Performance Assessment

Timothy M. Kowalewski and Thomas S. Lendvay

Introduction

Traditional surgical education has suffered from some long-standing challenges. These include a lack of objectivity or quantitative rigor in performance evaluation and a growing training gap due to tightening resource constraints and concomitant increase in number and diversity of skills requiring mastery. These are compounded by the constant influx of new technologies in the operating room [1], which further challenge the historically arduous, prolonged learning curves associated with surgical skill acquisition (5–7 years) [2]. In the past two decades, technology has augmented surgical education with a variety of simulators and robotic platforms. While these bring new training and learning challenges, they also promise a heightened level of scientific rigor for performance evaluation [3]. This offers further promise of semiautomated mentoring in skills training which can decrease the time, risk, and resource cost of training for students and faculty alike. The need for objective metrics remains pressing, and quantitative rigor is becoming increasingly available [4–6].

Need for Objective Measurements of Skill

Shifting healthcare reimbursement to performance-based compensation, increasing public awareness of variable healthcare quality, rapid adoption of new technologies, and a general trend toward continuous process improvement are all drivers of the need for increasing objectivity in surgical performance assessment.

T. M. Kowalewski (✉)
Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN, USA
e-mail: timk@umn.edu

T. S. Lendvay
Department of Urology, University of Washington, Seattle Children's Hospital, Seattle, WA, USA

The Training Need

Among novice surgeons in training, the ACGME and RRCs provide the direction for tracking individual's performance and maintaining standards for advancement. Despite standard core competencies against which all trainees in residencies are compared, a major challenge in this system has been that advancement – hinged to these core competencies – is still dictated by individual faculty within the program of the trainees [7]. This leads to variability of feedback to the trainees and to subjective biases based on personalities and leaves room for graduates not actually having all the necessary proficiencies to practice safe and effective healthcare.

In the most recent publication distributed by the ACGME regarding the core competency progression of residents from 1 year to the next, the trends were to be expected – residents achieved "graduation" benchmarks across the board for all milestones [8] (Fig. 1).

Using such grading systems alone can make it difficult to hold a trainee back from advancement as most faculty provide higher-"level" scores as the trainees ascend by the program year. In general, faculties are not experts on deciding whether a trainee is a 3 or a 4 out of 5 for interpersonal communication skills. This allows for a high degree of variable feedback scores and the benchmarks against which faculty grade the trainees are ill defined and left up to the Residency Program Directors of the residencies to instruct the faculty how to ideally score. This process is quite different than say a management consulting firm that applies psychological testing and customer feedback as metrics of success and advancement.

Whereas a trainee's advancement relies on faculty-only feedback, once a clinician is in practice, the primary feedback to the practicing clinician comes from self-selected peers usually within the practicing clinician's hospital network or community. Credentialing organizations around the country are struggling to standardize privileging and credentialing guidelines [9]. To date, there is no national standard. The concern is that with a growing number of high profile

D. Stefanidis et al. (eds.), *Comprehensive Healthcare Simulation: Surgery and Surgical Subspecialties*, Comprehensive Healthcare Simulation, https://doi.org/10.1007/978-3-319-98276-2_9

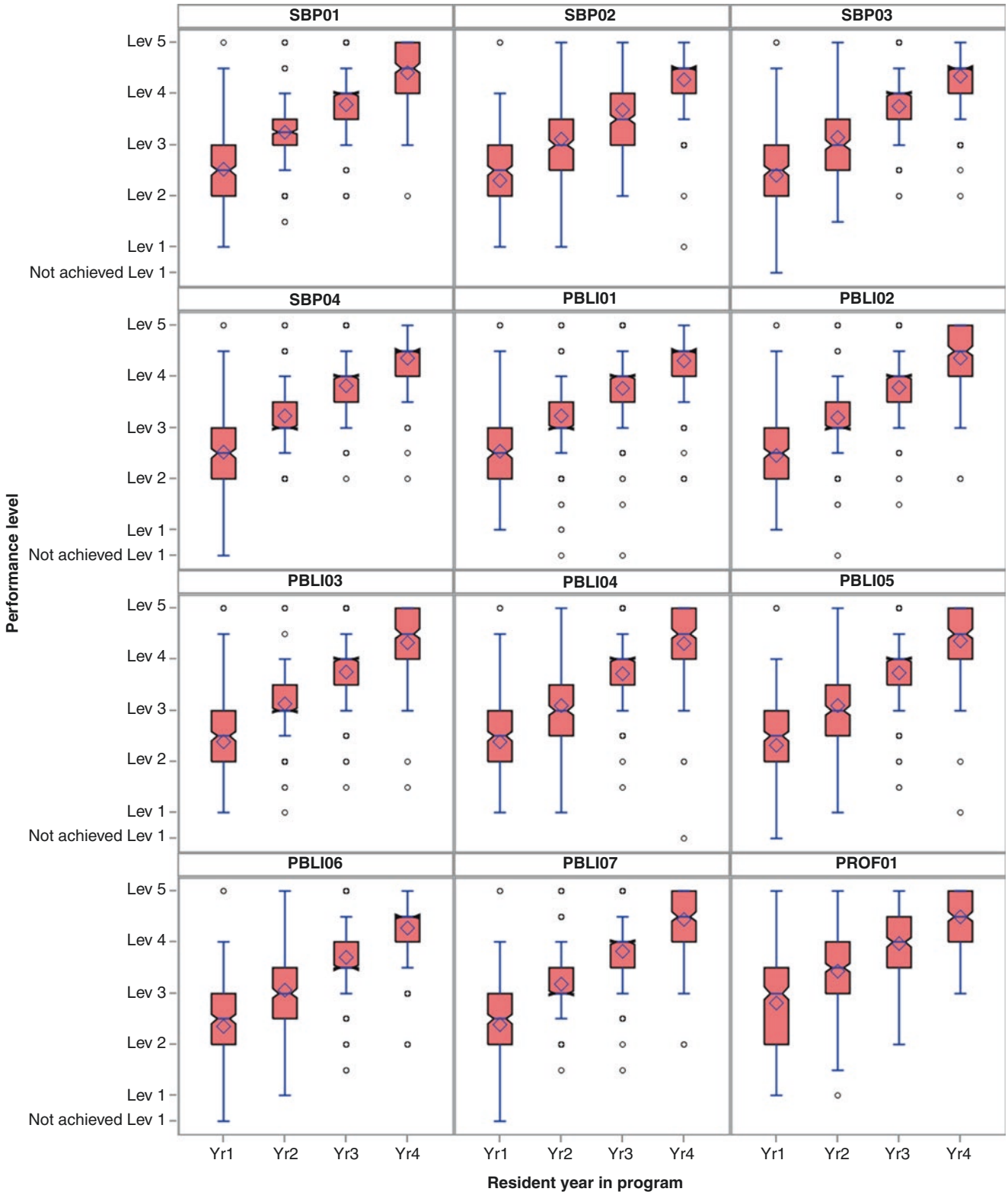

Fig. 1 Nationwide urology resident progression among the core competencies of systems-based practice, problem-based learning, and professionalism over the course of 1 year. (Reproduced with permission of ACGME)

and extremely costly malpractice suits [10] as well as the changing process by which payers reimburse hospitals are taking notice that these practice-granting processes need an infusion of objective methods. Furthermore, physician reimbursement is now being tied to patient satisfaction scores [11]. In order to ensure optimization of physician-patient communication, it will be imperative to utilize non-technical skills or communication skills assessment methods so that administrators overseeing the satisfaction scores can hone in on deficits and provide targeted remediation to these clinicians.

Need for Periodic Recertification of Existing Skills: Skill Decay and Use

Another growing concern among healthcare leaders and providers is the MOC process. Each surgical Board decides how to recertify their members and has a duty to the public to ensure safe and effective providers. In the beginning years of the American Board of Surgery (ABS), if a surgeon wanted to receive board certification, the Board would send a delegate to that surgeon's parent hospital and watch the clinician practice their craft in the operating room and on the wards [12]. After 2 years, this practice was abandoned. It was unscalable and unsustainable, yet the ABS knew that the practice of the clinician was a critical piece to ensuring the quality of the surgeon. The result was what most Boards do today which is administer 5- or 10-year written recertification exams as a means of quality control. These exams are based solely on cognitive skills and not on any technical skills appraisal. The only surrogate for technical skills is through case log submissions and complication reports which are put together by the clinicians themselves and not extracted from an independent data registry. The recertifying surgeon also needs to demonstrate that s/he is acquiring CME credit through regional and national conferences or hands-on course participation. These are passive learning processes and are not held to rigorous standards. Thus, the quality of the clinician recertifying can only be objectively ascertained through a single cognitive test – a sliding scale score based on clinical knowledge of the specialty.

This lack of technical skills appraisal provides evidence of the lag Boards demonstrate in their tracking methods behind the current reimbursement and regulatory environment that the parent hospitals are experiencing. In addition, there is significant variation in practices such that surgeons may have been granted certification or privileging at the beginning of their practices when fresh out of training, but as their practices change, the same recertification processes that were used upon initial certification remain identical. This has impacts on surgeons who sub-specialize, on surgeons who leave practice for a period of time (military deployment, leave of absence for personal reasons, infirmities, increasing administrative or teaching roles), and on the aging surgeon. The one size fits all recertification processes cannot objectively appraise the resultant variability from the above matters.

Technical skills decay as surgeons age [13] and as surgeons redistribute their clinical practices among other competing endeavors [14]. Evidence-based research provides insight into the skills decay phenomenon. No different than a professional athlete or a theatrical arts professional needs to warm up before performances or demonstrates a diminution of skill after long periods of rest, surgeons, too, experience such decays [15, 16]. Despite evidence supporting this reality, because we do not have systems in place to objectively quantify skills in practice, we cannot identify clinicians who may be experiencing skills degradation. And surgical Boards do not have the means to identify surgeons in need. It remains up to the surgeon himself/herself to recognize a skill deficit and either cease practicing that skill or seek remediation avenues.

Definition and Decomposition of Surgical Skills

In order to establish objective assessment of clinicians, a common language must be agreed upon for metrics. This section addresses how surgical skills are decomposed into constituent parts. Researchers have stratified surgical skills with varying degrees of resolution, incorporating insights form a variety of fields spanning education to aircraft pilot training. This has resulted in a nomenclature that can sometimes overlap but nonetheless help clarify the type and role of various skill components in surgery. This vocabulary can also help provide structured guidance to curriculum developers, hospital administrators, trainees, or researchers to focus resources where they may be most impactful.

Outcomes Versus Skills

We define surgical skill as the ability of a surgeon to consistently bring about a desired surgical outcome for a patient independent of patient-specific aspects. The importance of skills to surgery is irrefutable. But patient outcomes are the primary criterion for evaluating surgical success. Measures of skill – even a subset of overall skill like technical skill demonstrated in a single procedure as an indicator of overall practice – have shown to correlate directly to patient outcomes [17]. But "correlate" does not mean "equate." Skill is necessary but not sufficient for positive patient outcomes. There is more to surgery than surgical skill alone. Even a surgical master can make mistakes, and even procedures that are completed without error have unavoidable risks or complications. Having excellent surgical skills will thus maximize but not guarantee successful outcomes. With this in mind, the ultimate importance of different skills or their

constituent parts is determined by the degree to which they positively impact patient outcomes.

Cognitive Versus Psychomotor, Technical, and Nontechnical

Perhaps the most fundamental decomposition of surgical skill is into cognitive and psychomotor skills. Miller's pyramid reproduced in Fig. 2 spans this distinction and stratifies skill from the perspective of an instructor or evaluator [18].

Miller's four-layer pyramid implies that certain skills are foundational; they must be developed before others can be addressed. Typically, a finer degree of granularity is used in the surgical literature in reference to skill acquisition, particularly in simulation. The literature often distinguishes between cognitive and technical skills [19]. According to Miller's pyramid, this would place cognitive skills at the bottom two levels: "knows" and "knows how." Technical skills would belong to the top two layers, "shows how" and "does," with simulation typically falling into the "shows how" layer.

Many of these finer distinctions of technical skill arise due to a change in focus. Whereas Miller's pyramid was constructed primarily from the point of view of the evaluating clinician, the simulation literature moved toward stratifying skills from the perspective of the trainee and his perception. Technical skills are often further stratified into visuospatial and psychomotor skills [20, 21]. Visuospatial skills consist of being able to accurately reconstruct and navigate a 3D environment based on one's depth perception of 2D video that is typically displayed along a different axis than that of the tool interaction. In his comprehensive decomposition of skill categories, Satava further distinguishes psychomotor, visuospatial, perception, and haptic skills [3]. Haptics refers to a subject's ability to perceive haptic (tactile sensory) cues such that resolution of more subtle haptic cues implies stronger haptic abilities.

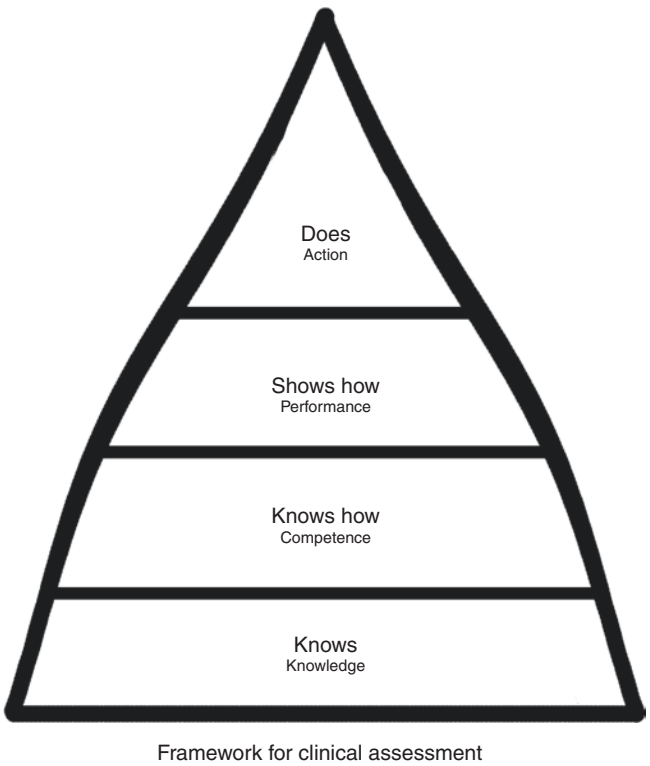

Fig. 2 Miller's pyramid: a "framework for clinical assessment" [18]

Gallagher et al. proposed a hypothetical map of attentional resources across different training levels, reproduced in Fig. 3 [22]. In this map, Gallagher et al. suggest that an individual surgeon has a fixed attentional capacity threshold. A novice surgeon must consciously attend to at least five items: psychomotor performance, depth and spatial judgments, operative judgment and decision-making, comprehending instruction, and gaining additional knowledge. For a typical novice surgeon, the simultaneous combination of these demands is beyond their attentional capacity. As a result, their ability to learn in at least some of these categories is significantly diluted. Gallagher et al. suggest that simulation-based pre-training of novice surgeons can refine technical skills like psychomotor performance and depth and spatial judgments such that most or all of the categories receive sufficient attention. This reasonably supposes that once trained, technical aspects will demand less attention, thus freeing attentional resources for the acquisition of other important skills or knowledge.

Gallagher et al. did not rigorously analyze the process of and neurophysiological elements involved in the relationships between attention, skill categories, and skill acquisition. But the hypothetical attentional resource map finds both conceptual and empirical support in the motor learning literature ("motor" in this field is synonymous with "muscle"). For example, the single channel theory of attention and its supporting evidence reveal that attention demand is usually estimated indirectly by the extent to which the tasks interfere with each other. Processing sensory stimuli (or performing other processes early in the sequence) can apparently be done in parallel, with little interference from other tasks. But processes associated with response selection or with response programming and initiation interfere greatly with other activities [23, p., 121].

Since early stages of surgical training deal heavily with response selection and programming, this supports Gallagher's notion of attentional resource strain. Moreover, "some evidence suggests that directing one's attention to movement or environmental cues may differ according to one's skill level" [23, p., 121]. Also of interest is that "other evidence, based on secondary task techniques, suggests that attention demands are highest at both the initiation and termination stages of movement" [23, p., 121]. Such

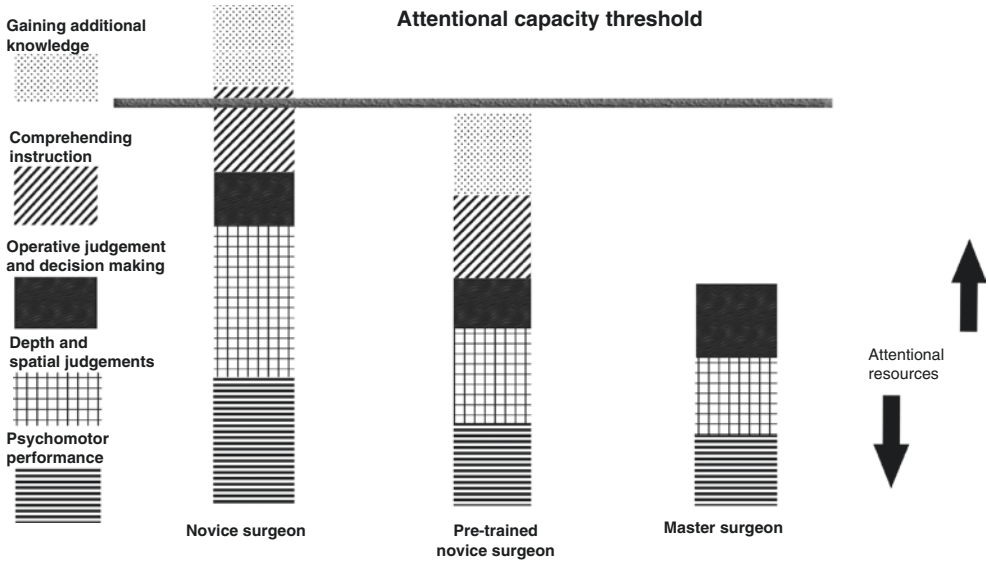

Fig. 3 Gallagher's hypothetical attention resource map indicates the benefits of simulation training. (Reproduced with permission [22])

observations suggest strategies for developing relevant dynamic metrics. However, Schmidt and Lee conclude that "even though attention has had a long history of thinking in psychology, we are still unclear about its nature and the principles of its operation—indeed, even its definition." The motor learning literature recognizes that "learners appear to pass through various stages phases when acquiring skill:

1. The *cognitive phase*, in which emphasis is on discovering what to do, e.g., observing the target motor skill. Trainees are most responsive to verbal instruction or feedback in this stage.
2. The *associative phase*, in which the concern is with perfecting the movement patterns.
3. The *autonomous phase*, in which attentional requirements of the movement appear to be reduced or even eliminated [23, p., 429]."

Human physiology employs numerous senses. Of these, however, the surgeon is essentially limited to three: sight, touch, and somesthesis (i.e., bodily perception). While balance, hearing, and possibly other senses are employed in surgery, the essential three senses identified in Table 1 – sight-related skills like visuospatial localization and depth perception – have often been the object of study in the surgical literature. However, no work exists investigating the role of proprioception in surgical skill acquisition and surgical performance. Yet their importance to technical skill can be elucidated.

Proprioception is crucial to the practice and acquisition of manual motor skills. This is vividly illustrated by the well-documented cases of Ms. G. L. and Mr. Ian Waterman (summarized in [24], original sources [25–28]). These individuals suffered from complete, permanent loss of somesthesis. They could not use proprioceptive senses to localize their body or limbs, only vision could provide this information. However, their efferent neural pathways – those sending control signals *to* muscles – were unaffected. Thus they could exert voluntary muscle control. The following symptoms and phenomena ensued in sequential order:

- Could not walk or stand upright.
- Could move limbs, but could not control them in a precise way.
- When not looking at limbs, did not know their location or if they were moving. Arms (particularly fingers) moved uncontrollably. Sometimes arms would unwittingly hit own self.
- Using constant visual tracking, could eventually learn some control over muscles, but learning was very slow, difficult, and demanded inordinate attention.
- Relearning to sit up took 2 months.
- Relearning to stand took 1.5 years longer.

Table 1 Human senses and the subset of senses available to surgeons

Sense categories	Human senses	Senses used in surgery
Exteroceptive senses	Sight	Sight
	Taste	
	Smell	
	Touch (tactile, heat, forces)	Touch
	Hearing	[a]
	Balance (vestibular sense)	[a]
Interoceptive senses	Pain	
	Movement of organs	
Proprioception	Somesthesis (body/limb localization)	Somesthesis

[a]While balance is critical for standing or sitting during surgery and providing orientation, beyond this, it does not contribute to dynamic surgical activity. Hearing is of utility in surgery, but not crucial to its performance

- Relearning to walk took several additional months. However, he could only walk with slow, somewhat awkward steps and only while looking at his feet.
- When visual information was suddenly removed, immediately fell to the floor (e.g., lights unexpectedly switched off).
- Decades later, still relies exclusively on vision for control. Controlled limb motions are still slow and ponderous, and hands are primarily restricted to only three fingers.
- Typically uses excessive force when holding objects, especially if not looking at them.
- Eventually learned to avoid falling to the floor due to sudden removal of visual information by exerting incredible, conscious effort to tense many muscles. Attempting this for a few minutes resulted in complete mental and physical exhaustion, requiring several days of rest and recovery.
- Tasks involving simultaneous cognitive load and fine motor control nearly exceeded the limits of his attentional capacity (e.g., could not write during dictation, had to constantly switch between listening and attempting to write).

The ramifications of these phenomena for surgical skill are profound. Clearly, proprioception is essential to surgical skill proficiency. This alone implies proprioception in surgery should be actively studied. The inordinate attention required in the above cases is empirical evidence that strongly corroborates Gallagher's hypothetical attentional resource hypothesis. Also, it is evident that somatosensory activity is a key component to surgical skill learning and performance. This strongly suggests that proprioception may yield a universal (cross-procedure, cross-modality) dynamic metric for surgical skill. Thus, proprioception should be better understood.

Proprioceptive somesthesis consists of several sensor groups and multiple neurological centers to which they relay data ([23], Chap. 5). These sensors include:

- Vestibular system: senses internal acceleration or rotation of the head (this sense infers the exteroceptive direction of gravity since gravity registers as an acceleration).
- Muscle receptors: muscle spindles innervate the fleshy part of the muscle and sense stretching position and velocity; Golgi tendon organs innervate the tendons, sense contraction, and have been shown to respond to forces less than 0.1 g.
- Joint receptors: are suspected to sense specific joint positions, joint extremes, continuous joint position, and/or joint velocity. However, there is much uncertainty about whether or how this comes to pass.
- Cutaneous receptors: sense deep or superficial pressure in the skin which often correlates to muscle or limb information as well as touch. Additionally, this group includes temperature, pain, and chemical stimuli. However, it has been shown that primary somesthesis in not affected by these later pathways.

The neurological centers where the sensors send their information to and along what pathway include *(listed reaction times are round trip)*:

- Spinal cord (via spindles): myotatic reflexes, effect individual muscles *(30–50 ms)*
- Cerebellum and cortex (via spindles): long loop reflexes, effect individual muscles *(50–80 ms)*
- Higher centers (via receptors): triggered reactions, effect associated musculature *(80–120 ms)*; reaction time, effect any musculature *(120–180 ms)*

Vision, on the other hand, is a much slower process. Motor control pathways that include vision feedback have reaction times ranging from *200 ms to 3 s*, depending on the type of visual stimulus and type of motion involved. These data apply to natural vision tasks. However, vision in MIS is significantly limited since it comes from a 2D image, typically viewed well off-axis from the original 3D task space. Of the typical visual cues for depth perception, only parallax, depth from motion, perspective, relative size, occlusion, texture gradient, and lighting/shading are available to the surgeon. Cues like familiar size, accommodation, foveal distortion, and inferred overhead lighting are not available. This, compounded with the typically imperfect lighting and picture quality in MIS video, implies that the data available to visual sense and perception is atypically limited and that visuospatial localization from depth perception requires more time, attention, and learning, especially for MIS trainees. This suggests that in MIS the *minimum* reaction time for the visual feedback loop is in fact longer than 200 ms. Moreover, in the case of novice surgeons, visual feedback loop times would be significantly longer, and the information may not be completely reliable as evidenced by common depth perception errors in early training.

MIS tools and the related fulcrum effect effectively alter the kinematic chain of the human limb and end effector. For a first time user, the immediate result is that proprioceptive perception and control must adapt to the novel kinematics. If a novice would not have somesthetic perception and somatomotor control well refined, he would depend exclusively on vision to track both tool and target – as was born out in the study. This would fall into the classic closed-loop motor learning theory reviewed in the motor learning literature, characterized by its precision and slow speeds. As the proprioception and related control adapt to the new kinematics and somesthetic tracking becomes more reliable, the subject needs to confirm tool tracking via vision less and less. At the

expert level, target gaze is dominant, and proprioception allows both faster overall tracking and faster, more accurate motor control. However, it is very unlikely this process would continue until a schema or open-loop control strategy is acquired. Unlike fast, precise schemas that have taken years to develop for virtuoso piano playing or high-speed professional sports activity – both are cases where high precision and high-speed performance are only possible via schemas – surgery requires a higher level of precision in more degrees of freedom, moves at a slower pace, and exhibits much greater variability. This essentially precludes the notion of surgical schemas.

The result of the above discussion implies a hierarchical control structure is chiefly active in surgical training, especially in MIS. The neurophysiological analysis and relevant evidence reviewed above allows us to construct a relatively accurate system diagram. The multiple feedback blocks and their respective reaction times suggest a major loop/minor loop control strategy exists [29]. This method is a classic, well-documented way of effectively combining dynamic systems of disparate reaction times. The inner, minor loop traditionally operates at much faster dynamics than the outer, major one (e.g., the stabilizers on supersonic jets require very fast dynamics to suppress vibration and turbulent disturbances, while the pilot's commands have a much smaller bandwidth). The inner one is tuned in such a way that the outer loop's optimal tuning is easy to realize. This can be implemented recursively, as illustrated in the system block diagram below (Fig. 4). Note the feedback loop response times are indicated.

Thus learning a surgical task first relies on vision-based feedback control. Progress involves learning to make sense of proprioceptive information and training somatomotor centers to use this information during motion. Eventually, dependency on visual tracking is reduced, as evidenced in the eye-tracking study. This enables target gaze, where eyes fix strictly on a target, while proprioceptive feedback motor control drives a tool to target. This affords at least two benefits. First, the eyes do not need to switch back and forth between target and tool to realize tracking. Since visual feedback takes (at least) twice as long to incorporate than somesthesis, this would seriously compound the delay time involved in task tracking. Second, the proprioceptive feedback can directly drive somatomotor control centers. Because this loop is 2–10 times as fast as the visual feedback loop, psychomotor performance can be significantly faster. Thus, proprioception is critical in surgical performance and skill acquisition. In fact, the degree to which a surgeon exploits internalized proprioception in favor of visual processing alone is a measure of psychomotor skill.

Typical Use Cases of Skill Metrics: Ideal Requirements

New Technology Certification

Since the introduction of laparoscopy in the mid-1980s, surgery has seen a rapid rate of new surgical technologies being employed, sometimes outpacing adequate training. In 1999, the FDA approved the use of the da Vinci surgical robot that has since transformed whole areas of surgical disciplines. Laparoscopy and robotic surgery never passed through a rigorous training and efficacy testing process. Surgeons who were early adopters decided to do their next cases using these technologies after fairly minimal training or proctoring. Despite high-profile malpractice cases and surgical complications related to the inadequate training of surgeons using these technologies and approaches, there are no standard pathways in place for surgeons to adopt new technologies. Each hospital decides which surgical approaches and technologies warrant special credentialing processes, and each hospital is different. Furthermore, the processes in place for new technology credentialing typically involves sign-off from peers in the institution with whom the surgeon is befriended, thus eliminating objectivity in the process of proficiency assessment.

Objective skills appraisal provides a common ground against which all surgeons adopting new technologies in the operating room can be compared. In an ideal professional situation, surgeons would need to show competency and proficiency in the use of a new technology before using it in a human patient. The reality is that access to physician expertise, available time and resources for the training, and the

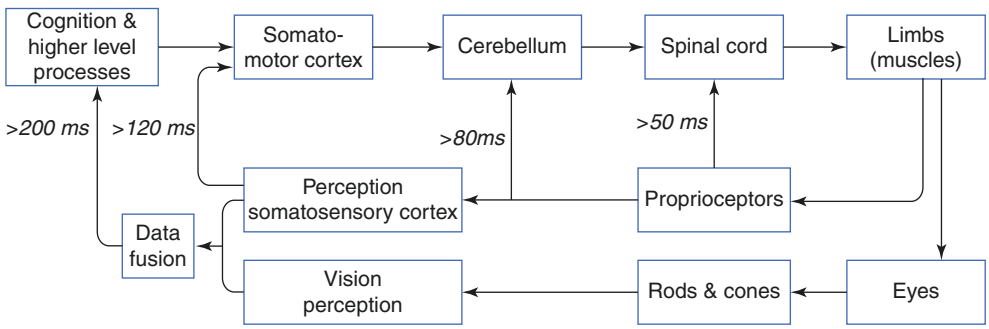

Fig. 4 System block diagram of human motor tracking

surgeon's overestimation of their own capabilities lead to new technology utilization before adequate skill is achieved. This may place patients at risk of harm. The FDA is working to establish guidelines for medical device companies to demonstrate that their new technology is not only safe and effective but also that it is usable by the surgeons. The FDA is calling out to professional leaders and key opinion leaders to encourage their hospitals to embrace [30].

Identifying Best Targets for Effective Remediation

The best targets for remediation are those skills that are universally required for competent performance and can be objectively assessed with validated tools and where clear feedback can yield change. The most basic and fundamental metric is task time. Although extremely easy to track and undisputable across any skill, speed of practice does not always confer safe surgery. Furthermore, giving feedback to a trainee that they are too slow may incentivize poor technique in exchange for improved task time. Other metrics such as bimanual dexterity represent a hard skill that can be objectively measured; directly influences task time, efficiency, and safety; and can be improved with training [3]. When watching a performance, it is immediately evident whether both hands are being used to complete a task. Also, poor bimanual dexterity has been validated a metric that can confer expertise – the higher the degree of bimanual dexterity, the higher the expertise. Bimanual dexterity can also be assessed in an automated way through tool motion or hand motion tracking; thus immediate feedback can be given to the learner about their performance in this skills domain. When deficits are observed, there are multiple low and high fidelity drills that exercise bimanual dexterity for minimally invasive and open approaches. Other areas of skill that meet the criteria of best targets for effective remediation include:

- Depth perception
- Control of instrumentation (laparoscopic, robotic, open)
- Force sensitivity or tissue handling (although this one requires human observation)

Summative Versus Formative Feedback

Assuming that the various components of skill can be measured accurately and readily, how will such information be presented to trainees, established surgeons, or risk assessment department to best improve patient outcomes? The amount of time elapsed from the completion of a procedure can govern this. Gallagher's hypothetical map (see Fig. 3) [22] was suggested primarily as a means to motivate pretraining of surgical trainees via simulation, that is, to hone their technical skills like tool handling and cognition of the procedural flow before joining their attending surgeon in the operating room. But this implies that implementing validated, objective metrics for technical skills can be used to evaluate whether surgeon trainees are ready for higher-level instruction or learning based on their available attentional resources. This would suggest proving skill evaluation information preemptively, *before* an operation ever takes place.

Another approach is to maintain records or data logs of surgical procedures (e.g., recorded videos, compiled ratings, simulator databases) and periodically process this data to provide a summative feedback to a trainee, practicing surgeon, or risk assessment department. This has the potential to link performance to outcomes but only retrospectively. This may occur with varying levels of delay: annually, monthly, or even shortly after the end of a procedure. Conversely, formative feedback would provide meaningful input on skill or performance more proximally – perhaps immediately upon the completion of a procedure or, even better, during a given procedure. This would have the benefit of making the skill evaluation data most relevant and actionable to a consumer. Individuals could learn more immediately from their mistakes or successes or even while they are occurring.

This leads to the ideal case for formative feedback of zero time delay, that is, virtually real-time measurement and structured feedback on skill, that is, "what is the skill rating at any moment within a surgical task?" and "what can one change in this instant and context to improve?" and not just summary (more summative) information such as total time upon completion of a task or procedure. This could ultimately accelerate or mitigate the prolonged, arduous learning curves associated with surgical skill acquisition.

Aggregation Versus Individuation of Skill and Context and Skill Decay

The impact of surgical skills to patient outcomes is a function of both context and time. For example, technical mastery of suturing can be targeted as a particularly critical skill. The manual dexterity required to master suturing in manual laparoscopy can ensure some dexterity in simpler technical tasks: if one masters suturing, he or she must implicitly be at least passable in other aspects like basic tissue manipulation. But while the success of some procedures hinges on suturing mastery, others may be able to completely avoid it or complete such procedures with comparable outcomes using tools like the Autosuture™ (Covidian Corp. Dublin, Ireland) that obviate the need for traditional suturing. The resulting impact that mastery of suturing has on ultimate patient outcomes is thereby also function of the specific procedure in question. While this relative importance of a specific skill like suturing

mastery to the patient outcome depends on the wider procedural context, the understanding of the skill in and of itself is also context dependent. For example, skill changes with time: a trainee's mastery of suturing increases with practice, but a master surgeon's level of technical proficiency can also decay with lack of use.

The level of granularity between aggregation and individuation can apply to an individual surgeon (e.g., their entire practice), a specific class of procedures (e.g., all of the appendectomies they have ever performed), a specific procedure on a particular date, specific steps or minutes within that procedure, and across the various components of skill (e.g., cognitive vs. psychomotor vs. visuospatial). In practice, aggregation (combining of performance evaluations from multiple dates or for a given performance from multiple raters using a method like averaging or median) provides more reliable, statistically stable results as it avoids the prevalence of outliers since spurious events like occasionally erroneous ratings or unusually extreme performances can cancel out. However, this introduces a necessary drawback: the more aggregation occurs over time intervals, the less formative (immediate) the feedback can be, somewhat hampering its possible utility. This delay can also overlook issues like identifying decayed skills that need a quick warm up. Feedback averaged over an entire practice may not provide the most up-to-date assessment of a surgeon's skill. Conversely, the evaluation of a single segment from a single procedure may not accurately reflect that surgeon's entire practice or aptitude in other procedural contexts. Independent of the level of aggregation, the principle of extremes can still apply. For example, a risk assessment department can look at a histogram of all surgical technical skills evaluated for a given procedure and identify the extremes: e.g., the top and bottom quartiles. This can identify individuals most and least deserving of additional resources for training and improvement. Then for a particular individual, a more individuated assessment, say, for the riskiest steps of a given procedure, can assess which of their component skills are weakest, e.g., "respect for tissue," and target very specific resources to improve them.

Methods of Surgical Skill Measurement

Determining methods to reliably, objectively, and quantitatively measure surgical skill remains an active area of research. While numerous approaches have been proposed over more than two decades, few have yet established widespread use. This is particularly true for more technology-dependent computational approaches that promise most quantitative rigor. However, with the increasing popularity of robotics and continual incursion of advanced technologies into the operating room, it is reasonable to expect that such methods will penetrate into practice.

Subjective Versus Objective Metrics

Barring technology and automation, earlier methods such as the objective structured assessment of technical skill (OSATS) employed manual, subjective evaluation of performance via expert review of video-recorded procedures [31, 32]. Objectivity was argued based on a consistent checklist and preset Likert scale evaluations with categories such as "respect for tissue," "time and motion," "instrument handling," "respect of instruments," etc. Such methods are equally applicable in both simulation and real surgical environments and scale well across the different tasks or modalities (e.g., robotics, laparoscopy, endoscopy, open surgery, etc.). However, they require a human proctor to manually evaluate each individual's tasks which is expensive and does not scale well to large numbers or concurrent trials. Multiple variants of OSATS have become practical de facto standards for skill assessment; the core concept of anonymized video review with structured survey instruments employing Likert scales remains the same, but some Likert domains may be slightly altered for specific surgical procedures or specialties. Examples include the global operative assessment of laparoscopic skills (GOALS) instrument for laparoscopy [33] and the global evaluative assessment of robotic surgery for robotic surgery [34]. Such approaches also invariably suffer from the subjectivity of the evaluator's judgment and imperfect inter- and intra-rater agreements. On the other hand, they are more objective than traditional in person "over the shoulder" subjective evaluations. This is due to blinding raters to the identity of surgeons whose performances they evaluate through videos, the aggregation of multiple ratings, and consistent textual descriptors used to anchor provided ratings. However, such tools are not as objective as rigorous quantitative algorithms. For example, the same panel of OSATS raters may provide slightly different scores to the same video at different times, whereas a quantitative method would provide the same deterministic score for each performance.

Methods to overcome barriers of scale for objective assessment of large groups of surgeons have been developed employing crowdsourcing to assess surgeon skill. Chen et al. first described posting a single robotic suturing video to a large group of distributed, independent, anonymous crowdworkers to rate the performance using a validated robotic skills assessment tool. When compared to a panel of expert robotic surgeons reviewing the same video, the crowd of presumably nonmedically trained crowdworkers agreed with the expert ratings. Furthermore, instead of the 3 weeks it took the experts to do the survey, it took the crowd of almost 500 people less than 24 h to complete the survey [35]. This methodology for objective skills assessment has since been validated for open, laparoscopic, and robotic animate, human, and dry lab surgery skills [36–44]. The enabling capability of crowdsourcing is evidenced by the consistently inexpensive and rapid results that mirror expert reviews.

Proficiency Benchmarks

Proficiency methods are based on the repetition of tasks or procedures until predetermined performance criteria have been met. To set the performance criteria on some criterion tasks like suturing, a pool of "expert surgeons" completes multiple repetitions, and their resulting scores are averaged. A trainee must score within 1 standard deviation of their average score at least two consecutive times to achieve proficiency. This approach deals well with the large amount of variability inherent within and among subjects, and applications of proficiency-based methods have spread beyond VR since their introduction to surgery. It is from within the corpus of VR surgical simulation studies that proficiency-based evaluation and training arose [22]. However, this approach suffers from some problems as well. The proficiency benchmarks tend to be highly task specific: two different tasks intended to evaluate suturing skills (e.g., a virtual reality simulation and a reality-based Fundamentals of Laparoscopic Surgery (FLS) suturing task) will provide different "task-specific" scores. This means that proficiency criteria must be established for each task. Furthermore, the choice of the "expert subjects" and their resulting performance can vary significantly as no universal criteria are established or espoused in selecting them: two groups of experts from different geographic locations may yield different proficiency criteria perhaps because they teach different suturing techniques, e.g., how to tie knots or hold the suture and needle. Ideally, skill evaluation metrics would move beyond tallying task-specific events to seeing the "skill" exhibited in the task – something that structured survey tools like OSATS can better cope with.

Technical Skills (Psychomotor, Visuospatial)

The act of surgery invokes numerous human physiological systems during its execution by a surgeon. Of those specifically identified in the surgical literature (e.g., Miller's pyramid, Gallagher's attentional resource chart), technical skills are most easily amenable to traditional scientific measurement and observation. Cognitive skills can, for the most part, be directly assessed with traditional examinations. While cognitive skills, knowledge, and sensory perception are important in surgery, their inaccessibility via direct observation precludes them from convenient scientific investigation. As a result, technical skills have received the most research effort to date.

In Simulation

Virtual reality (VR) was introduced into surgical simulation in 1993 [45] and continues to be adopted, evaluated, and improved as a tool for training and measuring surgical skill with varying degrees of granularity from its outset [6, 46–50]. In simulation, the benefits of VR include the ability to deploy the same environment between subjects and tasks and so offer a consistent training platform for trainees, low cost of long-term use, ease in data collection, and ease of tracking the virtual environment. Drawbacks include high initial cost, steep cost increases for better realism in visual representation, internal modeling or haptic rendering, and the inability to extract similar data from real cases. The bulk of the surgical literature in the VR simulation area has focused primarily on validation. That is, in establishing that skills acquired during simulation trials ultimately transfer to operating room (OR) performance. These validation studies rely almost exclusively on summary metrics like task time, path length, and economy of motion (path length divided by task time or similar efficiency measure) and provide typically positive but sometimes mixed results about the validity of simulators to train OR-transferable surgical skills [51].

In terms of metrics, VR natively supports automation and objectivity in recording metrics, more so than in reality-based procedures or simulations. Time to task is automatically computed along with more novel tool path metrics such as path length, economy of motion, smoothness, etc. Recording complete tool trajectories is trivial. Such information can provide a rich source for dynamic analysis, though this source of data and its subsequent, potential dynamic analysis are basically ignored. Because VR systems synthesize their environments, tracking of virtual tissue and objects and how they are interacted with is also trivial. Thus, once the expense of creating the environment is incurred, it is inexpensive to automate the accurate detection of both procedural and cognitive errors in VR. This is a major benefit of VR.

Reality-based (RB) simulators consist of physical objects that either mimic anatomy with varying degrees of realism or simply provide inexpensive, nonanatomical objects as a means for basic manipulation. These simulators employ real surgical tools used in the OR or slightly modified versions. Perhaps the most notable of these is the McGill Inanimate System for Training and Evaluation of Laparoscopic Skills (MISTELS). It originally consisted of seven laparoscopic tasks (peg transfers, pattern cutting, clip and divide, endolooping, mesh placement and fixation, suturing with intracorporeal or extracorporeal knots) executed on inexpensive materials like gauze, rubber grommets, latex gloves, tubing, and foam. The original purpose of MISTELS was to develop a series of structured tasks to objectively measure laparoscopic skills [20, 21]; these tasks were not necessarily developed to systematically accelerate or optimize the learning curves for skill acquisition. The chief metrics used in MISTELS are task time and an error penalty. These metrics are combined into a single score based on the following formula (Table 2):

Score = preset constant – completion time – penalty

Table 2 Equations used to compute FLS scores per Task with t for task time and E for task-specific error counts; derived from [20, 52, 53]

FLS task	FLS score
Peg transfer	$FLS_{Peg} = (300 - t - 17E_{dr})/237$
Cutting	$FLS_{Cut} = (300 - t - 2E_{a})/280$
Suturing	$FLS_{Sut} = (600 - t - E_{pd} - E_{g} - E_{q})/520$

Both the preset constant (cutoff time) and penalty are unique to each of the seven tasks. MISTELS was successfully validated with varying degrees of granularity [54–59]. Eventually, the Fundamentals of Laparoscopic Skills (FLS) committee, mandated in the late 1990s by the Society of American Gastrointestinal and Endoscopic Surgery (SAGES), adopted the MISTELS program with the exception of two tasks (clipping tubular structure and securing a mesh were found to lack utility) [19]. Since this adoption, a number of studies ensued to reinforce the validation of the MISTELS/FLS paradigm [52, 60–65]. Most notably, given proficiency-based training, translation of skills to the OR was established [66, 67] along with positive evidence for its utility in skill retention and maintenance [53, 68].

FLS and similar RB simulators are less expensive than VR simulators because they require less technology and do not need to invest resources to accomplish realism in accurate models or visual and haptic rendering. As such, validation only considers the metrics used for skill scoring and does not need to address the quality of realism in simulation since the subject is already interacting with real-world objects. However, the acquisition of metrics typically requires manual oversight for timing and particularly with evaluating errors for task-specific penalty scores. FLS trainers, like most RB methods, do not utilize tool path analysis, neither for summary metrics like path length and economy of motion nor for dynamic metrics or force information.

Robotics provides a platform in which dry lab simulations and OR procedures can both be logged in an identical manner and yield consistent, automatically generated metrics. This would be ideal for validation studies of dry lab or realistic VR training skills transferred to the operating theater. However, Intuitive Surgical, Inc. (Sunnyvale, CA), the company that currently deploys the vast majority of surgical robotic platforms, does not have universal open access to the data streams internally collected during operation. Some work is underway for creating VR tasks intended to train or evaluate robotic skills which resemble FLS constructs, but these are not as developed or validated as the FLS program and remain an active area of research at this time [69–71]. If dynamic metrics are successfully created based on tool trajectories from VR or RB simulation, they would be naturally well-suited to extend into surgical robotics.

Computational Metrics

Computational metrics obviate the need for human raters. They operate on quantitative data actively streaming or previously recorded from the operating room. This can include continuous video and a variety of tool tracking variables like tool tip and handle positions, orientations, and forces. Such data are generated either via customized sensors as in early work [72] but more commonly through existing computerized systems to which such data are already inherent; the increasingly ubiquitous da Vinci surgical robot (Intuitive Surgical, Inc.) is an example. This area of research has been highly active and continues to make significant progress [73, 74].

The basic approach employs methods from machine learning. This includes constructing a sophisticated mathematical model and "training" it with data captured from surgery that is labeled according to skill ranked level (e.g., novice, expert, intermediate). Then the ability of the model to quantify skill is evaluated by testing it with entries that were not part of the training set to emulate what a real-world situation would be like: the model must analyze data it has never seen before. This process is called cross-validation. The resulting models are typically said to classify skill level when referring to discrete predetermined skill levels such as novice or expert. Alternatively, they are said to quantify or score skill level when they provide a score that can take on a continuum of values instead of discrete categories. In this literature, the word metric and measure take on very specific, narrow mathematical meanings that are not compatible with the wider sense of the words in the surgical literature. This area of research is primarily hampered by a dearth of rich datasets that capture the massive variability of surgical practice, skills, and regionally varying techniques. To date, no computational methods have shown to predict patient outcomes. However, some techniques have recently been applied that effectively automate OSATS – a technique shown to correlate to patient outcomes – directly on raw video (from dry lab procedures) with surprising accuracy [75].

Among the most mature accomplishments in this area to date is the study by Ahmidi and colleagues [76] which summarizes the problems of automatically segmenting a surgical task into constituent sub-parts and atomic surgical gestures called "gestemes." More importantly, it also establishes a formal standard for validating the success of computational metrics, leave-surgeon-out cross-validation (also called leave-one-user-out or LOUO), and provides an open dataset captured from the da Vinci robot. This is particularly important given the scarcity of such data and the fact that surgeons vary so widely in their captured data.

Typical metrics such as procedural errors, task time, accuracy, blood loss, fluid use, etc. are specific not only to a par-

ticular task or procedure (e.g., FLS peg transfer or cutting, etc.) but are also specifically fixed to a certain modality. For example, the amount of blood loss may be cheaply computed in VR but may be difficult or impossible within RB, robotics, or traditional manual MIS.

Since the 1970s, hidden Markov models (HMMs) have enjoyed considerable success in computer speech recognition and voice identification [77]. They also showed promising results when applied to robotics problems such as human task segmentation or task identification [78–82]. Hannaford and Rosen successfully applied Markov modeling techniques to surgical skill/performance evaluation [83–86] in part by developing the Blue-DRAGON [87–89] data capture device and a subsequent, smaller version known as the Red-DRAGON [90] (see Fig. 5). The Blue-DRAGON employed a novel spherical mechanism and was used to record a large database of surgeon-tool interactions for common laparoscopic procedures executed in live porcine models. This exposed surgery to modern signal processing and led to validating the Markov modeling approach for surgical skill recognition [91]. Both the Red-DRAGON and the use of HMMs for surgical skill evaluation were eventually licensed and commercialized as the Electronic Data Generation for Evaluation (EDGE) machine by Simulab Corp (Seattle, WA).

The EDGE platform (Fig. 6) was used to collect data from hundreds of FLS task recordings across more than ten geographically diverse training hospitals in the United States. The motion data is ten-dimensional (tooltip position in x,y,z, tool rotation and grasp angle for both hands) and sampled at 30 Hz. Tool path plots of a peg transfer task for disparate skill levels reveal characteristic distinctions in refined vs. crude motion (see Fig. 7). Similar interesting nuances can be seen in the grasping force plots (Fig. 8).

The use of HMMs for surgical performance measurement and processing has gained considerable momentum since its inception at the Biorobotics Lab. This was primarily at Johns Hopkins University [92–94], but development has spread internationally [95, 96]. The strong reception of surgical Markov modeling in academia has spurred research activity in this field. While this academic success lends credibility to this method, it also may introduce alternative models which could potentially outmode classical HMMs by offering better performance in surgical applications [97].

Some earlier robotics studies from the University of Nebraska proposed some more intuitive metrics [98–100].

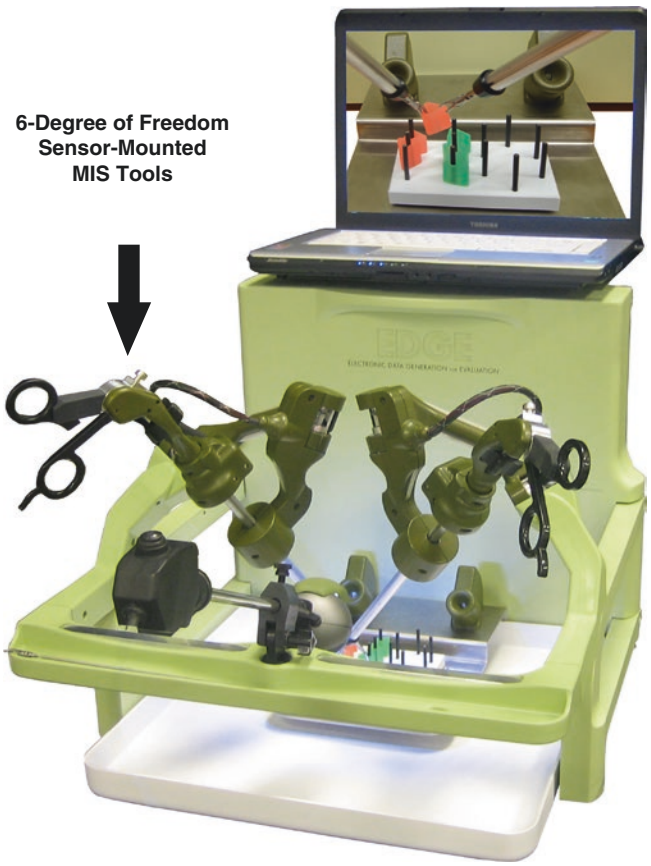

Fig. 6 Simulab's award-wining EDGE platform, a commercialized version of the Red-DRAGON. (Used with permission of Simulab Corporation)

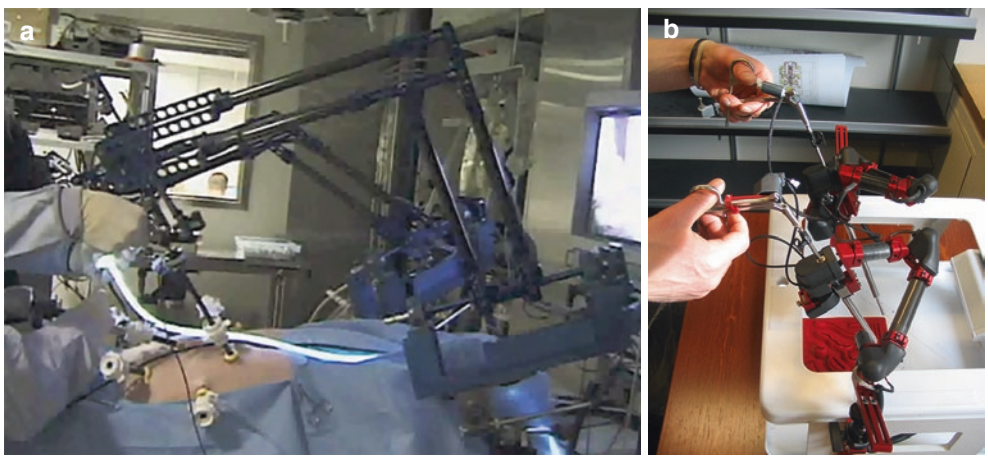

Fig. 5 The Blue-DRAGON collecting data during surgical training in live pigs (**a**) and the subsequent, smaller Red-DRAGON [90] (**b**) in use on an artificial tissue model. (**a**) Used with permission of Jacob Rosen; (**b**) used with permission of Scott Gunther

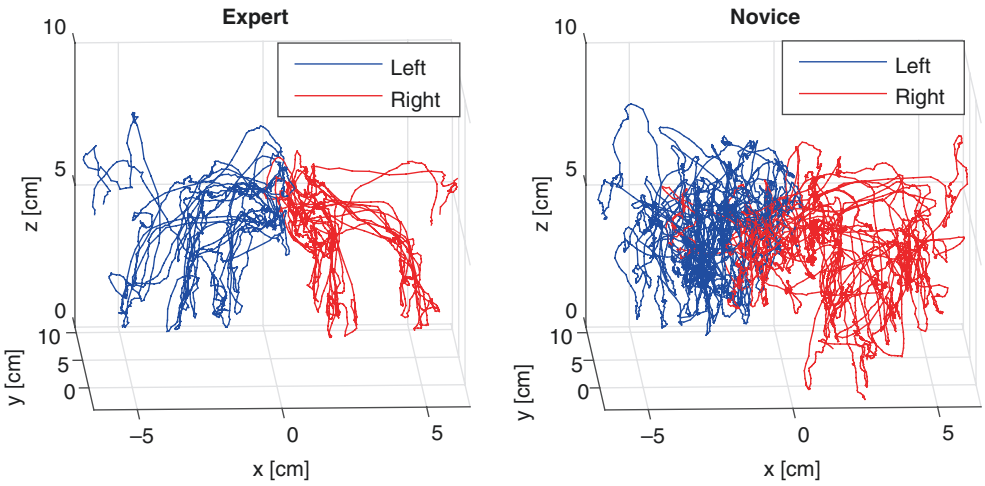

Fig. 7 3D plots of tool path in Cartesian space. The expert data (left) indicates refined, deliberate motion, ambidexterity, and economy of motion. The novice data (right) reveals the right hand (red line) dominates and consistently crosses into the left hand region to compensate

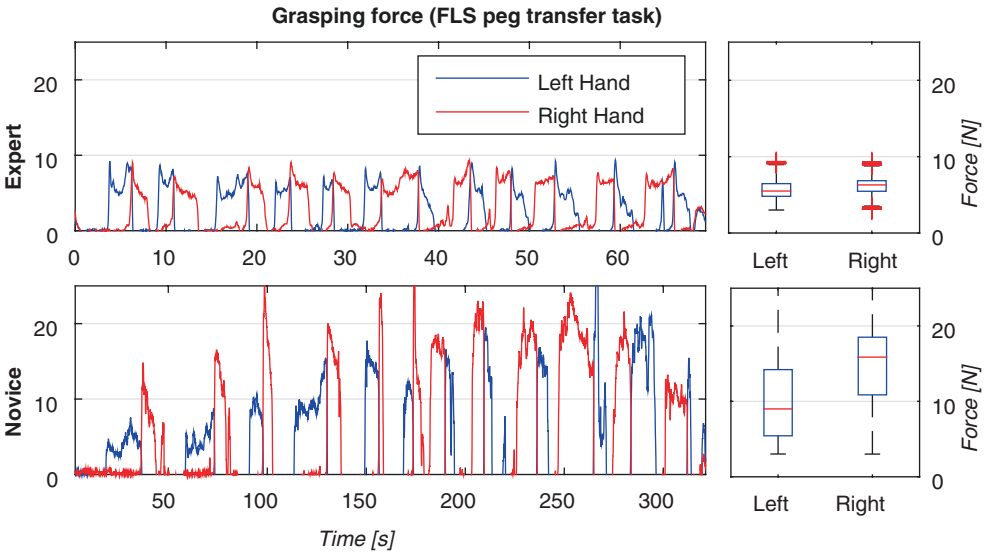

Fig. 8 Plots of left and right hand grasping force in FLS peg transfer tasks recorded with EDGE. Boxplots shown for left and right hands separately for all grasping forces recorded above 3 N to segment grasp events

Movement time intervals (e.g., time spent reaching for an object, time spent holding, etc.) and the coefficient of their variation allowed for finer granularity in temporal analysis. Another metric is the radius of curvature of the trajectory computed from the three-dimensional trajectory of a point and its time derivatives. Phase portraits of position vs. displacement were suggested for bimodal analysis. From the phase portrait, the suggested mean absolute relative phase (MARP) value, which measures the extent to which tools are out of phase (moving in opposite directions), was found to be significant (in phase registers with lower MARP, out of phase induces higher MARP). Moreover, electromyogram (EMG) signals were evaluated and also indicated a correlation to skill level. Historically, static metrics were predominant in the literature, with task time being the most prevalent. Any of the listed platforms that compute economy of motion (EoM) and/or tool path implicitly acquire and potentially log time-dependent tool path data. However, such metrics were potentially found to have little or no value over task time [101].

Another interesting branch of inquiry comes from eye tracking [102]. For example, five novices and five experts were presented with a VR laparoscopic targeting task where a target appeared in a laparoscopic simulation and they were to touch the target in minimal time with a laparoscopic tool. To see if the performance differences between groups were accompanied with eye movement differences, researchers looked at the amount of eye gaze on the tool and then characterized their eye behavior through eye and tool movement profiles. In terms of eye gaze behavior, novices tended to gaze at the tool longer than experts. Several eye gaze behaviors identified in this study, including target gaze, switching, and tool following, are similar to previous findings. The target gaze behavior was the preferred strategy for experts, and novices tended to follow the tool more frequently than experts [102]. Figure 9 and Table 3 demonstrate these phenomena.

There are several ramifications of this study in light of the surgical and motor learning literature reviewed above. First,

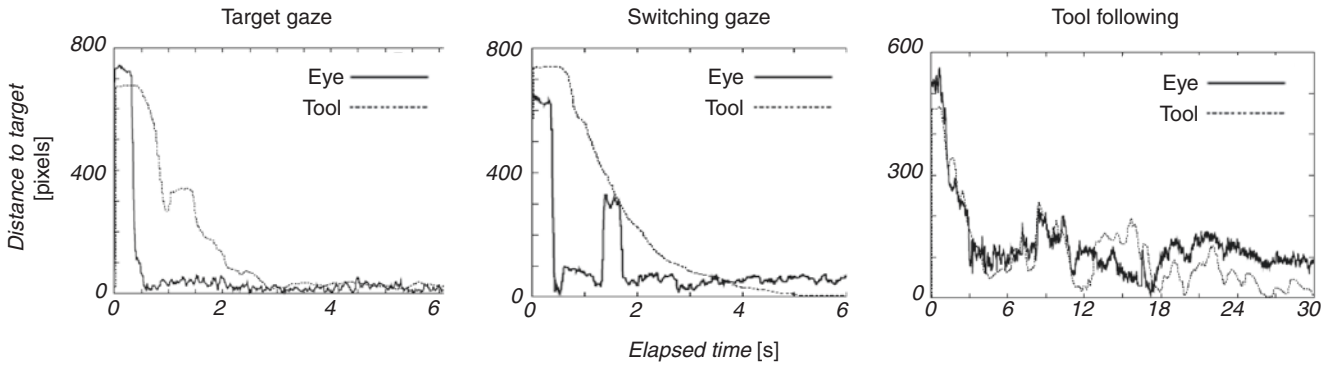

Fig. 9 Eye vs. tool movement profiles typifying different types of gaze patterns taken from [102]. (Reproduced with permission). Eyes gaze at (1) the target, left; (2) eyes switch between target and tool, center; and (3) eyes follow tool during motion, right

Table 3 Eye movement behavior distributions for expert and novices over all trials found in [102]

Group	Target gaze	Switching gaze	Tool following	Loss
Expert	73.3%	13.3%	8.9%	4.4%
Novice	53.3%	17.8%	26.7%	2.2%

the differences in the movement profiles and their associated task times corroborate the notion of Gallagher's attentional resources; the tool following profile of a novice indicates active attentional focus on the tool, while the target gaze of the expert suggests a level of autonomy in the manipulation task. Second, the difference in gaze and targeting patterns across skill levels, as suggested in the motor learning literature reviewed, is reproduced here in a VR laparoscopic setting. And third, this presents strong evidence of open-loop control in the expert (and hence faster performance) vs. closed-loop control in the novices, at least in the sense of a visual feedback loop.

This same study also makes the following two important observations [102]:

- Laparoscopic tool movement is unlike direct hand movement because proprioceptive feedback from hand position does not map directly to the tool tips necessitating additional visuomotor and spatial transformations [103].
- Tactile feedback from tool movement is minimal because of friction between the tool and the cannula (a tunnel-like structure surrounding the tool at the patient entry point), and thus, the surgeon has a greater reliance on the indirect visual information [104, 105].

Exploiting eye tracking in establishing metrics of surgical skill remains an active area of research and recently includes more rigorous methodologies for computational extraction [106].

Nontechnical Skills (NOTSs)

We have focused on technical skills which represent skills centered on a surgeon's kinematic signatures or hand/tool motions, yet surgical success also involves effective communication and human-human interaction – commonly referred to as nontechnical skills. Recent literature has started to address how to distinguish nontechnical surgical skills (NOTSs) such as effective communication, leadership, cooperation, read-backs, and team choreography [107]. These elements can be assessed through objective scoring tools validated in the literature. Clinical areas such as catastrophe or code environments, anesthesia team management, and urgent complex clinical care scenarios have been the initial benefactors of such assessment [108]. These types of scenarios tend to be practiced in simulation centers, yet some have advocated for in situ training scenarios so that any equipment or resource deficits existing on the wards/in the ORs can be unmasked during the simulated team training.

Operationalizing the assessment can be challenging, however, as video and audio from multiple vantage points may need to be obtained to capture the whole room, extensive time is required for coaches/instructors to debrief the teams, and the scenarios themselves can create quite stressful environments which subjects need to reconcile. In addition, in situ training involving patients introduces the concerns around maintaining patient privacy and HIPAA compliance. Thus most in situ scenarios still involve standardized patients or mannequins.

It is clear that effective communication leads to improved team dynamics. And the link to patient outcomes has been indirectly confirmed through malpractice evidence whereby a number of claims in surgery have been related to poor communication; whether between clinician and patient or provider-provider [16]. Operative choreography will become a metric for entire teams [109]. Systems-based training will parallel military training experience that has benefited from decades of evidence to support its value.

Currently there is a dearth of computational or quantitative tools to automatically process NOTSs. While such "soft skills" were traditionally only perceptible or analyzable to humans, this is slowly changing. For example, automatic speech recognition was historically perceived in the same way. But it is now a mature field of research with increasingly dependable algorithms that have become inexpensive and ubiquitous (e.g., Apple's Siri voice assistant). Key aspects of NOTSs are not just what is being said but how it is being said. This includes not only the efficiency of communication or correctness of language but also tone or emotional content – aspects that were historically incomputable. However, new branches of computer science and engineering are actively gaining momentum such as affective computing that can computationally grapple with such aspects [110, 111]. In the interim, however, crowdsourcing methods which have already found considerable success in evaluating surgical technical skills are immediately suitable for providing such evaluations more automatically and objectively than expert human raters [112].

Conclusions

The technology and knowledge exist to elevate the objectivity in a clinician's skill, both technical and nontechnical. And we know that the skill of the surgeon influences patient outcomes. Yet, the utilization of objective performance assessment has lagged awareness. There are many barriers to standard assessment including cost, time, and expertise. The onus is on thought leaders in the field of objective skills assessment to enlighten practicing surgeons and organizations tasked with establishing certification, credentialing, and privileging with a unified method for skills appraisal. Until there is agreement on cost-effective, universally agreed upon standards to capture surgeon performances and provide objective, iterative feedback that helps surgeons improve their skills, resistance will exist. Furthermore, we as surgeons should proactively figure out standard feedback methods before regulatory bodies comprised of non-clinicians decide for us how we are to be assessed.

References

1. Batalden P, Leach D, Swing S, Dreyfus H, Dreyfus S. General competencies and accreditation in graduate medical education. Health Aff. 2002;21(5):103.
2. Liu A, Tendick F, Cleary K, Kaufmann C. A survey of surgical simulation: applications, technology, and education. Presence Teleoperators Virtual Environ. 2003;12(6):599–614.
3. Satava RM, Cuschieri A, Hamdorf J. Metrics for objective assessment. Surg Endosc. 2003;17(2):220–6.
4. Darzi A, Smith S, Taffinder N. Assessing operative skill. BMJ. 1999;318(7188):887–8.
5. Satava RM. The need for metrics in surgical education. Surg Endosc. 1999;13(11):1082.
6. Gallagher AG, Satava RM. Virtual reality as a metric for the assessment of laparoscopic psychomotor skills. Surg Endosc. 2002;16(12):1746–52.
7. Nasca T, Philibert I, Brigham T. The next GME accreditation system—rationale and benefits. New Engl. J. 2012;366(11):1051–6.
8. Stanley LE Hamstra J, Yamazaki K, Holmboe ES. Milestones: Annual Report. Accreditation Council for Graduate Medical Education (ACGME), Chicago; 2016.
9. DeMaria EJ, El Chaar M, Rogers AM, Eisenberg D, Kallies KJ, Kothari SN. American Society for Metabolic and Bariatric Surgery position statement on accreditation of bariatric surgery centers endorsed by the Society of American Gastrointestinal and Endoscopic Surgeons. Surg Obes Relat Dis. 2016;12(5):946–54.
10. Tzafestas SG. Medical roboethics. In: Roboethics. Switzerland: Springer; 2016. p. 81–92.
11. Long C, Tsay EL, Jacobo SA, Popat R, Singh K, Chang RT. Factors associated with patient press ganey satisfaction scores for ophthalmology patients. Ophthalmology. Switzerland: Springer; 2016;123(2):242–7.
12. Buyske J. Forks in the road: the assessment of surgeons from the American Board of Surgery Perspective. Surg Clin North Am. 2016;96(1):139–46.
13. Bhatt NR, Morris M, O'Neil A, Gillis A, Ridgway PF. When should surgeons retire? Br J Surg. 2016;103(1):35–42.
14. Deering SH, Rush RM, Lesperance RN, Roth BJ. Perceived effects of deployments on surgeon and physician skills in the US Army medical department. Am J Surg. 2011;201(5):666–72.
15. Lendvay TS, et al. Virtual reality robotic surgery warm-up improves task performance in a dry laboratory environment: a prospective randomized controlled study. J Am Coll Surg. 2013;216(6):1181–92.
16. Levinson W, Roter DL, Mullooly JP, Dull VT, Frankel RM. Physician-patient communication: the relationship with malpractice claims among primary care physicians and surgeons. JAMA. 1997;277(7):553–9.
17. Birkmeyer JD, et al. Surgical skill and complication rates after bariatric surgery. N Engl J Med. 2013;369(15):1434–42.
18. Miller GE. The assessment of clinical skills/competence/performance. Acad Med J Assoc Am Med Coll. 1990;65(9 Suppl):S63.
19. Peters J, et al. Development and validation of a comprehensive program of education and assessment of the basic fundamentals of laparoscopic surgery. Surgery. 2004;135(1):21–7.
20. Derossis MD, et al. Development of a model for training and evaluation of laparoscopic skills. Am J Surg. 1998;175(6):482–7.
21. Derossis AM, Bothwell J, Sigman HH, Fried GM. The effect of practice on performance in a laparoscopic simulator. Surg Endosc. 1998;12(9):1117–20.
22. Gallagher AG, et al. Virtual reality simulation for the operating room: proficiency-based training as a paradigm shift in surgical skills training. Ann Surg. 2005;241(2):364.
23. Schmidt RA, Lee TD. Motor control and learning: a behavioral emphasis. Champaign: Human Kinetics Publishers; Switzerland: Springer; 2005.
24. Robles-De-La-Torre G. The importance of the sense of touch in virtual and real environments. Multimedia, IEEE. 2006;13(3):24–30.
25. Cole J. Pride and a Daily Marathon. A Bradford Book. MIT press. Cambridge, MA 1995.
26. Craig JC, Rollman GB. Somesthesis. Annu Rev Psychol. 1999;50(1):305–31.
27. Cole J, Paillard J. Living without touch and peripheral information about body position and movement: studies with deafferented subjects. In: The body and the self. Cambridge, MA: The MIT Press. p. 245–66.
28. Paillard J. Body schema and body image: A double dissociation in deafferented patients. Mot Control Today Tomorrow. 1999;48(3):197–214.

29. Nise NS. Control systems engineering, (With CD). Hoboken: Wiley; Switzerland: Springer; 2007.
30. Eydelman MB, Nguyen T, Green JA. The US Food and Drug Administration's new regulatory toolkit to bring medical device innovation back to the United States. JAMA Ophthalmol. 2016;134(4):353–4.
31. Martin JA, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg. 1997;84(2):273–8.
32. Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via an innovative 'bench station' examination. Am J Surg. 1997;173(3):226–30.
33. Vassiliou MC, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. Am J Surg. 2005;190:107–13.
34. Goh AC, Goldfarb DW, Sander JC, Miles BJ, Dunkin BJ. Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. J Urol. 2012;187(1):247–52.
35. Chen C, et al. Crowd-sourced assessment of technical skills: A novel method to evaluate surgical performance. J Surg Res. 2014;187(1):65–71.
36. Holst D, et al. Crowd-sourced assessment of technical skills: Differentiating animate surgical skill through the wisdom of crowds. J. Endourol. 2015;29(10):1183–8.
37. Aghdasi N, Bly R, White LW, Hannaford B, Moe K, Lendvay TS. Crowd-sourced assessment of surgical skills in cricothyrotomy procedure. J Surg Res. 2015;196(2):302–6.
38. Holst D, et al. Crowd-sourced assessment of technical skills: An adjunct to urology resident surgical simulation training. J Endourol. 2015;29(5):604–9.
39. Kirsch S, Comstock B, Warren J, Schaffhausen C, Kowalewski T, Lendvay T. Crowd Sourced Assessment of Technical Skills (CSATS): A Scalable Assessment Tool for the Nursing Workforce. J Invest Med. 2015;63(1):92.
40. Lendvay TS, White L, Kowalewski T. Crowdsourcing to assess surgical skill. JAMA Surg. 2015;150(11):1086–7.
41. Deal SB, et al. Crowd-sourced assessment of technical skills: an opportunity for improvement in the assessment of laparoscopic surgical skills. Am J Surg. 2016;211(2):398–404.
42. Chen SP, et al. Optical biopsy of bladder Cancer using crowd-sourced assessment. JAMA Surg. 2016;151(1):90–3.
43. Kowalewski TM, et al. Crowd-Sourced Assessment of Technical Skills for Validation of Basic Laparoscopic Urologic Skills Tasks. J Urol. 2016;195(6):1859–65.
44. Ghani KR, et al. Measuring to improve: peer and crowd-sourced assessments of technical skill with robot-assisted radical prostatectomy. Eur Urol. 2016;69(4):547–50.
45. Satava RM. Virtual reality surgical simulator. The first steps. Surg Endosc. 1993;7(3):203.
46. Healy GB. The college should be instrumental in adapting simulators to education. Bull Am Coll Surg. 2002;87(11):10.
47. Champion HR, Gallagher AG. Surgical simulation – a 'good idea whose time has come'. Br J Surg. 2003;90(7):767–8.
48. Gallagher AG, Richie K, McClure N, McGuigan J. Objective psychomotor skills assessment of experienced, junior, and novice laparoscopists with virtual reality. World J Surg. 2001;25(11):1478–83.
49. Watterson JD, Beiko DT, Kuan JK, Denstedt JD. A randomized prospective blinded study validating Acquisition of Ureteroscopy skills using a computer based virtual reality Endourological simulator. J Urol. 2002;168(5):1928–32.
50. Seymour NE, et al. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. Ann Surg. 2002;236(4):458.
51. Sutherland LM, et al. Surgical simulation: a systematic review. Ann Surg. 2006;243(3):291.
52. Fraser SA, Feldman LS, Stanbridge D, Fried GM. Characterizing the learning curve for a basic laparoscopic drill. Surg Endosc. 2005;19(12):1572–8.
53. Scott DJ, Ritter EM, Tesfay ST, Pimentel EA, Nagji A, Fried GM. Certification pass rate of 100% for fundamentals of laparoscopic surgery skills after proficiency-based training. Surg Endosc. 2008;22(8):1887–93.
54. Fraser SA, Klassen DR, Feldman LS, Ghitulescu GA, Stanbridge D, Fried GM. Evaluating laparoscopic skills: setting the pass/fail score for the MISTELS system. Surg Endosc. 2003;17(6):964–7.
55. Keyser EJ, Derossis AM, Antoniuk M, Sigman HH, Fried GM. A simplified simulator for the training and evaluation of laparoscopic skills. Surg Endosc. 2000;14(2):149–53.
56. Derossis AM, Antoniuk M, Fried GM. Evaluation of laparoscopic skills: a 2-year follow-up during residency training. Can J Surg. 1999;42(4):293.
57. Feldman LS, Hagarty SE, Ghitulescu G, Stanbridge D, Fried GM. Relationship between objective assessment of technical skills and subjective in-training evaluations in surgical residents* 1. J Am Coll Surg. 2004;198(1):105–10.
58. Feldman LS, Sherman V, Fried GM. Using simulators to assess laparoscopic competence: ready for widespread use? Surgery. 2004;135(1):28.
59. Fried GM, Derossis AM, Bothwell J, Sigman HH. Comparison of laparoscopic performance in vivo with performance measured in a laparoscopic simulator. Surg Endosc. 1999;13(11):1077–81.
60. Stefanidis D, Sierra R, Korndorffer JR, others. Intensive continuing medical education course training on simulators results in proficiency for laparoscopic suturing. Am J Surg. 2006;191(1):23–7.
61. Feldman LS, Cao J, Andalib A, Fraser S, Fried GM. A method to characterize the learning curve for performance of a fundamental laparoscopic simulator task: defining. Surgery. 2009;146(2):381–6.
62. Dauster B, et al. Validity of the MISTELS simulator for laparoscopy training in urology. J Endourol. 2005;19(5):541–5.
63. Swanstrom LL, Fried GM, Hoffman KI, Soper NJ. Beta test results of a new system assessing competence in laparoscopic surgery. J Am Coll Surg. 2006;202(1):62–9.
64. Stefanidis D, Korndorffer JR, others. Proficiency maintenance: impact of ongoing simulator training on laparoscopic skill retention. J Am Coll Surg. 2006;202(4):599–603.
65. Fried GM, et al. Proving the value of simulation in laparoscopic surgery. Ann Surg. 2004;240(3):518.
66. Korndorffer JR, others. Simulator training for laparoscopic suturing using performance goals translates to the operating room. J Am Coll Surg. 2005;201(1):23–9.
67. Ritter EM, Scott DJ. Design of a proficiency-based skills training curriculum for the fundamentals of laparoscopic surgery. Surg Innov. 2007;14(2):107.
68. Castellvi AO, Hollett LA, Minhajuddin A, Hogg DC, Tesfay ST, Scott DJ. Maintaining proficiency after fundamentals of laparoscopic surgery training: a 1-year analysis of skill retention for surgery residents. Surgery. 2009;146(2):387–93.
69. Sethi AS, Peine WJ, Mohammadi Y, Sundaram CP. Validation of a novel virtual reality robotic simulator. J Endourol. 2009;23(3):503–8.
70. Kenney PA, Wszolek MF, Gould JJ, Libertino JA, Moinzadeh A. Face, content, and construct validity of dV-trainer, a novel virtual reality simulator for robotic surgery. Urology. 2009;73(6):1288–92.
71. Lendvay TS, Casale P, Sweet R, Peters C. Initial validation of a virtual-reality robotic simulator. J Robot Surg. 2008;2(3):145–9.
72. Rosen J, MacFarlane M, Richards C, Hannaford B, Sinanan M. Surgeon-tool force/torque signatures evaluation of surgical skills in minimally invasive surgery. Med meets virtual reality-the Converg Phys Informational Technol options a New Era Healthc. 1999;62:290–6.
73. Reiley CE, Lin HC, Yuh DD, Hager GD. A review of methods for objective surgical skill evaluation. Surg Endosc. 2011;25(2):356–66.

74. Chmarra MK, Grimbergen CA, Dankelman J. Systems for tracking minimally invasive surgical instruments. Minim Invasive Ther Allied Technol. 2007;16(6):328–40.
75. Zia A, Sharma Y, Bettadapura V, Sarin EL, Clements MA, Essa I. Automated assessment of surgical skills using frequency analysis. In: International conference on medical image computing and computer-assisted intervention. Munich, Germany, 5–9 October 2015. p. 430–8.
76. Ahmidi N, Tao L, Sefati S, Gao Y, Lea C, Haro BB, Zappella L, Khudanpur S, Vidal R, Hager GD. A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. IEEE Trans Biomed Eng. 2017;64(9):2025–41.
77. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE. 1989;77(2):257–86.
78. Inamura T, Tanie H, Nakamura Y. From stochastic motion generation and recognition to geometric symbol development and manipulation. In: International conference on humanoid robots. Karlsruhe Munich, Germany. 2003.
79. Itabashi K, Hirana K, Suzuki T, Okuma S, Fujiwara F. Modelling and realization of the peg-in-hole task based on hidden Markov model. In: Robotics and Automation, 1998. Proceedings. 1998 IEEE International Conference on, 1998, vol. 2. Trieste, Italy. p. 1142–7.
80. Yang J, Xu Y, Chen CS. Human action learning via hidden Markov model. Syst Man Cybern Part A Syst Humans, IEEE Trans. 1997;27(1):34–44.
81. Hannaford B, Lee P. Hidden Markov model of force torque information in Telemanipulation. Int J Robot Res. 1991;10(5): 528–39.
82. Inamura T, Toshima I, Tanie H, Nakamura Y. Embodied symbol emergence based on mimesis theory. Int J Robot Res. 2004;23(4–5):363–77.
83. Kowalewski TM, Rosen J, Chang L, Sinanan M, Hannaford B. Optimization of a vector quantization codebook for objective evaluation of surgical skill. In: Proceeding of medicine meets virtual reality 12. Newport, CA. 2004. p. 174–9.
84. Rosen J, Chang L, Brown JD, Hannaford B, Sinanan M, Satava R. Minimally invasive surgery task decomposition – etymology of Endoscopic Suturing. Stud Heal Technol Informatics Med Meets Virtual Real. 2003;94:295–301.
85. Rosen J, Hannaford B, Richards CG, Sinanan MN. Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills. Biomed Eng IEEE Trans. 2001;48(5):579–91.
86. Rosen J, Solazzo M, Hannaford B, Sinanan M. Task decomposition of laparoscopic surgery for objective evaluation of surgical residents' learning curve using hidden Markov model. Comput Aided Surg. 2002;7(1):49–61.
87. Rosen J, Brown JD, Chang L, Barreca M, Sinanan M, Hannaford B. The {BlueDRAGON}-a system for measuring the kinematics and dynamics of minimally invasive surgical tools in-vivo. Proceedings 2002 IEEE International Conference on Robotics and Automation, vol. 2. Washington, DC. p. 1876–81.
88. Lum M. Kinematic optimization of a 2-{DOF} spherical mechanism for a minimally invasive surgical robot. Masters thesis. University of Washington, Department of Electrical Engineering; Switzerland: Springer; 2004. Accessible at: http://astro.ee.washington.edu/BRL_Pubs/Pdfs/Th029.pdf.
89. Rosen J, Lum M, Trimble D, Hannaford B, Sinanan M. Spherical mechanism analysis of a surgical robot for minimally invasive surgery – analytical and experimental approaches. Stud Heal Technol Informatics Med Meets Virtual Reality. Jan. 2005;111:422–8.
90. Gunther S, Rosen J, Hannaford B, Sinanan M. The {R}ed {DRAGON}: a multi-modality system for simulation and training in minimally invasive surgery. Stud Health Technol Inform. 2007;125:149.
91. Rosen J, Brown JD, Chang L, Sinanan MN, Hannaford B. Generalized approach for modeling minimally invasive surgery as a stochastic process using a discrete markov model. Biomed Eng IEEE Trans. 2006;53(3):399–413.
92. Kragic D, Marayong P, Li M, Okamura AM, Hager GD. Human-machine collaborative systems for microsurgical applications. Int J Robot Res. 2005;24(9):731–41.
93. Li M, Okamura AM. Recognition of operator motions for real-time assistance using virtual fixtures. In: Haptic Interfaces for Virtual Environment and Teleoperator Systems, 2003. HAPTICS 2003. Proceedings. 11th Symposium on. 2003, p. 125–31.
94. Lin HC, Shafran I, Yuh D, Hager GD. Towards automatic skill evaluation: detection and segmentation of robot-assisted surgical motions. Comput Aided Surg. 2006;11(5):220–30.
95. Megali G, Sinigaglia S, Tonet O, Dario P. Modelling and evaluation of surgical performance using hidden Markov models. Biomed Eng IEEE Trans. 2006;53(10):1911–9.
96. Dosis A, Bello F, Gillies D, Undre S, Aggarwal R, Darzi A. Laparoscopic task recognition using hidden markov models. Stud Health Technol Inform. 2005;111:115–22.
97. Reiley CE, et al. Automatic recognition of surgical motions using statistical modeling for capturing variability. Stud Health Technol Inform. 2008;132:396.
98. Judkins T, Oleynikov D, Stergiou N. Objective evaluation of expert performance during human robotic surgical procedures. J Robot Surg. 2008;1(4):307–12.
99. Oleynikov D, Judkins TN, Stergiou N. Objective evaluation of expert and novice performance during robotic surgical training tasks. Surg Endosc. 2009;23(3):590–7.
100. Narazaki K, Oleynikov D, Stergiou N. Robotic surgery training and performance: identifying objective variables for quantifying the extent of proficiency. Surg Endosc. 2006;20(1):96–103.
101. Kowalewski TM, et al. Beyond task time: Automated measurement augments fundamentals of laparoscopic skills methodology. J. Surg. Res. 2014;192(2):329–38.
102. Law B, Atkins MS, Kirkpatrick AE, Lomax AJ. Eye gaze patterns differentiate novice and experts in a virtual laparoscopic surgery training environment. In: Proceedings of the Eye tracking research applications symposium on Eye tracking research applications ETRA2004. 2004, vol. 1, no. 212, p. 41–8.
103. MacKenzie CL, Graham ED, Cao CG, Lomax AJ. Virtual hand laboratory meets endoscopic surgery. Stud Health Technol Inform. 1999;62:212–8.
104. Cuschieri A. Visual displays and visual perception in minimal access surgery. Semin Laparosc Surg. 1995;2(3):209–14.
105. Ibbotson JA, MacKenzie CL, Cao CG, Lomax AJ. Gaze patterns in laparoscopic surgery. Stud Health Technol Inform. 1999;62:154–60.
106. Ahmidi N, Hager G, Ishii L, Fichtinger G, Gallia G, Ishii M. Surgical task and skill classification from eye tracking and tool motion in minimally invasive surgery. Med Image Comput Comput Interv. 2010;2010:295–302.
107. Yule S, Flin R, Maran N, Rowley D, Youngson G, Paterson-Brown S. Surgeons' non-technical skills in the operating room: reliability testing of the NOTSS behavior rating system. World J Surg. 2008;32(4):548–56.
108. Marshall SD, Mehra R. The effects of a displayed cognitive aid on non-technical skills in a simulated 'can't intubate, can't oxygenate' crisis. Anaesthesia. 2014;69(7):669–77.
109. Bharathan R, Aggarwal R, Darzi A. Operating room of the future. Best Pract Res Clin Obstet Gynaecol. 2013;27(3):311–22.
110. Tao J, Tan T. Affective computing: A review. In: International conference on affective computing and intelligent interaction. 2005. p. 981–95.
111. Picard RW, Picard R. Affective computing, vol. 252. Cambridge: MIT press; 1997.
112. Borish M, Cordar A, Foster A, Kim T, Murphy J, Chaudhary N, Lok B. Utilizing real-time human-assisted virtual humans to increase real-world interaction empathy. KEER 2014 Conference. Linköping, Sweden. June 10–13.