Surg Endosc (2015) 29:3261–3266
DOI 10.1007/s00464-015-4070-8

External validation of Global Evaluative Assessment of Robotic Skills (GEARS)

Monty A. Aghazadeh · Isuru S. Jayaratna · Andrew J. Hung · Michael M. Pan · Mihir M. Desai · Inderbir S. Gill · Alvin C. Goh

Received: 21 September 2014 / Accepted: 8 January 2015 / Published online: 22 January 2015

Abstract
Background We demonstrate the construct validity, reliability, and utility of Global Evaluative Assessment of Robotic Skills (GEARS), a clinical assessment tool designed to measure robotic technical skills, in an independent cohort using an in vivo animal training model.
Methods Using a cross-sectional observational study design, 47 voluntary participants were categorized as experts (>30 robotic cases completed as primary surgeon) or trainees. The trainee group was further divided into intermediates (≥5 but ≤30 cases) or novices (<5 cases). All participants completed a standardized in vivo robotic task in a porcine model. Task performance was evaluated by two expert robotic surgeons and self-assessed by the participants using the GEARS assessment tool. Kruskal–Wallis test was used to compare the GEARS performance scores to determine construct validity; Spearman's rank correlation measured interobserver reliability; and Cronbach's alpha was used to assess internal consistency.
Results Performance evaluations were completed on nine experts and 38 trainees (14 intermediate, 24 novice). Experts demonstrated superior performance compared to intermediates and novices overall and in all individual domains ($p < 0.0001$). In comparing intermediates and novices, the overall performance difference trended toward significance ($p = 0.0505$), while the individual domains of efficiency and autonomy were significantly different between groups ($p = 0.0280$ and 0.0425, respectively). Interobserver reliability between expert ratings was confirmed with a strong correlation observed ($r = 0.857$, 95 % CI [0.691, 0.941]). Experts and participant scoring showed less agreement ($r = 0.435$, 95 % CI [0.121, 0.689] and $r = 0.422$, 95 % CI [0.081, 0.0672]). Internal consistency was excellent for experts and participants ($\alpha = 0.96, 0.98, 0.93$).
Conclusions In an independent cohort, GEARS was able to differentiate between different robotic skill levels, demonstrating excellent construct validity. As a standardized assessment tool, GEARS maintained consistency and reliability for an in vivo robotic surgical task and may be applied for skills evaluation in a broad range of robotic procedures.

Keywords Robotics · Validation studies · Clinical competence · Education

M. A. Aghazadeh · A. C. Goh (✉)
Department of Urology, Methodist Institute for Technology, Innovation, and Education, Houston Methodist Hospital, 6560 Fannin Street, Suite 2100, Houston, TX 77030, USA
e-mail: acg622@gmail.com

M. A. Aghazadeh · M. M. Pan
Scott Department of Urology, Baylor College of Medicine, Houston, TX, USA

I. S. Jayaratna · A. J. Hung · M. M. Desai · I. S. Gill
USC Institute of Urology, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

In the field of urology, robotic surgery has had an unprecedented rate of adoption, with over 360,000 surgical procedures performed with the da Vinci Surgical System (Intuitive Surgical; Sunnyvale, CA) in 2011. It is estimated that 67 % of all prostatectomies are currently being performed robotically with a clear upward trend [1]. With such rapid adoption of this new surgical modality, novel methods of robotic surgical training and skill assessment are needed.

In recent years, it has been shown that the traditional "apprenticeship" model of surgical training is no longer

tenable, due to financial, legal, and ethical issues [2]. There has been a new emphasis placed on developing standardized training curricula with objective benchmarks to define competency. While gains have been made in laparoscopic surgery [3], there is still no widely adopted assessment tool to gauge robotic surgical skill. As various robotic surgical simulators are being developed and validated [4–9], demonstration of true concurrent or predictive validity as it pertains to robotic clinical performance is impossible in the absence of such a standardized method to assess surgical skills. As such, there is limited data to support the actual impact that training tasks and simulators developed for robotic surgical skills have on clinical performance.

In recognition of this need, Global Evaluative Assessment of Robotic Skills (GEARS), a clinical assessment tool for robotic surgical skills, was developed and validated in an intraoperative environment. [10] Modeled after the Global Operative Assessment of Laparoscopic Skills (GOALS) [11], GEARS consists of six domains (depth perception, bimanual dexterity, efficiency, force sensitivity, autonomy, and robotic control) that are scored on a 5-point Likert scale with anchors at one, three, and five. When used by an observer experienced in robotic surgery to score trainees, GEARS has been shown to reliably quantify technical proficiency [9, 10, 12]. Despite its promise and potentially vital role in surgical education, data demonstrating the strength and reliability of GEARS as a robotic surgical assessment tool have been limited to few studies. [9, 10, 12, 13] We herein externally evaluate the reliability and consistency of GEARS as a robotic skills assessment tool in an in vivo animal model.

Materials and methods

Participants

After institutional review board approval, a cohort of residents and urologists who attended a surgical training course was asked to participate in this study. Study design is shown in Fig. 1. Once enrolled, participants completed a pre-study questionnaire to collect demographic and surgical experience information. Subjects were then categorized a priori as experts or trainees. Although no standard definition of an expert robotic surgeon exists, based on previous studies, [12, 14] experts (minimal standard) were defined as having completed >30 robotic cases as primary surgeon. To further investigate the ability of the assessment tool to differentiate between different skill levels, the trainee group was stratified into intermediates (≥5, but ≤30 cases) and novices (<5 cases).

Standardized task

During a single event, all participants completed a standardized in vivo task using the da Vinci® surgical robot in a porcine model (Fig. 2). The task was developed using expert robotic surgeon input to include sufficient complexity, incorporate multiple basic technical skills, and ensure repeatability. The objective of the task was to locate a defined section of small bowel, maneuver it to a marked area of peritoneum overlying the kidney, and suture it in place with a secure square knot. Each performance was independently evaluated using GEARS by two expert robotic surgeons and self-evaluated by the operator themselves, generating three GEARS scores for each participant. Expert robotic surgeon observers did not know the identity of the operators and were therefore blinded to their experience level.

Statistical analysis

Demographics were compared across groups using Chi-squared/Fisher's exact tests for categorical variables and Kruskal–Wallis test for continuous variables. Construct validity was evaluated using the Kruskal–Wallis test to compare expert, intermediate, and novice GEARS performance scores. The mean of the two expert evaluator scores was used for this comparison. *P* values for pairwise comparisons were adjusted using the Holm method for multiple comparisons. Interobserver reliability was assessed between expert and participant observations using Spearman's rank correlation. Internal consistency of each domain of GEARS was estimated using Cronbach's alpha. A p value <0.05 was considered as statistically significant. All statistical analysis was performed with SPSS® (Chicago, IL, USA).

Results

Demographics and surgical experience of the participants are shown in Table 1. Sixty-four participants completed the standardized in vivo task. Complete scoring information was available for 47 subjects (24 novices, 14 intermediates, and nine experts), which was used for the final analysis. Novices (median age 30 years) were 75 % male, had a median of 0 (0–2) years experience with robotic surgery, and completed a median number of 0 (0–3) robotic cases as primary surgeon. Intermediate participants (median age 30.5 years) were 78.6 % male, had a median of 1.5 (0.5–4) years robotic experience, and had completed a median of 10 (5–30) cases. Lastly, experts (median age 42.5 years) were all male, had 5.5 (4–9) years robotic experience, and had completed a median of 350 (150–2,000) cases as

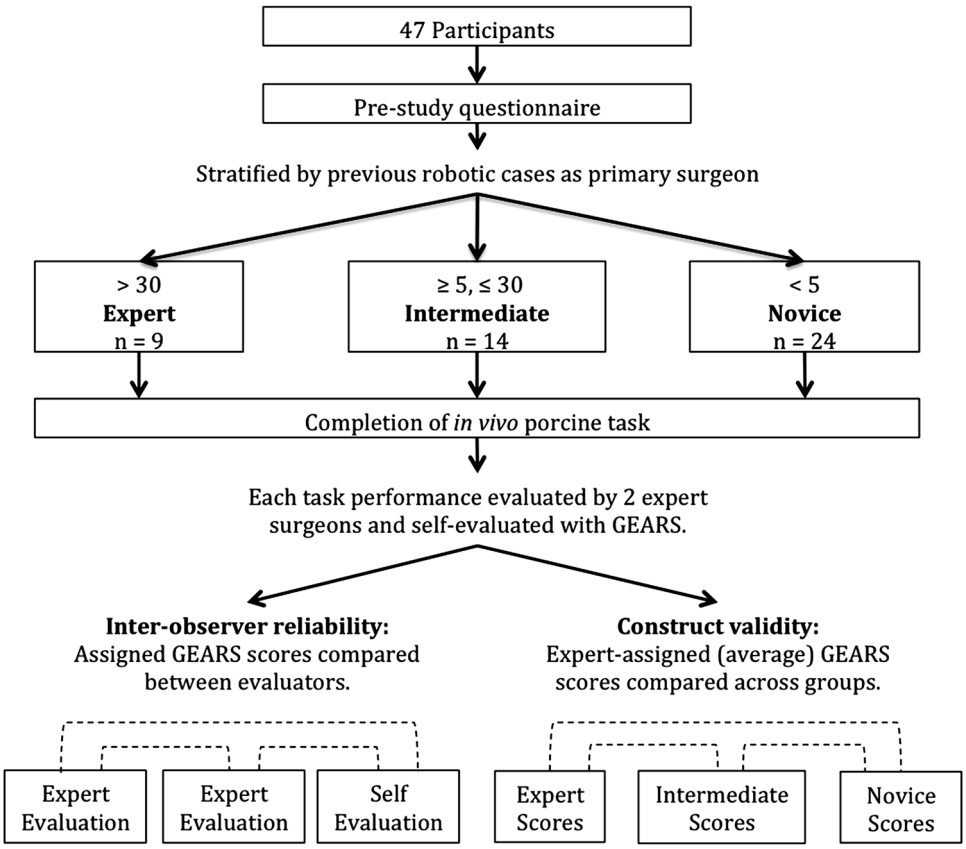

Fig. 1 Study design

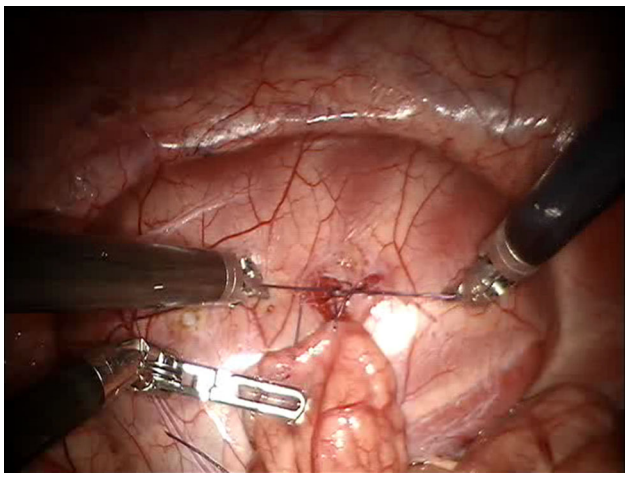

Fig. 2 In vivo robotic porcine task

primary surgeon. By classification, the novice group was comprised of 23 residents and one fellow, the intermediate group was comprised of only residents, and the expert group was comprised of eight attending physicians and one fellow.

Table 2 shows the median overall and domain-specific scores for the groups. Experts demonstrated superior performance compared to novices and intermediates for the overall GEARS score as well as for all individual domains ($p < 0.0001$). When comparing novice and intermediate groups, differences in performance scores were noticeably smaller. Intermediates did outperform novices in overall median GEARS score (20.75 vs. 19.0), and this difference approached statistical significance ($p = 0.0505$). Intermediates also demonstrated superior performance in the individual domains of efficiency ($p = 0.0280$) and autonomy ($p = 0.0425$). However, scoring differences (if present) in depth perception, bimanual dexterity, force sensitivity, and robotic control were not significant ($p > 0.05$ for all).

Thirty-four participants had complete scoring information from a self-assessment and two expert observers. These data were included in the interobserver reliability and internal consistency analyses. A strong correlation between experts' ratings confirmed excellent interobserver reliability ($r = 0.857$, 95 % CI [0.691, 0.941]). Less agreement was present between operator self-assessments compared to expert observations ($r = 0.435$, 95 % CI [0.121, 0.689]; $r = 0.422$, 95 % CI [0.081, 0.672]). Even less agreement was seen when analyzing only trainee self-assessments compared to expert observations ($r = 0.270$, 95 % CI [−0.164, 0.583]; $r = 0.282$, 95 % CI [−0.113, 0601]). In fact, trainees as a whole rated themselves on

Table 1 Participant demographics

	Novice	Intermediate	Expert
n	24	14	9
Age (years)*	30 (28–36)	30.5 (29–41)	42.5 (34–53)
# Female (%)	6 (25)	3 (21.4)	0 (0)
Classification (%)			
Residents	23 (95.8)	14 (100)	0 (0)
Fellows	1 (4.2)	0 (0)	1 (11.1)
In practice	0 (0)	0 (0)	8 (88.9)
MIS fellowship*	0 (0)	0 (0)	7 (77.8)
Robotic experience			
Years of experience*	0 (0–2)	1.5 (0.5–4.0)	5.5 (4–9)
Case number			
Surgeon*	0 (0–3)	10 (5–30)	350 (150–2,000)
Assistant*	17.5 (0–150)	20 (4–100)	100 (60–250)
Self-rated skill level¥	2 (1–4)	3 (2–4)	5 (3–5)
Previous robotic course*	0 (0)	3 (12.5)	3 (33.3)
Laparoscopic experience			
Years of experience*	2 (0–6)	2.5 (0–4)	10 (4–15)
Case number			
Surgeon*	2.5 (0–40)	20 (5–60)	600 (100–1,500)
Assistant*	20 (9–80)	30 (0–100)	200 (30–800)
Self-rated skill level¥	2 (1–4)	3 (2–3)	5 (3–5)
Previous laparoscopic course	4 (16.7)	2 (14.3)	4 (44.4)

Presented as median (range) or counts (%)

* $p < 0.05$

¥ 5-point Likert scale with one being novice, five being expert

Table 2 Pairwise comparison of GEARS performance scores across groups

GEARS	Novice (24)	Intermediate (14)	Expert (9)	*p* values		
				Novice versus int.	Int. versus expert	Novice versus expert
Overall Score	19.0 (11–27)	20.75 (16.5–27)	30 (29–30)	0.0505	<0.0001	<0.0001
Depth Perception	3.5 (2–4.5)	3.5 (3–4.5)	5 (5–5)	0.1115	<0.0001	<0.0001
Bimanual dexterity	2.5 (1–5)	3.25 (2–5)	5 (5–5)	0.0555	<0.0001	<0.0001
Efficiency	2.75 (1–4.5)	3 (2–5)	5 (4.5–5)	0.0280	<0.0001	<0.0001
Force sensitivity	3 (1–4.5)	3.25 (2.5–4.5)	5 (4.5–5)	0.1660	<0.0001	<0.0001
Autonomy	3.5 (2–5)	4 (3–5)	5 (5–5)	0.0425	<0.0001	<0.0001
Robotic control	3.5 (2–4.5)	3.5 (2.5–4.5)	5 (5–5)	0.0910	<0.0001	<0.0001

Data presented as median (range)

average 1.5 points better than the average expert score evaluation. Internal consistency for the assessment tool was excellent for expert observations, ($\alpha = 0.958$, 0.975), participant scoring ($\alpha = 0.924$), and overall ($\alpha = 0.959$). Omitting each item did not significantly increase internal consistency.

Discussion

With the rapid adoption of robotic surgery, the necessity of an objective assessment tool to gauge and evaluate robotic surgical skills has become more apparent. GEARS was specifically developed to fulfill this very need; however, beyond its initial validation in robotic prostatectomy, [8] data supporting its consistency, reliability, and adaptability are limited. Moreover, in recognizing the vital role that GEARS will likely play in the future of robotic surgical education, this relative lack of validation becomes critical. Thus, the primary goals of the current study were to externally evaluate the construct validity and interobserver reliability of GEARS in an in vivo animal model.

Indeed, excellent construct validity was demonstrated in this study. In this model, not only was GEARS able to

differentiate between experts and novices but also experts and intermediates. An ability to discern between intermediate and novice skill levels was also identified, although the difference overall only trended toward statistical significance. It is likely that an expanded cohort would allow for confirmation of this finding. Nevertheless, in the context of surgical education, such excellent construct validity will prove critical in tracking and evaluating trainee performance over time. Currently, the number of cases performed during training and the subjective assessment of expert surgeons are the predominant methods for determining surgical proficiency. Although surgical skill may increase with the number of cases, a direct correlation to surgical competency may not be present [15]. Our findings can help to establish minimum proficiency levels for different experience levels so that trainees have an objective measure of surgical competence. Even in the postgraduate re-training and maintenance of skills settings, acceptable proficiency levels can be developed to ensure that a minimal standard is met.

Additionally, with growing interest in skills acquisition outside of the operating room, our findings further support the use of GEARS in the evaluation of novel training models. Many robust surgical simulators have been developed for laparoscopic- and robot-assisted surgery, with proven face, content, and construct validity [4–8]. However, in the absence of a standardized method to assess surgical skills, none of these simulators have been able to accurately demonstrate concurrent or predictive validity. Without the correlation of simulator scores with clinical performance, the real educational impact of these training tools cannot be defined. In order to validate and compare performance between these new simulators, a standard assessment method needs to be widely adopted.

Adding to the strength of GEARS as an assessment tool, this study confirms excellent interobserver reliability between expert scorers. The reliability of any assessment tool hinges on its reproducibility regardless of the qualified evaluator. Interestingly, participants self-scoring did not correlate as well with experts, likely reflecting a participant self-scoring bias overall. This finding was further exaggerated with regards to trainee self-assessment, which may speak to relative trainee inexperience in evaluating technical skills. However, this trainee expert assessment discrepancy may be useful in highlighting what aspects of surgery the individual finds challenging. Further investigation with observers of different skill levels may help to clarify the influence of surgical skill on scoring results. There were limitations to this study that deserve discussion. While this was a prospective study, our cohort size was defined by personnel availability and external logistical factors and, as such, groups were not equal. Additionally, although the overall sample size was modest, it was larger than the initial validation study. We also examined the use of GEARS for a single task. Future investigation will be directed toward application of GEARS as a tool to track robotic skill acquisition over time. Finally, while we have reported an absolute difference between the scoring of the different skill levels, a larger dataset is required to develop benchmarks for expert performance. Nevertheless, this study confirms that GEARS is a robust skills assessment tool and suggests that it may have applicability for a range of surgical procedures and training tasks.

In conclusion, GEARS was able to differentiate between different robotic skill levels, while maintaining consistency and reliability for an in vivo robotic surgical task. The results of this study, combined with previous development and validation studies, support GEARS as a valid, reliable, and versatile skills assessment tool. As such, GEARS may be applied to provide formative feedback to trainees in a range of training environments and may be used to evaluate surgical simulators and measure their impact on clinical performance.

Acknowledgment Authors received Fund Source from Institutional, Ethicon, and Intuitive.

Conflict of interest Drs. Aghazadeh, Jayaratna, Hung, Desai, Gill, Goh, and Mr. Pan have no conflicts of interest or financial ties to disclose.

References

1. Lowrance WT, Eastham JA, Savage C, Maschino AC, Laudon VP, Dechet CB, Stephenson RA, Scardino PT, Sandhu JS (2012) Contemporary open and robotic radical prostatectomy practice patterns among urologists in the United States. J Urol 187(6): 2087–2093
2. Scott DJ (2006) Patient safety, competency, and the future of surgical simulation. Simul Healthc 1(3):164–170
3. Sweet RM, Beach R, Sainfort F, Gupta P, Reihsen T, Poniatowski LH, McDougall EM (2012) Introduction and validation of the American Urological Association basic laparoscopic urologic surgery skills curriculum. J Endourol 26(2):190–196. doi:10.1089/end.2011.0414
4. Seixas-Mikelus SA, Stegemann AP, Kesavadas T, Srimathveeravalli G, Sathyaseelan G, Chandrasekhar R, Wilding GE, Peabody JO, Guru KA (2011) Content validation of a novel robotic surgical simulator. BJU Int 107(7):1130–1135. doi:10.1111/j.1464-410X.2010.09694.x
5. van der Meijden OA, Broeders IA, Schijven MP (2010) The SEP "robot": a valid virtual reality robotic simulator for the Da Vinci surgical system? Surg Technol Int 19:51–58
6. Jonsson MN, Mahmood M, Askerud T, Hellborg H, Ramel S, Wiklund NP, Kjellman M, Ahlberg G (2011) ProMIS can serve as a da Vinci(R) simulator—a construct validity study. J Endourol 25(2):345–350. doi:10.1089/end.2010.0220
7. Kenney PA, Wszolek MF, Gould JJ, Libertino JA, Moinzadeh A (2009) Face, content, and construct validity of dV-trainer, a novel virtual reality simulator for robotic surgery. Urology 73(6): 1288–1292. doi:10.1016/j.urology.2008.12.044

8. Hung AJ, Patil MB, Zehnder P, Cai J, Ng CK, Aron M, Gill IS, Desai MM (2012) Concurrent and predictive validation of a novel robotic surgery simulator: a prospective, randomized study. J Urol 187(2):630–637. doi:10.1016/j.juro.2011.09.154
9. Ramos P, Montez J, Tripp A, Ng CK, Gill IS, Hung AJ (2014) Face, content, construct and concurrent validity of dry laboratory exercises for robotic training using a global assessment tool. BJU Int 113(5):836–842. doi:10.1111/bju.12559
10. Goh AC, Goldfarb DW, Sander JC, Miles BJ, Dunkin BJ (2012) Global Evaluative Assessment of Robotic Skills: validation of a clinical assessment tool to measure robotic surgical skills. J Urol 187(1):247–252. doi:10.1016/j.juro.2011.09.032
11. Vassiliou MC, Feldman LS, Andrew CG, Bergman S, Leffondre K, Stanbridge D, Fried GM (2005) A global assessment tool for evaluation of intraoperative laparoscopic skills. Am J Surg 190(1):107–113. doi:10.1016/j.amjsurg.2005.04.004
12. Goh A, Goldfarb DW, Sander J, Miles B, Dunkin B (2012) Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. J Urol 187:247–252
13. Hung AJ, Jayaratna IS, Teruya K, Desai MM, Gill IS, Goh AC (2013) Comparative assessment of three standardized robotic surgery training methods. BJU Int 112(6):864–871. doi:10.1111/bju.12045
14. Patel VR, Tully AS, Holmes R (2005) Robotic radical prostatectomy in the community setting—the learning curve and beyond: initial 200 cases. J Urol 174(1):269–272
15. Nordin P, van der Linden W (2008) Volume of procedures and risk of recurrence after repair of groin hernia: national register study. BMJ 336(7650):934–937. doi:10.1136/bmj.39525.514572.25