situation
2	2	Availability bias	Choosing a diagnosis because it is in the forefront of your mind due to an emotionally charged memory of a bad experience
3	3	Premature closure	Accepting a diagnosis prematurely, failure to consider reasonable differential of possibilities
5	4	Confirmation bias	Seeking or acknowledging only information that confirms the desired or suspected diagnosis
7	5	Commission bias	Tendency toward action rather than inaction. Performing un-indicated maneuvers, deviating from protocol. May be due to overconfidence, desperation, or pressure from others
8	6	Overconfidence bias	Inappropriate boldness, not recognizing the need for help, tendency to believe we are infallible
10	7	Sunk costs	Unwillingness to let go of a failing diagnosis or decision, especially if much time/resources have already been allocated. Ego may play a role
12	8	Zebra retreat	Rare diagnosis figures prominently among possibilities, but physician is hesitant to pursue it

scenarios within a clinical domain was significant only for trauma. A difference between the two scenarios was significant for cognitive success rates only amongst PGY3 residents for the OR (p=0.02); and for non-cognitive performance amongst PGY4 residents for trauma (p=0.02).

When we evaluated the performance grades of <0.6 for each item amongst PGY4 group, and the error in each item related to known possible cognitive errors [35]. The most common cognitive errors observed in the Resuscitation scenarios were availability bias (choosing a diagnosis because it is in the forefront of your mind due to an emotionally charged memory of a bad experience) and premature closure (accepting a diagnosis prematurely, or failure to consider reasonable differential of possibilities); in the Trauma scenarios anchoring (focusing on one issue at the expense of understanding the whole situation) and premature closure and in the OR scenarios anchoring, availability bias, premature closure, confirmation bias (seeking or acknowledging only information that confirms the desired or suspected diagnosis).

Phase 3

PGY-4 residents error rates were lower for the cognitive items compared to basic and technical item performance in resuscitation and OR domains ($P < 0.05$), but not in each scenario and after controlling for PGY and field (using similar mixed models as described in a previous paper [18,33]). In all three clinical domains, the cognitive error rate by PGY-4 residents was fairly low (0.08-0.22) and the cognitive success rate by PGY-4 residents was high (0.83-1.00). These findings were significantly better ($P < 0.05$) compared to the previous annual assessment.

Figure 2 [33] presents the change in mean error rates across years. The reduction in error rates is primarily due to decreases in non-technical errors. The differences between cognitive and technical errors were not the same between years, with this effect more predominant in resuscitation and trauma. Error rates in trauma and resuscitation were higher in 2011–2012 compared to OR, but these error rates showed greater decreases in 2012–2013, thus the error rates within all three fields were similar in 2012–2013. The differences between cognitive and technical error rates across years in trauma and resuscitation are significantly higher ($P < 0.05$ and < 0.003, respectively) in 2011–2013 compared to OR, but they both show significant decreases, thus the differences between cognitive and technical error rates are similar in 2012–2013.

In all three clinical domains, the cognitive success rate by PGY-3 and PGY-4 residents in 2012–2013 was higher (range, 0.74–1.00) than the previous year's value (range, 0.39–0.87), and significantly better ($P < 0.05$) compared to the previous annual assessments. Cognitive error rate values within the same examinees in trauma and resuscitation were initially higher compared to OR, but they both showed greater decreases in 2012–3, thus the error rates within all three fields were similar in 2012-13.

The success rate, major (with the grouped success rates of <0.7 in the PGY-4 group) cognitive errors within list of items tested, during the second (cognitive) stage of each scenario, and the possible error type contributing to each deficient item [37], we isolated 16 major cognitive errors in 4 scenarios (out of 88 items = 17% error rate). Such major cognitive errors were decreased in the trauma scenarios during two subsequent years. The amount and quality of cognitive errors in the resuscitation and OR domains were minimally changed compared to the previous (2011–2012) annual evaluation. The cognitive errors ranked as 1, 3 “cognitive-errors” (anchoring, premature closure) in the trauma domain were decreased compared to the previous (2011–2012) annual evaluation. The most common cognitive errors in the resuscitation scenarios remained Similar between the 2 succssesive years. Thus, the most common cognitive errors observed remained anchoring, availability bias, premature closure, and confirmation bias.

Discussion

In the process of assessing a teaching/testing tool the validation should take into consideration the case selection, case-subject interaction, rater variability, “construct” validation, and the rating system. Our study was performed in an academic department, which allowed assessment of the construct-related validity of the scenarios - i.e. the ability of each scenario to differentiate between participants on the hypothesis that more senior residents will generally have better scores. The study demonstrated that PGY 4 residents had superior results compared to PGY2 residents in: error rate, total scenarios score, general evaluation score, critical items error rate, and final pass rate. Our results are very similar to the results of the original tests performed amongst graduating (PGY4) residents by the Israeli

Examination [26,27], These comparable results demonstrate the "generalizability" of the scenarios.

In the second phase our main findings revealed that PGY-3 and PGY-4 residents' error rates were higher and success rates lower for the cognitive items as compared to the non-cognitive ones in each domain tested. The most common cognitive errors in all three domains were ranked within top "cognitive-errors" (anchoring, availability bias, premature closure, confirmation bias) [37].

The most interesting finding is that our simulations' cognitive deficiencies mirrored the "top anesthesiology cognitive errors." Thus, similar cognitive errors were found also in other American program [37]. These findings are very similar to those found in a preliminary study that showed higher error rates in the evaluated clinical domains that had more cognitive or advanced knowledge items [18]. Also, comparing the results of OSCE scenarios previously used with non-American (Israeli) graduating – PGY-4 equivalent residents [28], to our program – we observed comparable error rates, performance grades and pass rates.

In the final phase we demonstrated that not only simulation is effective at identifying these errors, but also that simulation may be a valuable way to teach and combat these errors. We describe observed improvements in non-technical or "higher order" deficiencies, and cognitive performance skills as discerned from an item and scenario analysis within OR, trauma, and resuscitation domains in an anesthesia residency cohort over a 1-year time interval.

Learning theories in medical education offer insights into memory formation, motor skills acquisition, diagnostic decision-making, and instructional design [40]. In spite of a "non-consistent" approach to applying learning theories [41,42], the Accreditation Council for Graduate Medical Education (ACGME) has instituted an initiative that requires training programs to assess each resident's competence in several domains of medical practice (ACGME Outcomes Project, 2007) [43]. The ACGME toolbox for evaluation lists simulation training as the most effective evaluation strategy for medical procedures [44]. The up and coming ACGME "Developmental Milestones" for internal medicine residency is playing an even more prominent role in assessing clinical skills & reasoning and consultative care than in the Outcomes project [45]. As of 2015, all specialties began reporting milestones; however, simulation was still listed as an assessment method for both Patient Care (PC) and Interpersonal and Communication Skills (ICS), and is still very relevant today. In our 2018 publication, we bring as an example PGY-3 & PGY-4 anesthesiology resident's growth in both areas of PC and ICS [33].

Nontechnical skills should be specifically taught and evaluated in all anesthesia training programs [30,31,46]. Understanding and correcting cognitive errors cannot be overemphasized. Cognitive errors are thought-process errors which lead to incorrect diagnoses and/or treatments. The psychology of decision-making has received little formal attention in the anesthesiology literature. Only 7 years ago (2012), a cognitive error catalogue specific to anesthesiology practice was created [37]. This catalogue with the original ranking was matched with the cognitive errors found in the Operating-Room, Trauma and Resuscitation Scenarios – in a later (2014) anesthesia teaching program [18] (Table). The most common cognitive errors in all three tested domains were ranked within top "cognitive-errors" [18,37]. The most common higher-order errors in the OR scenarios as well as all 3 domains were anchoring, availability bias, premature closure, and confirmation bias [18]. Some items that were scored as critical by the authors when the cognitive- error anesthesiology 'top 10' was created but were observed relatively infrequently in these 2 comparative studies [18,37]. A goal for each anesthesiology training program should be to explore, define, and pinpoint its own cognitive learning errors and then plan an education strategy designed to decrease these errors.

If we view optimal performance as a combined ANTS integrated with technical skills, we should then expect anesthesiology residents to perform on the same high level for both technical and nontechnical skills. In order to achieve that level, learning objectives and curriculum/teaching should be adjusted to address the deficiencies identified in these learning skills. To reach this objective, educational training in cognitive errors, meta-cognition, and de-biasing strategies is needed [37]. However, there are still many questions regarding which errors are most important to address and which "adjustment" learning strategies are the most appropriate and effective in anesthesiology. Further research in this area is needed to reduce decision-making errors and improve patient safety [37]. Unfortunately, education research is not rocket science, which is built on a structured linear system with a straightforward set of factors which can be inserted into a well-articulated formula to predict a clearly defined outcome. Rather, if we must make analogies to the physical sciences, we might do better to look to quantum mechanics, or the "chaos" theory [47]. Such analogies might lead us away from the search for proof of simple generalized solutions to the observed problems/errors.

A typical process of building "adjustment" learning strategies might follow this strategy: identify a content area that needs to be taught; develop a teaching module to match the content and implement the module; test to see if it works try to figure out what went wrong; tweak the design and delivery; test to see if it works now (if it does not, go back…) [47]. There are few suggestions in the literature for "adjustment" learning strategies in order to improve cognitive/higher-order learning or performance:

1. **Problem-Based Learning (PBL)** is a well-known technique used in education for three decades [48]. This PBL approach can facilitate the students' processes of acquisition, organization, and retrieval of knowledge, and, to a certain degree, the transfer of knowledge and competencies across different problems [49].

2. **Focus groups** [50] involve physicians with a variety of clinical experience in conducting and analyzing broad clinical headings while focusing on certain themes, such as transferring knowledge into practice, and decision-making and uncertainty.

3. **Cognitive Task Analysis (CTA)** Simulation-based Modules [51] were developed using a framework of tasks, and the CTA theory as a guide [52]. The underpinnings of this theory are based on the assumption that every performable task consists of a series of basic and irreducible cognitive and perceptual operations that enable the human mind [53].

4. **Conceptual frameworks** [54] represent ways of thinking about a problem, or ways of representing how complex things work. Different frameworks will emphasize different variables and

outcomes, and their inter-relationship.

5. Cognitive simulators [55] use a generic framework for design, development, and evaluation of such simulators. This framework is generalizable, and can be applied to different task domains. It is independent of the types of sensors, simulation environment, and feedback mechanisms that simulators use.

6. Script Concordance Test (SCT) [23]**,** could be a new tool of clinical reasoning assessment, which may test the elaborated networks of knowledge that experienced surgeons / anesthesiologists acquire over the years. It allows for multiple different approaches to the same problem and could be developed.

Sharing scenarios can provide an objective comparative view of trainees in American and non-American residencies [28,32,34] and the potential for universal applicability of such scenarios, and learning from the mistakes detected [28]. When investigators used simulation-based assessment to highlight cognitive mistakes, these models also provided real-time feedback for the tested residents at the end of each scenario [18,28]. Exposing and revealing the mistakes found in the assessment during the debriefing stage can serve as an "adjustment" learning strategy. Defects or mistakes that are recurring themes should inform curriculum development [13].

When investigators based the assessment on testing for Minimal Requirement Task Performance (used in the OSCE [18,26-28]), it appears that even though a smaller number of the tasks/items were advanced/applied knowledge and skills, this type of task was more problematic for all residents [18,28], including the graduating residents [18,26-28]. These comparable results between studies demonstrate the "generalizability" or the feasibility of "sharing" formative or summative assessment scenarios. This feasibility of sharing scenarios between different residency programs has been previously demonstrated [26,34]. Although simulation in anesthesia has become part of the teaching curricula [4,56,57], only 14% of simulation centers used simulation for evaluation of competence in 2002 [58]. Reasons for this underutilization include lack of standardized, valid, and reliable tests.58 Communication and collaboration among centers involved in simulation programs (including sharing of validated scenarios) is important to the future of this technology and approach [59]. In the last 2 decades the use of simulation, communication and collaboration among centers increased dramatically: a 2011 national survey reported that 91% of pre-licensure nursing programs in the US are utilizing high- or medium-fidelity simulation [60,61].

In the last 2 decades, various subspecialties and fields associated with Anesthesia used simulation modalities (Cardiovascular, Vascular, Pain & Regional, Critical Care, Trauma, Transplants, Neuro and Preoperative Medicine) [62]. Also other medical professions beside Anesthesia – Nursing [60,61], Pediatric [62,63] Internal Medicine [45] – endorsed and adapted the use of simulation for testing and evaluation. Since 2004 the National Board of Medical Examiners (NBME) has required a simulation-based clinical skills examination for medical students [64]. The Accreditation Council for Graduate Medical Education (ACGME) Outcome and Milestones Projects [43-45], led the American Board of Anesthesiology (ABA) to give diplomates enrolled in Maintenance of Certification in Anaesthesiology (MOCA) from 2000 through 2007 the option to complete an endorsed simulation course in lieu of references to support their knowledge and skills [65]. More recently the ABA implemented a simulation-based OSCE as part of its certification exams [66,67].

Study Limitations

The **phase 1** study was conducted at a single institution; we did not perform a comprehensive calculation of sample size, and a limited sample of residents was used. This study is also limited in its ability to differentiate learning (by the residents) from teaching (according to a systematic curriculum). We did not consider the more advanced clinical year as the sole indicator of competence.

We attempted to minimize the variances in our study, but did not control or prove similarity or equivalence for PGY year- three levels and clinical domain -three levels. By the equivalence & non-inferiority tests the two scenarios in each domain (except for trauma) were not different in difficulty or performance level, thus the two scenarios in each field were treated as one unit, for the OR and resuscitation domains. The evaluator had only a single level of variability, because one experienced evaluator evaluated all residents; therefore PGY and domain between groups were compared using ANOVA.

Although data on mistakes performed during resident training were very close to the graduating residents (PGY4 equivalent) from the Israeli study [26], such a comparison between American and non-American PGY2-3 years is difficult and certainly not an aim of our study. Although the data represents a well-defined progression of knowledge and skill acquisition in the Americans this is not easily equated to Israeli residents. Most Israeli residents are older and more experienced than their American counterparts because many were previously trained and worked as medical doctors in the former Soviet Union, and some were in anesthesia practice for several years. but are required to repeat their training. A previously observed lower incidence of OSCE mistakes by Israeli junior residents [32] (as compared to the US junior residents) may be due to these differences in the medical background and experience. However, we did not find any differences at the graduating level between the 2 training programs in general scores, critical mistakes or incidence of mistakes.

The phases 2 & 3 studies had several limitations. The study was conducted at a single institution, we did not perform a comprehensive calculation of sample size, and a limited sample of residents was used, limiting generalizability. This study is also limited in its ability to differentiate learning (by the residents) from teaching (according to a Systematic curriculum). Despite a consensus that anesthesia acute care skills should be taught systematically and perhaps using simulation [46], these skills are taught sporadically rather than systematically. If they had been taught systematically, it would highlight a very different problem and suggest a problem with the teaching methods. We did not consider the more advanced clinical year as the sole indicator of competence. The simulation-based format of our practical examination tested the upper level of competence—the "does" stage of Miller's model of medical competence [19].

We do not have a very good understanding of how cues in the simulated environment affect decision making and problem solving. Thus, what we are witnessing may in part be due to the limitations of using simulation for summative assessment. For example, residents often perform relatively well in resuscitation scenarios because the cues received in the simulated environment are often clear-cut (e.g., arrhythmia on monitor) and the treatment follows well-known algorithmic approaches (e.g., Advanced Cardiovascular Life Support). Scenarios that are less clear-cut (e.g., evolving hypertension or hypotension) may depend on multiple cues from various sources with varying degrees of fidelity.

Using simulation for assessment may have its limitations. We do not have a very good understanding of how cues in the simulated environment affect decision-making and problem-solving. Thus, what we are witnessing may in part be due to the limitations of using simulation for summative assessment. For example, residents often perform relatively well in resuscitation scenarios because the cues received in the simulated environment are often clear-cut (e.g., arrhythmia on monitor), and the treatment follows well known algorithmic approaches (e.g., ACLS). Scenarios that are less clear-cut e.g. evolving hypertension or hypotension may depend on multiple cues from various sources with varying degrees of fidelity.

In summary, cognitive and non-cognitive simulation based skills assessment that included the so called Anaesthesia Nontechnical Skills (ANTS) can help to identify areas of strength and weakness that can be used guide the residency curriculum, especially with regard to deficiencies in tasks requiring higher-order processing. Any such deficiencies need to be addressed in any training program.

References

1. Devitt JH, Kurrek MM, Cohen MM, Fish K, Fish P, Noel AG, et al. Testing internal consistency and construct validity during evaluation of performance in a patient simulator. Anesth Analg. 1998; 86: 1160-1164.
2. Murray DJ, Boulet JR, Kras JF, Woodhouse JA, Cox TC, McAllister JD. Acute care skills in anesthesia practice. a simulation-based resident performance assessment. Anesthesiology. 2004; 101: 1084-1095.
3. Forrest FC, Taylor MA, Postlethwaite K, Aspinall R. Use of a high-fidelity simulator to develop testing of the technical performance of novice anaesthetists. Br J Anaesth. 2002; 88: 338-344.
4. Weller JM, Bloch M, Young S, Maze M, Oyesola S, Wyner J, et al. Evaluation of high fidelity patient simulator in assessment of performance of anaesthetists. Br J Anaesth. 2003; 90: 43-47.
5. Gaba DM, Howard SK, Flanagan B, Smith BE, Fish KJ, Botney R. Assessment of clinical performance during simulated crises using both technical and behavioral ratings. Anesthesiology. 1998; 89: 8-18.
6. Eisenach JC. Editorial: Excellence in Anesthesiology. The Role of Nontechnical Skills. Anesthesiology 2009; 110: 201-203.
7. Flin R, Patey R, Glavin R, Maran N. Anaesthetists' non-technical skills, British Journal of Anaesthesia. 2010; 105: 38-44.
8. Fletcher G, Flin R, McGeorge P, Glavin R, Maran N and Patey R. Anaesthetists' Non-Technical Skills (ANTS): evaluation of a behavioural marker system. Br J Anaesth. 2003; 90: 580-588.
9. Patey R, Flin R, Fletcher G, Maran N, Glavin R. Developing a Taxonomy of Anesthetists' Nontechnical Skills (ANTS). In: Henriksen K, Battles JB, Marks ES, Lewin DI, editors. Advances in Patient Safety: From Research to Implementation (Volume 4: Programs, Tools, and Products). Rockville (MD): Agency for Healthcare Research and Quality (US); 2005.
10. University of Aberdeen. ANTS website. 2019.
11. Reich DL, Uysal S, Bodian CA, Gabriele S, Hibbard M, Gordon W, et al. The relationship of cognitive, personality, and academic measures to anesthesiology resident clinical performance. Anesth Analg. 1999; 88: 1092-1100.
12. Castanelli DJ. The rise of simulation in technical skills teaching and the implications for training novices in anaesthesia.. Anaesth Intensive Care. 2009; 37: 903-910.
13. J M Rodriguez-Paz, M Kennedy, E Salas, A W Wu, J B Sexton, E A Hunt, P J Pronovost. Beyond ''see one, do one, teach one'': toward a different training paradigm. Postgrad Med J. 2009; 85: 244-249.
14. Objective Structured Assessment of Technical Skill (OSATS) in ACGME Outcomes Project Toolbox of Assessment Methods.
15. Rovai AP, Wighting MJ, Baker JD, Grooms LD. Development of an instrument to measure perceived cognitive, affective, and psychomotor learning in traditional and virtual classroom higher education settings. Internet and Higher Education. 2009: 7-13.
16. Flin R, Goeters K, Amalberti R et al. The development of the NOTECHS system for evaluating pilots' CRM skills. Human Factors Aerospace Saf. 2003; 3: 95-117.
17. Reader T, Flin R, Lauche K, Cuthbertson BH. Non-technical skills in the intensive care unit. Br J Anaesth. 2006; 96: 551-559.
18. Sidi A, Ozrazgat Baslanti T, Gravenstein N, Lampotang S: Simulation-based assessment to evaluate cognitive performance in an Anesthesiology residency training program. JGME. 2014; 6: 85-92.
19. Miller GE. The assessment of clinical skills/competence/performance. Acad Med. 1990; 65: S63-67.
20. Foerde K, Shohamy D. The role of the basal ganglia in learning and memory: insight from Parkinson's disease. Neurobiol Learn Mem. 2011; 96: 624-636.
21. Morgan PJ, Cleave-Hogg DM, Guest CB, Herold J. Validity and reliability of undergraduate performance assessments in an anesthesia simulator. Can J Anaesth. 2001; 48: 225-233.
22. Lampotang S, Lizdas D, Rajon D, Luria I, Gravenstein N, Bisht Y, et al. Mixed Simulators: Augmented Physical Simulators with Virtual Underlays. Proceedings of the IEEE Virtual Reality 2013 meeting. 2013; 13: 7-10.
23. Howard SK, Gaba DM, Fish KJ, Yang G, Sarnquist FH. Anesthesia crisis resource management training: teaching anesthesiologists to handle critical incidents. Aviat Space Environ Med. 1992; 63: 763-770.
24. Meterissian SH. A novel method of assessing clinical reasoning in surgical residents. Surg Innov. 2006; 13: 115-119.
25. Epstein RM, Hundert EM. Defining and assessing professional competence. JAMA. 2002; 287: 226-235.
26. Berkenstadt H, Ziv A, Gafni N, Sidi A. Incorporating simulation-based OSCE into the Israeli National Board Examination in Anesthesiology. Anesth Analg. 2006; 102: 853-858.
27. Berkenstadt H, Ziv A, Gafni N, Sidi A. The validation process of incorporating simulation based accreditation into the anesthesiology Israeli national board exams. Isr Med Assoc J. 2006; 8: 728-733.
28. Sidi A, Lampotang S, Gravenstein N. Construct Validity and Generalizability of Simulation-based Objective Structured Clinical Examination Scenarios for Trainee Assessment. J Graduate Med Education. 2014; 6: 489-494.
29. Ziv A, Rubin O, Sidi A, Berkenstadt H. Credentialing and certifying with simulation. Anesthesiol Clin 2007; 25: 261-269.
30. Schwid HA, Rooke GA, Carline J, Steadman RH, Murray WB, Olympio M, et al. Anesthesia Simulator Research Consortium. Evaluation of anesthesia residents using mannequin-based simulation: a multiinstitutional study. Anesthesiology. 2002; 97: 1434-1444.
31. Scavone BM, Sproviero MT, McCarthy RJ, Wong CA, Sullivan JT, Siddall VJ, Wade LD. Development of an objective scoring system for measurement of resident performance on the human patient simulator. Anesthesiology. 2006; 105: 260-266.
32. Berkenstadt H, Kantor GS, Yusim Y, Gafni N, Perel A, Ezri T, et al. The feasibility of sharing simulation-based evaluation scenarios in anesthesiology. Anesth Analg. 2005; 101: 1068-1074.

33. Sidi A, Vasilopoulos T, Gravenstein N, Lampotang S: Simulation-Based Assessment Identifies Longitudinal Changes in Cognitive skills and “Higher-Order” Deficiencies in an Anesthesiology Residency Training Program. J Patient Saf. In print. 2018.

34. Boulet JR, Murray DJ. Simulation-based assessment in anesthesiology: requirements for practical implementation. Anesthesiology. 2010; 112: 1041-1052.

35. Angoff WH. Scales, norms, and equivalent scores. In: Thorndike RL, ed. Educational Measurement. 2nd ed. Washington, DC: American Council on Education. 1971; 508-600.

36. Yudkowsky R, Downing SM, Wirth S. Simpler standards for local performance examinations: The yes/no Angoff and whole-test Ebel. Teach Learn Med. 2008; 20: 212-217.

37. Stiegler MP, Neelankavil JP, Canales C, Dhillon A. Cognitive errors detected in anaesthesiology. Br J Anaesth. 2012; 108: 229-235.

38. James Hung HM, Wang S-J, Tsong Y, Lawrence J, O’Neill RT. Some fundamental issues with non-inferiority testing in active controlled trials. In James Hung HM, ed. Proceedings of the Annual Meeting of the American Statistical Association, American Statistical Association. 2011.

39. Luˇzar-Stiffler V, Stiffler C. Equivalence Testing the Easy Way. Journal of Computing and Information Technology - CIT 10. 2002; 3: 233-239.

40. Hodges BD, Kuper AI. Theory and practice in the design and conduct of graduate medical education. Acad Med. 2012; 87: 25-33.

41. McGaghie WC, Issenberg SB, Petrusa ER. Simulation--Savior or Satan? A rebuttal. Adv Health Sci Educ Theory Pract. 2003; 8: 97-103.

42. Holzinger A, Kickmeier-Rust MD, Wassertheurer S, Hessinger M. Learning performance with interactive simulations in medical education: Lessons learned from results of learning complex physiological models with the HAEMOdynamics SIMulator. Comput Educ. 2009; 52: 292-301.

43. SR Swing. The ACGME outcome project: retrospective and prospective. Accreditation Council for Graduate Medical Education, Chicago, IL, USA. Medical Teacher. 2007; 29: 648-654.

44. Recommended Assessment Tools for the General Competencies. ACGME>Specialties>Recommended Assessment Tools for the General Competencies.

45. ACGME Milestones project: The Internal Medicine Milestone Project. A Joint Initiative of The Accreditation Council for Graduate Medical Education and The American Board of Internal Medicine.

46. Matveevskii AS, Gravenstein N. Role of simulators, educational programs, and nontechnical skills in anesthesia resident selection, education, and competency assessment. J Crit Care. 2008; 23: 167-172.

47. Regehr G. It’s NOT rocket science: rethinking our metaphors for research in health professions education. Med Educ. 2010; 44: 31-39.

48. Schmidt HG. Problem-based learning: rationale and description. Med Educ. 1983; 17: 11-16.

49. Van Gessel E, Nendaz MR, Vermeulen B, Junod A, Vu NV. Development of clinical reasoning from the basic sciences to the clerkships: a longitudinal assessment of medical students’ needs and self-perception after a transitional learning unit. Med Educ. 2003; 37: 966-974.

50. Tallentire VR, Smith SE, Skinner J, Cameron HS. Understanding the behaviour of newly qualified doctors in acute care contexts. Med Educ. 2011; 45: 995-1005.

51. Pugh CM, DaRosa S, Santacaterina S, Clark RE. Faculty evaluation of simulation-based modules for assessment of intraoperative decision making. Surgery. 2011; 149: 534-542.

52. Chipman SF, Schraagen JM, Shalin VL. Introduction to cognitive task analysis. In: Schraagen JM, Chipman SF, Shute VJ, eds. Cognitive Task Analysis. Mahwah, NJ: Lawrence Erlbaum Associates; 2000; 3-23.

53. Anderson JR. Cognitive Psychology and Its Implications. 4th ed. New York, NY: Freeman. 1995.

54. Bordage G. Conceptual frameworks to illuminate and magnify. Med Educ. 2009; 43: 312-319.

55. Kahol K, Vankipuram M, Smith ML. Cognitive simulators for medical education and training. J Biomed Inform. 2009; 42: 593-604.

56. Seropian MA. General concepts in full scale simulation: getting started. Anesth Analg. 2003; 97:1695-1705.

57. Blum RH, Raemer DB, Carroll JS, Sunder N, Felstein DM, Cooper JB. Crisis resource management training for an anaesthesia faculty: a new approach to continuing education. Med Educ. 2004; 38: 45-55.

58. Morgan PJ, Cleave-Hogg D. A worldwide survey of the use of simulation in anesthesia. Can J Anaesth. 2002; 49: 659-662.

59. Girard M, Drolet P. Anesthesiology simulators: networking is the key. Can J Anaesth 2002; 49: 647-649.

60. Kiernan LC. Evaluating competence and confidence using simulation technology. Nursing. 2018; 48: 45-52.

61. McCabe DE, Gilmartin MJ, Goldsamt LA. Student self-confidence with clinical nursing competencies in a high-dose simulation clinical teaching model. J Nurs Educ Pract. 2016; 6: 52-58.

62. Blasius KR, Streiff A, Singh D et al· Patel S. Chapter Simulation in Pediatrics. In book: Comprehensive Healthcare Simulation: Anesthesiology. Edited by B Mahoney, RD Minehart, MCM Pian-Smith. 2019.

63. Fehr JJ, Boulet JR, Waldrop WB, Snider R, Brockel M, Murray DJ. Simulation-based assessment of pediatric anesthesia skills. Anesthesiology. 2011; 115: 1308-1315.

64. Papadakis MA. The step 2 clinical skills examination. N Engl J Med. 2004; 350: 1703-1705.

65. Weinger MB, Burden AR, Steadman RH, Gaba DM. MOCA Part IV requirements update. American Board of Anesthesiology, Misconceptions Surrounding the Maintenance of Certification in Anesthesiology Simulation Course. Anesthesiology 09. 2014; 121: 655-659.

66. Applied Examination (new part 2 Oral Examination). American Board of Anesthesiology (ABA).

67. Sidi A. Editorial: Challenges in Learning and Assessing Anesthesia Cognitive & Non-technical Skills. Austin J Anesthesia and Analgesia. 2014; 2: 4.