

# Global Rating Scales and Motion Analysis Are Valid Proficiency Metrics in Virtual and Benchtop Knee Arthroscopy Simulators

Justues Chang MD, Daniel C. Banaszek MD, Jason Gambrel MD,  
Davide Bardana MD

Published online: 18 August 2015

© The Association of Bone and Joint Surgeons® 2015

## Abstract

**Background** Work-hour restrictions and fatigue management strategies in surgical training programs continue to evolve in an effort to improve the learning environment and promote safer patient care. In response, training programs must reevaluate how various teaching modalities such as simulation can augment the development of surgical competence in trainees. For surgical simulators to be most useful, it is important to determine whether surgical proficiency can be reliably differentiated using them. To

our knowledge, performance on both virtual and benchtop arthroscopy simulators has not been concurrently assessed in the same subjects.

**Questions/purposes** (1) Do global rating scales and procedure time differentiate arthroscopic expertise in virtual and benchtop knee models? (2) Can commercially available built-in motion analysis metrics differentiate arthroscopic expertise? (3) How well are performance measures on virtual and benchtop simulators correlated? (4) Are these metrics sensitive enough to differentiate by year of training?

**Methods** A cross-sectional study of 19 subjects (four medical students, 12 residents, and three staff) were recruited and divided into 11 novice arthroscopists (student to Postgraduate Year [PGY] 3) and eight proficient arthroscopists (PGY 4 to staff) who completed a diagnostic arthroscopy and loose-body retrieval in both virtual and benchtop knee models. Global rating scales (GRS), procedure times, and motion analysis metrics were used to evaluate performance.

**Results** The proficient group scored higher on virtual ( $14 \pm 6$  [95% confidence interval {CI}, 10–18] versus  $36 \pm 5$  [95% CI, 32–40],  $p < 0.001$ ) and benchtop ( $16 \pm 8$  [95% CI, 11–21] versus  $36 \pm 5$  [95% CI, 31–40],  $p < 0.001$ )

---

Each author certifies that he or she, or a member of his or her immediate family, has no funding or commercial associations (eg, consultancies, stock ownership, equity interest, patent/licensing arrangements, etc) that might pose a conflict of interest in connection with the submitted article.

All ICMJE Conflict of Interest Forms for authors and *Clinical Orthopaedics and Related Research*® editors and board members are on file with the publication and can be viewed on request.

*Clinical Orthopaedics and Related Research*® neither advocates nor endorses the use of any treatment, drug, or device. Readers are encouraged to always seek additional information, including FDA-approval status, of any drug or device prior to clinical use.

Each author certifies that his or her institution approved the human protocol for this investigation, that all investigations were conducted in conformity with ethical principles of research, and that informed consent for participation in the study was obtained.

---

J. Chang, D. C. Banaszek  
School of Medicine, Queen's University, Kingston, ON, Canada

J. Chang  
Division of Orthopaedic Surgery, University of Toronto,  
Toronto, ON, Canada

D. C. Banaszek (✉)  
Division of Orthopaedic Surgery, Queen's University,  
905 Roshan Drive, Kingston, ON K7P 0B1, Canada  
e-mail: 12db24@queensu.ca

J. Gambrel  
Brockville General Hospital, Brockville, ON, Canada

D. Bardana  
Surgical Foundations Program, Division of Orthopaedic Surgery,  
Queen's University, Kingston, ON, Canada

D. Bardana  
Kingston General Hospital, Kingston, ON, Canada

GRS scales. The proficient subjects completed nearly all tasks faster than novice subjects, including the virtual scope ( $579 \pm 169$  [95% CI, 466–692] versus  $358 \pm 178$  [95% CI, 210–507] seconds,  $p = 0.02$ ) and benchtop knee scope + probe ( $480 \pm 160$  [95% CI, 373–588] versus  $277 \pm 64$  [95% CI, 224–330] seconds,  $p = 0.002$ ). The built-in motion analysis metrics also distinguished novices from proficient arthroscopists using the self-generated virtual loose body retrieval task scores ( $4 \pm 1$  [95% CI, 3–5] versus  $6 \pm 1$  [95% CI, 5–7],  $p = 0.001$ ). GRS scores between virtual and benchtop models were very strongly correlated ( $\rho = 0.93$ ,  $p < 0.001$ ). There was strong correlation between year of training and virtual GRS ( $\rho = 0.8$ ,  $p < 0.001$ ) and benchtop GRS ( $\rho = 0.87$ ,  $p < 0.001$ ) scores.

**Conclusions** To our knowledge, this is the first study to evaluate performance on both virtual and benchtop knee simulators. We have shown that subjective GRS scores and objective motion analysis metrics and procedure time are valid measures to distinguish arthroscopic skill on both virtual and benchtop modalities. Performance on both modalities is well correlated. We believe that training on artificial models allows acquisition of skills in a safe environment. Future work should compare different modalities in the efficiency of skill acquisition, retention, and transferability to the operating room.

## Introduction

Work-hour restrictions established by regulatory bodies such as the Accreditation Council for Graduate Medical Education [1] in the United States have made a large impact on surgical training programs. In conjunction with fatigue management strategies such as mandatory postcall days, a larger emphasis is placed on promoting a healthier learning environment where learners are less fatigued and can provide safer patient care. As a result, surgical training programs must take the opportunity to develop alternative training methods that allow learners to gain surgical competence more efficiently despite potential reductions in intraoperative case times [23].

In orthopaedic surgery, training on various virtual reality and benchtop simulators has been shown to improve performance based on motion analysis measures such as economy of motion and care with soft tissues in knee and shoulder arthroscopy [5, 6]. Additionally, multiple studies have shown that motion analysis and global rating scales were able to differentiate arthroscopic proficiency in knee and shoulder models and cadavers based on years of residency training [5, 11, 13, 15, 16, 22, 28, 29]. Finally, Martin et al. [20] have shown that performance on a shoulder simulator was similar to performance on a cadaver model. Howells et al. [14] have additionally shown that

intraoperative knee arthroscopy training with adjunctive benchtop training leads to improved objective structured assessments of technical skills and global rating scale (GRS) scores as well as improved motion analysis in the operating room.

For a surgical simulator to be most useful, it is important to determine how reliably and accurately surgical proficiency can be differentiated. Current literature has shown that both virtual and benchtop models, tested in isolation, can distinguish among novices, intermediate learners, and experienced arthroscopists [5, 11, 13, 15, 16, 22, 28, 29]. However, to our knowledge, both virtual and benchtop models have not been compared concurrently in the same subjects. Before establishing a simulator-based curricula, it is important to determine if one modality is more readily able to differentiate skill and promote development of proficiency. Because the cost of setting up virtual and benchtop simulations is quite variable, this knowledge will help training programs decide what modalities would be most appropriate for their learners. We therefore evaluated and sought to validate one benchtop knee model and one virtual reality arthroscopy simulator and specifically asked: (1) Do global rating scales and procedure time differentiate arthroscopic expertise in both virtual and benchtop knee models? (2) Can commercially available built-in motion analysis metrics differentiate arthroscopic expertise? (3) How well are performance measures on virtual and benchtop simulators correlated? (4) Are these metrics sensitive enough to differentiate by year of training?

## Materials and Methods

This cross-sectional study was completed at a single teaching hospital. Data were collected prospectively over a period of 1 year from 2012 to 2013.

Volunteers were recruited from the undergraduate medical students ( $n = 4$ ), orthopaedic surgery residents ( $n = 12$ ), and orthopaedic staff surgeons ( $n = 3$ ). Two staff surgeons are fellowship-trained sports surgeons and one is a nonsports surgeon. Participants were divided into two groups of 11 novice arthroscopists (senior medical students, Postgraduate Year [PGY] 1–3 residents) and eight proficient arthroscopists (PGY 4–5 and staff). Differentiation of proficiency at the PGY 4 level was decided on based on the current arthroscopic curriculum at our institution and the senior author's (DB) experience. In total, graduating orthopaedic residents at our institution have 8 months of sports rotations, on average 2 months per year from PGY 2, where they have exposure to knee arthroscopy. As experience accumulates, opportunity to be acting primary surgeon is also expected to increase. Participants were oriented to the two models, but none of the participants had

**Table 1.** Demographics and experience

Demographics and experience	Novice (n = 11)	Proficient (n = 8)	p value
Level of training	MS3-PGY 3	PGY 4-staff	–
Age (years)	26.0 ± 3.0	35.6 ± 5.2	0.001
Gender	Male = 9	Male = 8	–
Hand dominance	Right = 11	Right = 7	–
Previous observed*	2.1 ± 1.7	4.3 ± 1.3	0.010
First-assisted*	1.3 ± 1.6	4.3 ± 1.3	0.001
Primary surgeon*	0.4 ± 0.8	3.9 ± 1.6	0.000
Hours previous VR used*	0.6 ± 0.5	0.8 ± 0.5	0.477
Average hours game played*	2.1 ± 1.7	0.1 ± 0.4	0.003

\* 1 = 0–5 hours; 2 = 6–10 hours; 3 = 11–15 hours; 4 = 16–20 hours; 5 = > 20 hours; MS3 = medical student, third year; PGY = Postgraduate Year; VR = virtual reality.

pretraining on the models as part of this study. However, both models have been available at our institution for the purpose of self-practice outside of scheduled work time. Medical students are not expected to have any experience with arthroscopic surgery. At our institution, PGY 1 and 2 residents gain limited exposure to arthroscopic surgery and expand on these skills in PGY 3. It is not until PGY 4, however, that our senior author (DB) felt proficiency was beginning to be met.

Participant consent, baseline demographics, and arthroscopic experience data were collected before beginning the study (Table 1). Overall, within the last year, the proficient group had more arthroscopic experience as observers (2.1 ± 1.7 hours versus 4.3 ± 1.3 hours,  $p = 0.01$ ), assistants (1.3 ± 1.6 hours versus 4.3 ± 1.3 hours,  $p = 0.001$ ), and primary surgeons (0.4 ± 0.08 hour versus 3.9 ± 1.6 hours,  $p < 0.001$ ). However, there was no difference in prior experience with the virtual simulator with most subjects having less than 1 to 5 hours of previous use (0.6 ± 0.5 hour versus 0.8 ± 0.5 hour,  $p = 0.48$ ). The novice group had more video game exposure than did the experts (2.1 ± 1.7 hours versus 0.1 ± 0.4 hour,  $p = 0.003$ ). Note that exposure was measured from a scale of 1 to 5 (1 = 0–5 hours, 2 = 6–10 hours, 3 = 11–15 hours, 4 = 16–20 hours, 5 = > 20 hours).

### Equipment and Tasks

All participants completed a single diagnostic arthroscopy and loose-body retrieval in the benchtop knee (Sawbones®, Model 1517, Vashon Island, WA, USA) and insight ArthroVR™ Knee Simulator (GMV, Cary, NC, USA).

The Sawbones model included replicas of both menisci, medial cruciate ligament, lateral cruciate ligament, and

patella with quadriceps and patellar tendon. The original model was modified by adding an anterior cruciate ligament and posterior cruciate ligament using a Penrose drain. A decommissioned arthroscopic camera (Arthrex, Naples, FL, USA) was used with imaging viewed on a personal laptop computer. Anteromedial working and anterolateral viewing portals were premade for all participants. A standard probe and grasper were used for the probing examination and loose body removal, respectively. In the benchtop model, loose bodies were simulated by popcorn kernels placed in the medial and lateral compartments.

In both models, a checklist described by Elliott et al. [9] was used to guide participants through a standard diagnostic arthroscopy if needed. The checklist was not used directly for evaluation. The evaluator was allowed to provide assistance and guidance as they would in the operating room. Participants were allowed 10 minutes for each benchtop task. Participants were allowed unlimited time for virtual tasks because built-in motion analysis data were not collected without task completion.

### Outcome Measures

A subjective GRS score based on a validated tool described by Reznick et al. [24] was used. Similar GRS have been validated outside orthopaedics [3, 8, 20] and in trauma orthopaedics [18]. At the time of our data collection, arthroscopy-specific GRSs developed by Slade Shantz et al. (Objective Assessment of Arthroscopic Skills) [27], Koehler et al. (Arthroscopic Surgery Skill Evaluation Tool) [16], and Alvand et al. (Basic Arthroscopic Knee Skills) [5] were not yet published. Therefore, a modified GRS based on the Reznick model was used and included the following assessment categories: respect for tissue, time and motion, instrument handling, knowledge of instruments, flow of operation, knowledge of specific procedure, overall performance, and quality of final product. However, arthroscopy-focused modifications to the rubric scale were added below each of the Likert scales (scored 1–5).

Evaluations were completed by one of two expert observers (JG, MA) who were blinded to learners' level of training because they have not been involved with their residency training to date. The first, who completed nearly all assessments, was a community sports-focused orthopaedic surgeon without affiliation to our institution (JG). The second was a new internationally trained sports fellow (MA), who completed a small number of assessments when JG was not available. Separate GRS were completed immediately after direct observation of tasks in both virtual and benchtop modalities. The evaluators were asked to rate the participants with the assumption that a 5 was the level expected of an independently practicing surgeon. The

amount of assistance required to complete each task was accounted for in the GRS evaluation. The cumulative score (maximum 40) was obtained for each task. Arthroscopic video was not recorded, and all assessments were completed after live assessments. We felt live assessments were more realistic for day-to-day assessments, but this precluded repeat assessments for intra- and interobserver reliability testing.

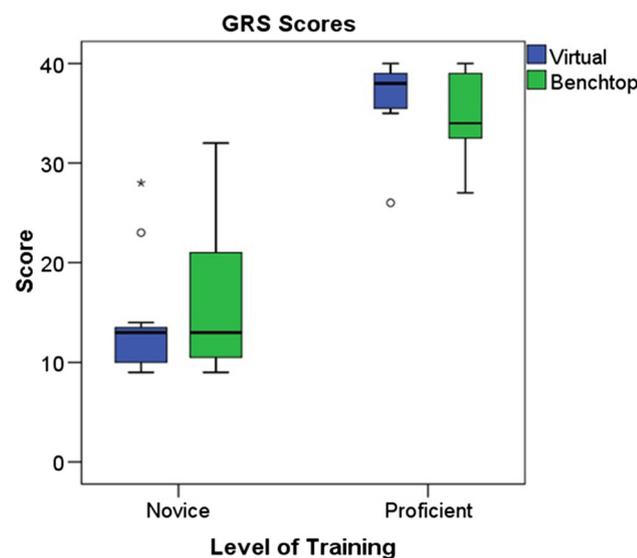
Objective scores included procedure time (seconds) in all models. Additionally, the virtual simulator provides preprogrammed motion analysis parameters including camera, probe, and grasper movement distance and roughness measured in millimeters and force in Newtons, respectively. Additionally, the software provides an overall performance score out of 10.

Statistical Analysis

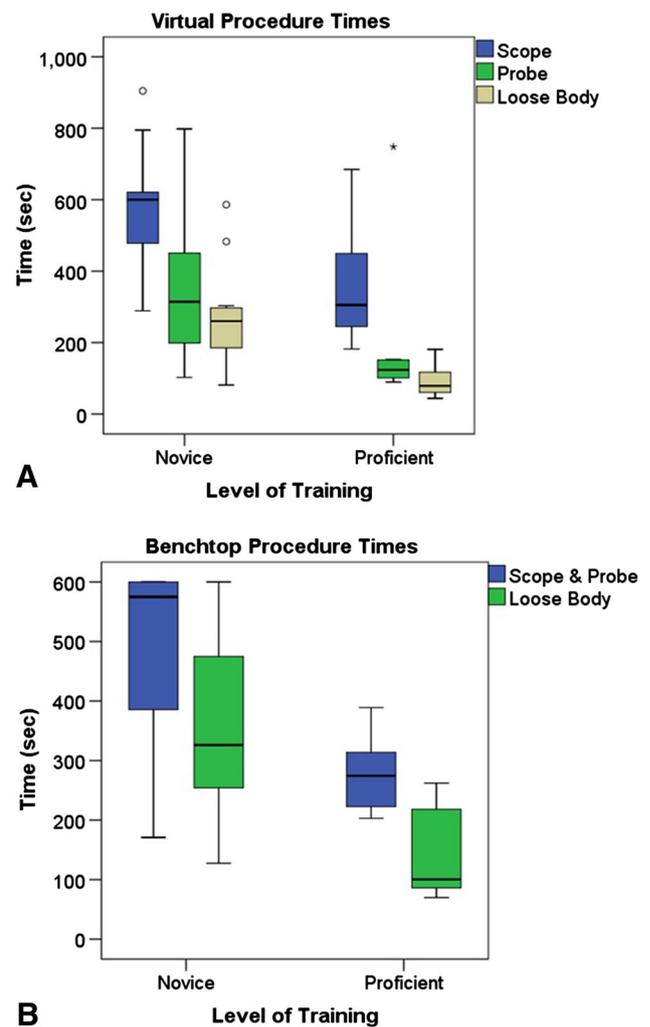
Nonparametric data such as the GRS scores were compared between groups using a Mann-Whitney U-test. Procedure time and motion analysis scores were compared using Student’s t-tests with an alpha of 0.05. All presented data in results, the table, and figures were analyzed based on initial group assignment. Statistical analysis was completed using SPSS software (Version 22; SPSS, Chicago, IL, USA). Spearman and Pearson correlation coefficients were used according to data type in comparison of level of training and GRS scores to procedure time and motion analysis results.

Results

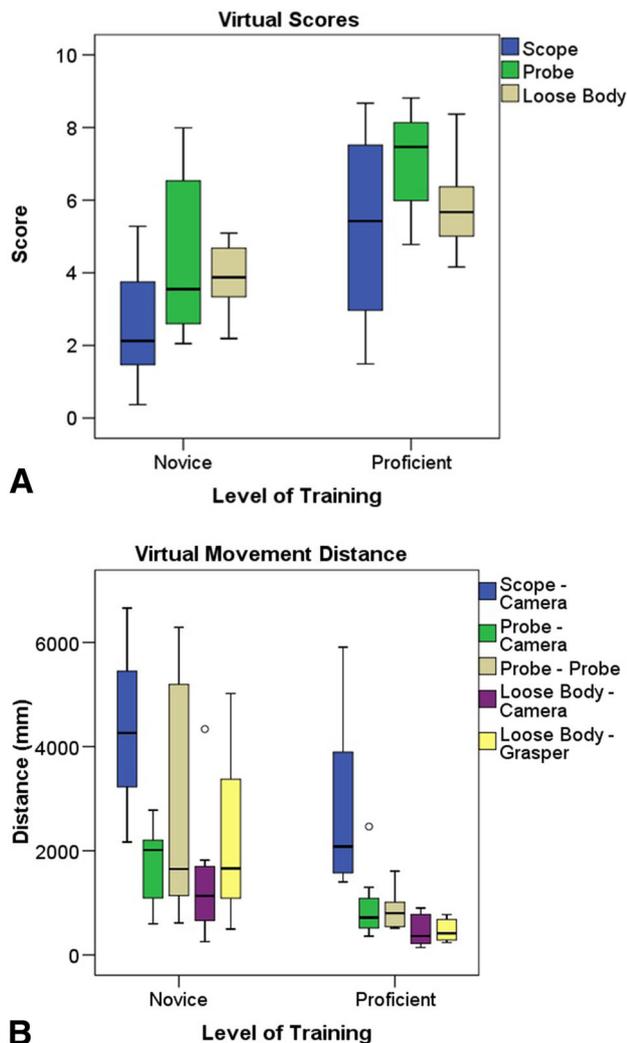
In both virtual and benchtop models, GRS scores and procedure times consistently differentiated novice and proficient subjects. By the subjective GRS, the proficient group performed better on virtual ( $14 \pm 6$  [95% confidence interval {CI}, 10–18] versus  $36 \pm 5$  [95% CI, 32–40],  $p < 0.001$ ) and benchtop ( $16 \pm 8$  [95% CI, 11–21] versus  $36 \pm 5$  [95% CI, 31–40],  $p < 0.001$ ) models (Fig. 1). Objectively, the proficient group had shorter procedure times in all except the probe-only task in the virtual simulator (Fig. 2A–B). On the virtual simulator, novices required more time to complete the blue sphere scope ( $579 \pm 169$  [95% CI, 466–692] versus  $358 \pm 178$  [95% CI, 210–507] seconds,  $p = 0.02$ ) and loose-body retrieval tasks ( $269 \pm 150$  [95% CI, 168–370] versus  $92 \pm 45$  [95% CI, 55–130]



**Fig. 1** This figure shows a summary boxplot of GRS scores in novice and proficient subjects in both virtual and benchtop models.



**Fig. 2A–B** (A) This summarizes virtual procedure times for the scope only, scope + probe, and loose-body tasks. (B) This summarizes benchtop procedure times for combined scope + probe and loose-body tasks.



**Fig. 3A–B** (A) This shows the virtual simulator built-in motion analysis scores for scope, scope + probe, and loose-body tasks. (B) This shows equipment movement distances for the camera (in all three tasks), probe, and grasper.

seconds,  $p = 0.005$ ). No difference was found on the probe-only task ( $349 \pm 203$  [95% CI, 213–485] versus  $199 \pm 223$  [95% CI, 12–385] seconds). Similarly, on the benchtop, novices were slower than the proficient group in combined scope + probe ( $480 \pm 160$  [95% CI, 373–588] versus  $277 \pm 64$  [95% CI, 224–330] seconds,  $p = 0.002$ ) and loose-body retrieval tasks ( $366 \pm 157$  [95% CI, 260–471] versus  $143 \pm 77$  [95% CI, 78–207] seconds,  $p = 0.002$ ).

Built-in motion analysis metrics from the virtual simulator were able to distinguish novice and proficient arthroscopists by overall scores and economy of motion, but not in care with soft tissues. The novice group had lower overall scores in the virtual scope ( $3 \pm 2$  [95% CI, 2–4] versus  $5 \pm 3$  [95% CI, 3–8],  $p = 0.018$ ), virtual probe ( $5 \pm 2$  [95% CI, 3–6] versus  $7 \pm 1$  [95% CI, 6–8],  $p = 0.009$ ), and virtual loose-body tasks ( $4 \pm 1$  [95% CI, 3–5] versus

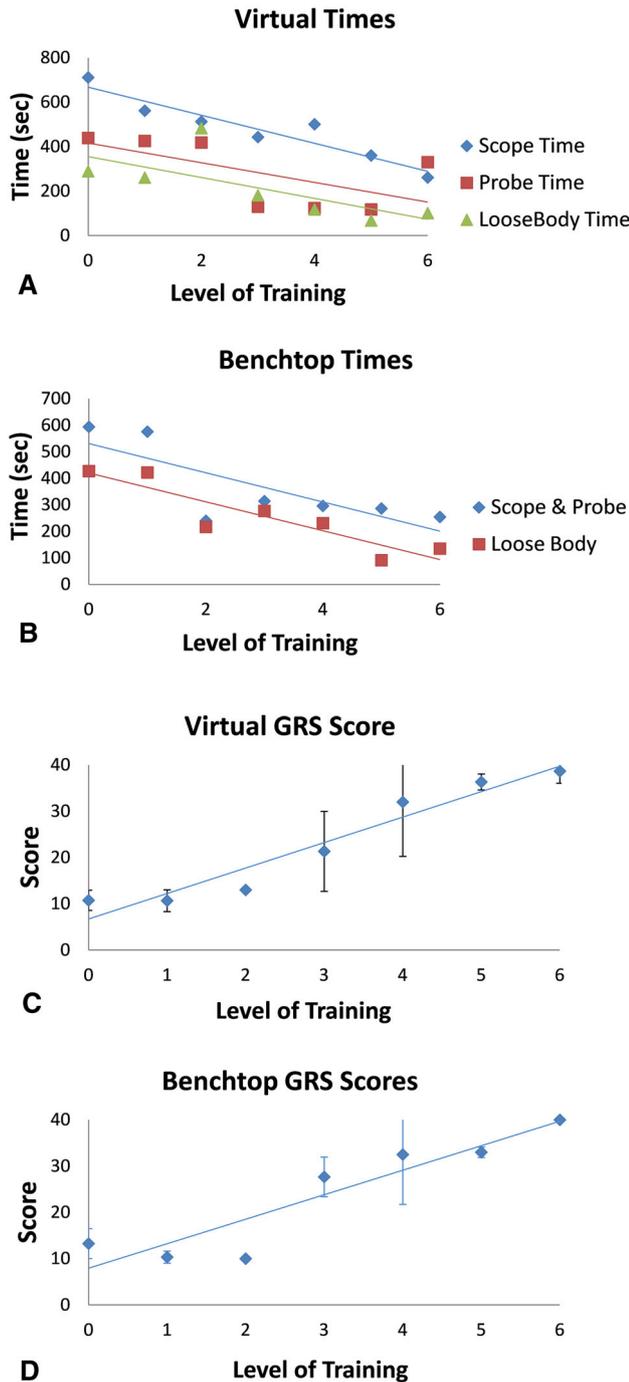
$6 \pm 1$  [95% CI, 5–7],  $p = 0.001$ ) (Fig. 3A–B). Novices were less efficient with equipment movement and had poorer economy of motion in all tasks. Novices moved the camera, probe, and grasper larger distances in the virtual scope (camera distance:  $4393 \pm 1491$  [95% CI, 3327–5459] versus  $2801 \pm 1662$  [95% CI, 1411–4191] mm,  $p = 0.048$ ), virtual probe (camera distance:  $1739 \pm 746$  [95% CI, 1238–2240] versus  $934 \pm 686$  [95% CI, 360–1508] mm,  $p = 0.028$ ; probe distance:  $2698 \pm 2285$  [95% CI, 1163–4233] versus  $854 \pm 370$  [95% CI, 545–1164] mm,  $p = 0.024$ ), and virtual loose-body tasks (camera distance:  $1400 \pm 1098$  [95% CI, 569–2225] versus  $470 \pm 301$  [95% CI, 218–721] mm,  $p = 0.033$ ; grasper distance:  $2062 \pm 1402$  [95% CI, 1178–3197] versus  $473 \pm 216$  [95% CI, 292–654],  $p = 0.004$ ). Interestingly, except for probe roughness ( $39 \pm 18$  [95% CI, 27–51] versus  $17 \pm 17$  [95% CI, 3–31] N,  $p = 0.018$ ), there was no difference with numbers available in camera and grasper roughness between the two groups (camera roughness: novice range [15–37  $\pm$  12–15] versus proficient range [8–27  $\pm$  9–21] N,  $p = 0.200$ – $0.807$ ; grasper roughness:  $47 \pm 26$  versus  $29 \pm 24$  N,  $p = 0.131$ ). Finally, with the numbers available, we observed no differences between accuracy in loose body retrieval between the novice and more-experienced arthroscopists ( $11 \pm 12$  [95% CI, 4–19] versus  $4 \pm 3$  [95% CI, 1–6] attempts,  $p = 0.09$ ).

Performance on virtual and benchtop models showed strong correlation based on subjective GRS scores and objective motion analysis metrics, suggesting that subjects performed similarly on both models. GRS scores between the virtual and benchtop models for the same subject, irrespective of surgical skill, were very strongly correlated ( $\rho = 0.93$ ,  $p < 0.001$ ). Higher subjective virtual GRS and benchtop GRS scores were associated with higher objective built-in motion analysis metrics (Virtual-Scope, probe, loose body:  $\rho = 0.70$ – $0.78$ ,  $p = 0.001$ ; Benchtop-Scope, Probe, Loose body:  $\rho = 0.59$ – $0.74$ ,  $p = 0.008$ ). Procedure times on both virtual and benchtop models were similar on scope and probe tasks ( $r = 0.52$ – $0.72$ ,  $p = 0.001$ – $0.21$ ), but the loose-body procedure times between the two models were weakly correlated ( $r = 0.31$ ,  $p = 0.21$ ). The built-in virtual procedure scores for scope, probe, and loose-body tasks were also compared with the task-specific measures such as equipment movement distance and roughness. A higher overall score was strongly correlated with shorter procedure times ( $r = -0.77$  to  $-0.88$ ,  $p < 0.001$ ) and shorter equipment movement distances ( $r = -0.57$  to  $-0.92$ ,  $p < 0.001$  to  $0.01$ ). To a lesser extent, lower roughness scores were also associated with higher overall scores ( $r = -0.46$  to  $-0.74$ ,  $p = 0.001$  to  $0.05$ ).

Performance scores on both virtual and benchtop models increased as trainee experience increased such that level of training based on PGY was strongly correlated with

subjective and objective performance measures. Higher level of training was strongly associated with higher GRS scores on the virtual ( $\rho = 0.8, p < 0.001$ ) and benchtop ( $\rho = 0.87, p < 0.001$ ) simulators. Higher PGY was also modestly associated with higher virtual overall task scores ( $\rho =$

$0.62-0.63, p = 0.004-0.005$ ) and shorter procedure times ( $\rho = -0.54$  to  $-0.75, p < 0.001$  to  $0.017$ ) (Fig. 4A–D). Finally, increased gaming exposure, seen in the novice group, was associated with lower GRS ( $\rho = -0.52$  to  $-0.63, p = 0.004$  to  $0.024$ ) and motion analysis scores ( $\rho = -0.44$  to  $-0.50, p = 0.029-0.067$ ).



**Fig. 4A–D** (A–B) These show average virtual and benchtop procedural times as stratified by year of training. (C–D) These show total GRS ratings based on level of training (medical student = 0, PGY 1–5 = 1–5, staff = 6).

**Discussion**

Simulated surgical training offers the advantage of providing a risk-free environment where learning from repetition, experimentation, and mistakes is encouraged. In recent years, there has been increased interest on the use of simulated models for evaluation of surgical proficiency. In a randomized controlled trial in general surgery, Franzeck et al. [10] found that simulator-based training may be more efficient than intraoperative training. After experiencing similar on-screen practice times, the laparoscopic simulator group performed similarly to the intraoperatively trained group. However, the intraoperatively trained group required much longer times in the operating room where they were not getting hands-on practice. For a simulator modality to be an effective training tool, trainees should demonstrate improvement in technical tasks practiced and trained tasks should lead to enhanced performance in the operating room. Furthermore, technical skills should be easily evaluated and expertise differentiated on the simulated models. Various simulation models have been explored in orthopaedic training [19, 21]; however, to our knowledge there have been no studies specifically comparing different surgical simulators in teaching arthroscopic skills or assessing surgical proficiency. The first step, and purpose of this study, is to validate the use of both virtual and benchtop models for differentiation of surgical proficiency in the same subjects. In our study, subjective GRS, objective procedure time, and motion analysis metrics were able to distinguish between novice and proficient arthroscopists. Subjects performed very similarly on both virtual and benchtop models and more arthroscopic experience was associated with better performance.

This study has a number of limitations. First, sample size was limited in this study by the availability of participants at our single center and thus further division of groups into year of training was not possible. Given the small sample size, outcome measures as correlated with level of training are of reduced statistical power. Particularly, given the number of assessed variables, the relatively small sample size poses the concern of pseudoreplication bias resulting in decreased p values. Although perhaps preliminary, our findings are in agreement with prior work with larger sample sizes that support that GRS and motion

analysis can differentiate arthroscopic skill by postgraduate year of training [9, 11, 15, 16]. Given the consistency of improved subjective and objective measures correlated with year of training within our study, it is predicted that a larger cohort would better elucidate the precision of the current measures. Second, the vast majority of evaluations were completed by a single observer (JG), but a small number was completed by another blinded observer (MA) as a result of scheduling conflicts. Because the GRS scores are subjective in nature, interobserver reliability is of importance because individual ratings may be biased as a result of different expectations from the independent observers; however, detailed descriptions and expectations under each GRS subcategory were provided to minimize this bias. Furthermore, strong correlation between subjective GRS and objective procedure time and motion analysis scores suggests that any bias was of limited effect.

Another potential limitation was that assessors were allowed to provide guidance and assistance as warranted and were requested to take into account the extent of assistance when completing the GRS. In prior studies, to more accurately assess a learner's arthroscopic proficiency, some authors have limited the observer-subject interaction by using video recordings of hand motion and on-screen displays [3, 16]. The latter option may minimize bias and more accurately capture one's true skill but is less practical in day-to-day assessments of learners and therefore we felt live evaluations are more appropriate for assessment of real-life validity. Unfortunately, this meant that only one set of evaluations was able to be completed for each subject as a result of time and resource constraints because no video recording was available for review. However, the purpose of this study is to determine whether performance on two different arthroscopic models is comparable rather than validating a subjective metric. Therefore, although important, intraobserver reliability is not a critical requirement for the purposes of this study. An additional limitation was that the division of novice and proficient subjects at the PGY 4 level was decided based on the arthroscopic curricula at our single center. Learners at other centers may have different exposure in their training and thus have different levels of expected proficiency. As defined in our study, not surprisingly, the proficient group had more arthroscopic experience compared with the novice group (Table 1). Notably, the groups had the same exposure to virtual knee simulators, which limits confounding tool-specific learning effects. Even with the impact of staff performance removed from data analysis, subgroup analysis showed that GRS scores, procedure times, and equipment movement distances were still able to differentiate novice and proficient arthroscopists on nearly all tasks, supporting the separation of groups at the PGY 4 level. Finally, each subject was tested only once on each

modality. Despite advances in simulator fidelity, it is clear that these models are not perfect imitations of the *in vivo* experience, because even expert surgeons show a clear learning curve during testing [2, 12]. Howells et al. [12] have shown that in fellowship-trained surgeons, performance on an arthroscopic shoulder simulator reaches a plateau or baseline after 12 trials [12]. Compared with novices, however, expert laparoscopic surgeons plateau much faster [2]. This suggests that even in the proficient surgeons, a number of repeated trials may be necessary to account for learning effects of a new instrument such as a simulated model. Thus, performance on simulators, particularly in nonexperts, may underestimate true intraoperative proficiency. Average performance after repeated trials would allow a more precise estimate of true skill.

Both GRS and procedure times were able to differentiate arthroscopic skill in our study. In agreement with prior work [15], a subjective GRS was easily able to distinguish novice and proficient arthroscopists in both virtual and benchtop models. Since initiation of our study, several others have developed and demonstrated the validity of arthroscopic-specific GRS with additional assessment categories for field of view, depth perception, bimanual dexterity, and autonomy [5, 16, 27]. Presumably these versions would allow even more precision in differentiating arthroscopic proficiency. It has been previously shown that a GRS was able to distinguish among PGY 1–2, 2–3, and 4–5 on orthopaedic trainees' performance on a cadaver knee for diagnostic arthroscopy and meniscectomy [15]. In our study, with respect to procedure time, it is notable that a 10-minute maximum time was allowed on the benchtop model tasks. Only seven of 11 novice subjects were able to complete the task in time. If allowed to proceed to completion, the differences in procedure time would be expected to be larger.

In our study, the built-in (nonmodified) motion analysis metrics including an overall performance score, procedure time, and equipment movement distance from the virtual simulator was able to consistently differentiate novice from proficient knee arthroscopists. Procedure time and economy of motion have been shown to differentiate novice, intermediate, and skilled surgeons in other commercially available shoulder simulators [11, 28]. Additionally, equipment movement and velocity captured by custom-made motion capture devices on benchtop knee models have also been able to distinguish arthroscopic skill [13, 29]. Interestingly, both novice and proficient subjects had similar roughness measures on the virtual simulator. In contrast, the "respect of tissue" subcategory of the GRS for both virtual and benchtop models indicated that the more experienced members were more careful. One explanation for this discrepancy is that subjectively the experienced

subjects performed more confidently and appeared safer with tissues, but objectively there was no difference. Alternatively, it may be that novice arthroscopists, although less skilled, accommodate for this by manipulating equipment more slowly to avoid inadvertent tissue damage. Subjective and objective measures of performance on both benchtop and virtual knee models have been shown to differentiate basic knee arthroscopic skill independently. To our knowledge, however, this is the first study to directly correlate performance on both modalities in the same subjects. We have shown that intrasubject performance was similar on both models. This has important implications because for the general orthopaedic training program, benchtop models are usually easier and cheaper to obtain than more expensive virtual simulators. Additionally, a GRS scale and procedure time are quite easily obtained measures, whereas motion analysis at the very least will require specialized equipment. An important question, however, is how accurate and precise these measures are. Although the tools are able to differentiate between novice and proficient surgeons, it is important to establish if the same measures can detect more subtle differences among each PGY or between fellowship-trained and general orthopaedic staff.

Previous authors have shown that more subtle differences in arthroscopic skill can be detected by GRS, procedure time, or economy of motion. The GRS and procedure checklist has been shown to differentiate PGY 1–2, 2–3, and 4–5 on basic knee arthroscopies [15]. Motion analysis metrics from a virtual shoulder simulator could differentiate skill among three groups separated into PGY 2/3, PGY 4/5, and staff [11]. Although our smaller sample size was not amenable to further division into smaller groups, our results were in general agreement with current literature. As PGY level increased, GRS scores also increased accordingly. To a lesser degree, higher PGY level was associated with better virtual procedure scores and shorter procedure time. Finally, although previous research has correlated video game exposure with laparoscopic surgical skill [25], this trend was not observed in our data. Increased gaming experience in the nonexperts was not enough to accommodate for limited arthroscopic exposure.

In conclusion, to our knowledge, this is the first study to evaluate performance on both benchtop and virtual knee arthroscopy simulators. Additionally, it is the first to validate the built-in motion analysis measures of the insight ArthroVR in orthopaedic trainees and staff surgeons in knee arthroscopy. Furthermore, we found that intrasubject performance on both virtual and benchtop knee models is highly correlated. GRS and procedure times on both virtual and benchtop knee models is consistently able to differentiate novice and proficient arthroscopists. Furthermore,

our results support current literature showing correlation between [3, 5] GRS and motion analysis metrics. To date, there are a plethora of arthroscopic training modalities and measures of performance. We believe that training on artificial models allows acquisition of skills in a safe environment. At our center, both models have been available for residents to train independently. However, as identified (Table 1), the use of these adjuncts is limited and a formal simulator curriculum has not yet been established at our center. Future work should compare different modalities in the efficiency of skill acquisition, retention, and transferability to the operating room. Ultimately, as previously attempted in general surgery [4, 7, 17, 26, 30], patient safety may be improved by determining predefined levels of competence on simulated modalities. Such a measure would allow safe and measurable transition of residents from practicing in the laboratory to performing on patients in the operating room.

**Acknowledgments** We thank Dr Majed Alrowaili for his volunteered time in completing subject evaluations, Carolyn Heymans for technical assistance related to the ArthroVR setup, and Heather Grant for guidance on statistical analysis and interpretation.

## References

1. ACGME. Accreditation Council for Graduate Medical Education: Common Program Requirements (resident duty hours). 2003. Available at: [https://www.acgme.org/acgmeweb/Portals/0/PFAssets/ProgramRequirements/CPRs\\_07012015.pdf](https://www.acgme.org/acgmeweb/Portals/0/PFAssets/ProgramRequirements/CPRs_07012015.pdf). Accessed June 2015.
2. Aggarwal R, Grantcharov TP, Eriksen JR, Blirup D, Kristiansen VB, Funch-Jensen P, Darzi A. An evidence-based virtual reality training program for novice laparoscopic surgeons. *Ann Surg*. 2006;244:310–314.
3. Aggarwal R, Ward J, Balasundaram I, Sains P, Athanasiou T, Darzi A. Proving the effectiveness of virtual reality simulation for training in laparoscopic surgery. *Ann Surg*. 2007;246:771–779.
4. Ahlberg G, Enochsson L, Gallagher AG, Hedman L, Hogman C, McClusky DA 3rd, Ramel S, Smith CD, Arvidsson D. Proficiency-based virtual reality training significantly reduces the error rate for residents during their first 10 laparoscopic cholecystectomies. *Am J Surg*. 2007;193:797–804.
5. Alvand A, Logishetty K, Middleton R, Khan T, Jackson WF, Price AJ, Rees JL. Validating a global rating scale to monitor individual resident learning curves during arthroscopic knee meniscal repair. *Arthroscopy*. 2013;29:906–912.
6. Andersen C, Winding TN, Vesterby MS. Development of simulated arthroscopic skills. *Acta Orthop*. 2011;82:90–95.
7. Brunner WC, Korndorffer JR Jr., Sierra R, Dunne JB, Yau CL, Corsetti RL, Slakey DP, Townsend MC, Scott DJ. Determining standards for laparoscopic proficiency using virtual reality. *Am Surg*. 2005;71:29–35.
8. Doyle JD, Webber EM, Sidhu RS. A universal global rating scale for the evaluation of technical skills in the operating room. *Am J Surg*. 2007;193:551–555; discussion 555.
9. Elliott MJ, Caprise PA, Henning AE, Kurtz CA, Sekiya JK. Diagnostic knee arthroscopy: a pilot study to evaluate surgical skills. *Arthroscopy*. 2012;28:218–224.

10. Franzeck FM, Rosenthal R, Muller MK, Nocito A, Wittich F, Maurus C, Dindo D, Clavien PA, Hahnloser D. Prospective randomized controlled trial of simulator-based versus traditional in-surgery laparoscopic camera navigation training. *Surg Endosc.* 2012;26:235–241.
11. Gomoll AH, O'Toole RV, Czarniecki J, Warner JJ. Surgical experience correlates with performance on a virtual reality simulator for shoulder arthroscopy. *Am J Sports Med.* 2007;35:883–888.
12. Howells NR, Auplish S, Hand GC, Gill HS, Carr AJ, Rees JL. Retention of arthroscopic shoulder skills learned with use of a simulator. Demonstration of a learning curve and loss of performance level after a time delay. *J Bone Joint Surg Am.* 2009;91:1207–1213.
13. Howells NR, Brinsden MD, Gill RS, Carr AJ, Rees JL. Motion analysis: a validated method for showing skill levels in arthroscopy. *Arthroscopy.* 2008;24:335–342.
14. Howells NR, Gill HS, Carr AJ, Price AJ, Rees JL. Transferring simulated arthroscopic skills to the operating theatre: a randomised blinded study. *J Bone Joint Surg Br.* 2008;90:494–499.
15. Insel A, Carofino B, Leger R, Arciero R, Mazzocca AD. The development of an objective model to assess arthroscopic performance. *J Bone Joint Surg Am.* 2009;91:2287–2295.
16. Koehler RJ, Amsdell S, Arendt EA, Bisson LJ, Braman JP, Butler A, Cosgarea AJ, Harner CD, Garrett WE, Olson T, Warne WJ, Nicandri GT. The Arthroscopic Surgical Skill Evaluation Tool (ASSET). *Am J Sports Med.* 2013;41:1229–1237.
17. Korndorffer JR, Jr., Scott DJ, Sierra R, Brunner WC, Dunne JB, Slakey DP, Townsend MC, Hewitt RL. Developing and testing competency levels for laparoscopic skills training. *Arch Surg.* 2005;140:80–84.
18. Leong JJ, Leff DR, Das A, Aggarwal R, Reilly P, Atkinson HD, Emery RJ, Darzi AW. Validation of orthopaedic bench models for trauma surgery. *J Bone Joint Surg Br.* 2008;90:958–965.
19. Mabrey JD, Reinig KD, Cannon WD. Virtual reality in orthopaedics: is it a reality? *Clin Orthop Relat Res.* 2010;468:2586–2591.
20. Martin KD, Belmont PJ, Schoenfeld AJ, Todd M, Cameron KL, Owens BD. Arthroscopic basic task performance in shoulder simulator model correlates with similar task performance in cadavers. *J Bone Joint Surg Am.* 2011;93:e1271–1275.
21. Modi CS, Morris G, Mukherjee R. Computer-simulation training for knee and shoulder arthroscopic surgery. *Arthroscopy.* 2010;26:832–840.
22. Pedowitz RA, Esch J, Snyder S. Evaluation of a virtual reality simulator for arthroscopy skills development. *Arthroscopy.* 2002;18:E29.
23. Pedowitz RA, Marsh JL. Motor skills training in orthopaedic surgery: a paradigm shift toward a simulation-based educational curriculum. *J Am Acad Orthop Surg.* 2012;20:407–409.
24. Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via an innovative 'bench station' examination. *Am J Surg.* 1997;173:226–230.
25. Rosser JC Jr., Gentile DA, Hanigan K, Danner OK. The effect of video game 'warm-up' on performance of laparoscopic surgery tasks. *JSLS.* 2012;16:3–9.
26. Seymour NE, Gallagher AG, Roman SA, O'Brien MK, Bansal VK, Andersen DK, Satava RM. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. *Ann Surg.* 2002;236:458–463; discussion 463–464.
27. Slade Shantz JA, Leiter JR, Collins JB, MacDonald PB. Validation of a global assessment of arthroscopic skills in a cadaveric knee model. *Arthroscopy.* 2013;29:106–112.
28. Srivastava S, Youngblood PL, Rawn C, Hariri S, Heinrichs WL, Ladd AL. Initial evaluation of a shoulder arthroscopy simulator: establishing construct validity. *J Shoulder Elbow Surg.* 2004;13:196–205.
29. Tashiro Y, Miura H, Nakanishi Y, Okazaki K, Iwamoto Y. Evaluation of skills in arthroscopic training based on trajectory and force data. *Clin Orthop Relat Res.* 2009;467:546–552.
30. Zendejas B, Cook DA, Bingener J, Huebner M, Dunn WF, Sarr MG, Farley DR. Simulation-based mastery learning improves patient outcomes in laparoscopic inguinal hernia repair: a randomized controlled trial. *Ann Surg.* 2011;254:502–509; discussion 509–511.