# Comparing Automatic Simulator Assessment with Expert Assessment of Virtual Surgical Procedures

S. Tuchschmid[1], M. Bajka[2], and M. Harders[1]

[1]Computer Vision Laboratory, ETH Zurich, Switzerland
[2]Clinic of Gynecology, University Hospital Zurich, Switzerland
[1]{tuchschmid, mharders}@vision.ee.ethz.ch,
[2]michael.bajka@hin.ch

**Abstract.** This study focuses on the comparison of expert assessment of virtual surgical procedures through Objective Structured Assessment of Technical Skills (OSATS) with the automatic assessment and feedback provided by a surgical simulator for hysteroscopic procedures. The existing multi-metric scoring system of the simulator was extended to include hysteroscopic myomectomy. The original OSATS was also modified for the examined surgical procedure. OSATS reliability, expert coherence, and interrater agreement with simulator feedback were investigated in a study with eight experts. The same selection of six movies showing virtual procedures performed at a hysteroscopy training course was rated by each expert. For the task-specific checklist, the reliability of the simulator was significantly higher than that of the individual human raters ($p=0.006$). In addition, the ranked order of the overall scores of all movies was the same for both simulator and expert consensus opinion. This is a first step to providing simulator feedback with the same reliability as an expert panel, thus facilitating competency-based surgical education and assessment in the near future.

## 1 Introduction

The traditional apprenticeship model for surgical education is increasingly put to test with regard to economy, reliability, and ethics [1]. The assessment of proficiency in this model is mainly based on the subjective impression of the surgical educator as well as on the number of performed procedures (e.g. logbooks). This paradigm is currently being replaced by more formal competency-based assessments, at least at the early stages of a surgeon's education [2]. In addition to this, the demand for certified continuous medical education (CME) is also growing, thus further increasing the need for new approaches to training and assessment.

Competency-based education requires objective measures of proficiency. To this end, Reznick et al. [3] have developed in the late 1990s the Objective Structured Assessment of Technical Skills (OSATS) method, which mainly targeted skills assessment on bench trainers. Since first presented, OSATS has been modified and successfully tested for reliability in many different surgical specialties,

such as general surgery [4], urology [5], or ophthalmology [6]. In obstetrics and gynecology, a team of the University of Washington has put its focus on the development of customized OSATS for a new hysteroscopy curriculum [7, 8].

Virtual Reality (VR) simulators have been reported to be a reliable and valid means of measuring surgical competency, in addition to traditional methods such as direct observation, animal models, or procedure logs [1]. VR trainers track all actions of a user in the virtual environment, with complete knowledge of the state of the procedure. Darzi et al. [9] pointed out in 1999 that *"a system that can provide unbiased and objective measurement of surgical precision (rather than just speed) could help training, complement knowledge based examinations, and provide a benchmark for certification."* Since then, some effort has been made to develop assessment parameters which relate better to the final objective of the given surgery. Rosen et al. used Markov models for the clustering of force/torque signatures in order to reveal the internal structure of a surgical task [10]. Moorthy et al. [11] analyzed the dexterity of a surgeon to provide performance feedback on the psychomotor skills. The CELTS system [12] developed at Harvard University in the early 2000s succeeded in providing objective expert-novice differentiation of procedural skills based on motion tracking of laparoscopic instruments. Ritter et al. [13] showed that time to completion is a poor metric for the objective assessment of intracorporal knot-tying performance, whereas an automated knot quality score could accurately distinguish well-tied knots from poorly tied ones.

However, the development of metrics that evaluate cognitive decision making and compare behaviour to that of experts remains an important research area [14]. Especially the correspondence of any simulator-based metrics with assessments of experts has not been well studied so far.

Therefore, we compared expert assessment through OSATS to the automatic assessment and feedback provided by a surgical simulator. The VR-based Surgical training system used in our study is a simulator for hysteroscopic interventions [15]. The HystSim trainer [16] allows full procedure training of diagnostic and therapeutic interventions and showed high acceptance ratings by experienced and novice surgeons [17].

## 2 Methods

### 2.1 Simulator Assessment of Virtual Hysteroscopic Procedures

A number of different assessment metrics for VR diagnostic hysteroscopy have been implemented in the simulator. We have extended a previously existing version of a multi-metric scoring system [18] with new parameters for therapeutic hysteroscopy, e.g. removal of myoma. Construct validity of the applied system has successfully been shown [19].

The simulator assessment parameters are grouped into five modules, namely Myomectomy, Visualization, Ergonomics, Fluid Handling, and Safety (see Table 2.1). Each parameter is weighted and linearly interpolated between a lower and upper limit. Two expert surgeons, each having performed more than 500

hysteroscopic interventions were responsible for choice, weighting, implementation, and configuration of the metrics into the scoring and grading system. In the resulting feedback report, module scores and the individual metrics for each module are provided to the trainee. In order to group the overall scores, gradings with letters from A (best, $> 90\%$) to E (worst, $< 60\%$) were given.

**Table 1.** Extended multi metric scoring table for therapeutic hysteroscopy on the HystSim VR trainer (adapted from [19]).

| **Scoring & Grading** | maximum score | upper value | lower value | target value |
|---|---|---|---|---|
| **Myomectomy** | **100** | | | |
| *removed pathology* | 100 | 95 % | 60 % | high |
| **Safety** | **100** | | | |
| *tool active without contact* | 25 | 10 s | 3 s | low |
| *cuts from front to back* | 25 | 1 | 0 | low |
| *cutting while view obscured* | 25 | 3 s | 0.1 s | low |
| *cutting while uterus collapsed* | 25 | 3 s | 0.1 s | low |
| **Economy** | **80** | | | |
| *procedure time* | 40 | 600 s | 240 s | low |
| *path length* | 30 | 3200 mm | 1600 mm | low |
| *camera tilted* | 10 | 30 s | 10 s | low |
| **Visualization** | **60** | | | |
| *visualized surface* | 40 | 85 % | 50 % | high |
| *left tube visualized* | 5 | 1.0 s | 0.0 s | high |
| *right tube visualized* | 5 | 1.0 s | 0.0 s | high |
| *upper cavum visualized* | 5 | 1.0 s | 0.0 s | high |
| *time out of focus* | 5 | 45 s | 5.0 s | low |
| **Fluid Handling** | **40** | | | |
| *time view obscured* | 20 | 60 s | 10 s | low |
| *time uterus collapsed* | 10 | 30 s | 5 s | low |
| *distension media used* | 10 | 3000 ml | 800 ml | low |
| **Overall score** | **380** | (100%) | | |
| **Grading** | A 90-100%, B 80-89%, C 70-79%, D 60-69%, E <60% | | | |

### 2.2 Adapted OSATS for Virtual Hysteroscopic Procedures

The OSATS introduced by Reznick et al. [3] consists of a task-specific checklist and a global rating scale (GRS). The checklist is comprised of binary assessment questions and needs to be adapted for each different test station. The GRS has seven items, each evaluated on a 5-point Likert scale. The middle and extreme points are anchored with explicit verbal descriptions. Finally, a Pass/Fail rating is provided for each test. The overall score for OSATS is computed by taking

the sum of the checklist score (1 point for each passed/completed item) and the global rating scale (7 items with maximal 5 and minimal 1 point).

The previously presented and validated OSATS for hysteroscopy [8] has been developed for the assessment on bench stations. Specifically, in that study residents were asked to assemble an operative hysteroscope and to resect a large polyp from an inanimate uterine model. Several of the suggested items do not apply for a fully virtual simulated procedure (e.g. "tool assembly", "placing of obturator", "removing of specimen"). Therefore, a new checklist was agreed upon by a panel of four experts with broad teaching experience in hysteroscopy. In addition, the panel extended the OSATS of the Toronto group with further items for evaluating video recordings [20]. To this end, GRS categories were removed, which cannot be rated based on video recordings only. These were substituted with points which were more appropriate for the given surgical procedure. We have exchanged the items "Knowledge of Instruments", "Use of Assistants", and "Knowledge of Specific Procedure" with the categories "Resection Skills", "Visualisation", and "Fluid Handling". The task specific checklist of the resulting OSATS is shown in Figure 1 and the GRS in Figure 2.

### 2.3 Study Setup

An important element of our study is the movies of the simulated procedures to be assessed. In order to avoid any bias which potentially could be introduced by artificial generation of such movies (e.g. by asking our expert panel to act like novices and perform common mistakes), we decided to obtain recordings during a hysteroscopy training course. The recordings were acquired during a course with 62 participants in Lugano, Switzerland from June 25 - 27, 2009, organized by gynecologie suisse (Swiss Society of Obstetrics and Gynecology). Virtual training sessions were carried out on the hysteroscopy simulator, focusing on myoma removal. The training setup used in the course is depicted in Figure 3. Thereafter, six representative movies were selected from all recordings for our study together with an expert. Selection criteria were set such as to cover a wide range of performances. The selected movies reflected varying prior exposure and experience with hysteroscopy. In addition, one movie was including during which a uterus perforation occurred – one of the main complications in hysteroscopy. In the scoring system of the simulator, these movies covered the whole scoring range from grade A through E.

Next, the movies were provided online in random order to national and international experts with teaching function for residents. The experts were provided with the OSATS form and asked to carefully review the movies consecutively on the webpage. Thereafter, they were asked to fill out first the task-specific checklist and then the GRS. Finally, the Pass/Fail rating had to be provided for each case. The first eight expert responses were used in this study. All participants completed the questionnaire with checklist, GRS, and Pass/Fail score for all movies.

| Task-Specific Checklist for Hysteroscopic Myomectomy | | |
|---|---|---|
| | | |
| **Instructions to candidates:** Cervix is already dilated. Perform first a diagnostic hysteroscopy and then safely remove myoma. | | |
| *Item* | *Not done or incorrect* | *Done correctly* |
| Performs complete diagnostics before cutting | 0 | 1 |
| Clearly visualizes left and right tubal ostia | 0 | 1 |
| Visualizes upper fundus by $180^0$ rotation | 0 | 1 |
| Only starts cutting when view is clear | 0 | 1 |
| Only starts cutting when loop close to pathology | 0 | 1 |
| Always pulls activated loop toward scope rather than away | 0 | 1 |
| Completely removes pathology | 0 | 1 |
| Does not cut into myometrium | 0 | 1 |
| Does not perforate | 0 | 1 |
| Always keeps horizon stable | 0 | 1 |
| Completes task within 5 min | 0 | 1 |

**Fig. 1.** Task-specific checklist, adapted for hysteroscopic removal of myoma.

| Global Rating Scale of Operative Performance (Reznick 1997), Adaptation for VR Hysteroscopy | | | | |
|---|---|---|---|---|
| | | | | |
| **1** | **2** | **3** | **4** | **5** |
| **Respect for Tissue:** *frequently used unnecessary force on tissue or risk of perforation by inappropriate use of instrument* | | *careful handling of tissue but minimal risk of perforation* | | *consistently handled tissues approprately with no risk of perforation* |
| **Time and Motion:** *many unecessary moves* | | *efficient time/motion but some unecessary moves* | | *clear economy of movement and maximum efficiency* |
| **Instrument Handling:** *repeatedly makes tentative or awkward moves with instruments by inappropriate use of instruments* | | *competent use of instruments but occasionally appeared stiff or awkward* | | *fluid moves with instruments and no awkwardness* |
| **Flow of Operation:** *frequently stopped operating and seemed unsure of next move* | | *demonstrated some forward planning with reasonable progression of procedure* | | *obviously planned course of operation with effortless flow from one move to the next* |
| **Resection Skills:** *pathology removal not appropriately caried out* | | *adequate resection performed, e.g. myoma mostly removed* | | *optimal resection, e.g. little tissue damage, complete removal* |
| **Visualisation:** *not the entire cavity is inspected, e.g. missing of crucial landmarks* | | *most of the cavity has been inspected under adequate viewing conditions* | | *cavity is fully visualized, all crucial landmarks inspected* |
| **Fluid Handling:** *innapproriate control and utilitization of distention fluid* | | *adequate utilization of distention fluid with limited loss of clear viewing conditions* | | *optimal establishing and sustaining of clear viewing conditions* |
| **1** | **2** | **3** | **4** | **5** |
| **Overall, should this candidate:** | | | **Pass** | **Fail** |

**Fig. 2.** Global Rating Scale, adapted for therapeutic hysteroscopy.

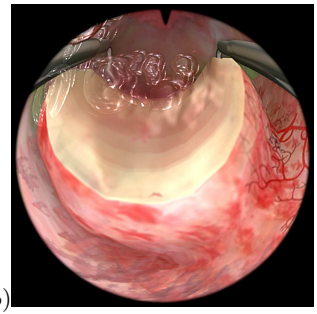
(a)  (b) 

**Fig. 3.** Hardware configuration during recording (a) and example simulation screenshot(b) of setup used in this study.

## 3 Results

In the first step we verified the reliability of the OSATS itself. We calculated Cronbach's $\alpha$ for the eight experts. According to [21], a test should have at least a Cronbach $\alpha$ of 0.7 to be reliable. We obtained 0.841 for the checklist, 0.894 for the global rating scale, and 0.858 for the Pass/Fail criteria. This is in line with other OSATS studies, where Cronbach $\alpha$ values in the range of 0.71 [3] to 0.97 [8] were reported.

Next, we examined the coherence of the experts' opinions. To this end, we first defined the consolidated expert panel opinion for the checklist. The majority decision of all eight experts on every item of the checklist for all six movies was determined. In a sense, this can be seen as an "expert panel" assessment. This expert panel decision was termed "consensus opinion". If the consensus opinion was tied on an item, it was treated as "passed". Note that treating such a tie as "failed" did not significantly change the results. In 25.7% of the items, the decision was unanimous. One or two experts disagreed in 22.8% and 22.7% of the items, respectively. In 18.2% the decision was with 5 to 3 votes, and in 10.6% experts were tied on an item. The obtained consensus report was then used as a measure of interrater variability for each expert.

Several methods for establishing interrater differences were presented in the literature [21]. We consider interrater agreement as the number of observations in agreement with the consensus opinion, divided by the total number of observations. Values above 80% are generally considered sufficiently high for reliable assessment [21]. Table 2 shows the results of comparing each of the individual experts to the consensus opinion for the eleven checklist items of each movie (N=11) and for all the six movies together (N=66). Depending on the movie, the mean interrater agreement for the checklist was in the range between 0.68 (Movie F) and 0.93 (Movie D). It can be seen that in the checklist ratings there are sometimes large discrepancies between experts on single movie assessments. Interrater agreement values as low as 0.45 resulted.

**Table 2.** Interrater agreement in checklist rating of 6 procedures (Movies A - F) between individual experts and consensus opinion as well as between simulator and consensus opinion. In addition, the mean of all experts' agreement with consensus opinion is shown.

| *Movie* | *A* [N=11] | *B* [N=11] | *C* [N=11] | *D* [N=11] | *E* [N=11] | *F* [N=11] | *All* [N=66] |
|---|---|---|---|---|---|---|---|
| *Expert 1* | 0.91 | 0.82 | 0.91 | 0.91 | 0.73 | 0.73 | 0.83 |
| *Expert 2* | 0.91 | 0.82 | 0.82 | 0.82 | 0.91 | 0.64 | 0.82 |
| *Expert 3* | 0.82 | 1.00 | 0.64 | 1.00 | 0.64 | 0.73 | 0.80 |
| *Expert 4* | 0.91 | 0.91 | 0.64 | 1.00 | 0.73 | 0.36 | 0.76 |
| *Expert 5* | 0.82 | 1.00 | 0.73 | 1.00 | 0.55 | 0.64 | 0.79 |
| *Expert 6* | 0.91 | 0.91 | 0.91 | 0.82 | 0.64 | 0.82 | 0.83 |
| *Expert 7* | 0.91 | 0.82 | 0.82 | 0.91 | 0.91 | 1.00 | 0.89 |
| *Expert 8* | 0.45 | 0.64 | 0.55 | 1.00 | 0.55 | 0.55 | 0.62 |
| *Mean Experts* | 0.83 | 0.86 | 0.75 | 0.93 | 0.70 | 0.68 | 0.79 |
| *Simulator* | 1.00 | 0.91 | 0.73 | 1.00 | 0.73 | 0.91 | 0.88 |

A similar trend is also visible when analyzing the results of the GRS for the different procedures. While there was wide agreement on some aspects of the procedures (e.g. Movie C and D provided unsatisfactory visualization), the 95% confidence interval of the median value is usually quite large (see Figure 4). As an example, the difference between upper and lower limit for the checklist item "Flow of Operation" was ranging from 0.747 points of the Likert scale (Movie A; 95% confidence interval 1.252 - 1.999) to 2.521 points (Movie F; 95% confidence interval 1.989 - 4.51).

We also compared agreement on the Pass/Fail criteria. For Movies A - F, the number of Pass/Fail ratings were 0/8, 7/1, 3/5, 7/1, 1/7, and 3/5, respectively. Movie F was a perforation of the uterine cavity. It is interesting to note that this serious complication was assessed quite differently by the experts in checklist, GRS, and Pass/Fail rating. One reason for this could be that the perforation was not fully realized by some of the experts. It could also be that some experts adjusted their scores for items unrelated to the actual perforation downwards, in order to ensure a low overall score.

In the next step, we compared the expert ratings to the automated assessment of the simulator. In order to compare consensus opinion with the simulator assessment, we defined corresponding checklist values based on the individual metrics of the simulator multi-metric scoring system. This was done by setting appropriate thresholds, and sometimes combining two simulator metrics for one checklist item. As an example, the checklist entry for "Clearly visualizes left and right tubal ostia" was true when both simulator metrics "Left tube visualized" and "Right tube visualized" were larger than one second. For all movies, the agreement between consensus opinion and simulator opinion on the checklist was 0.88 (see last row of Table 2). Comparing all human ratings with simulator assessment for all movies, the simulator agreement with the consensus opinion

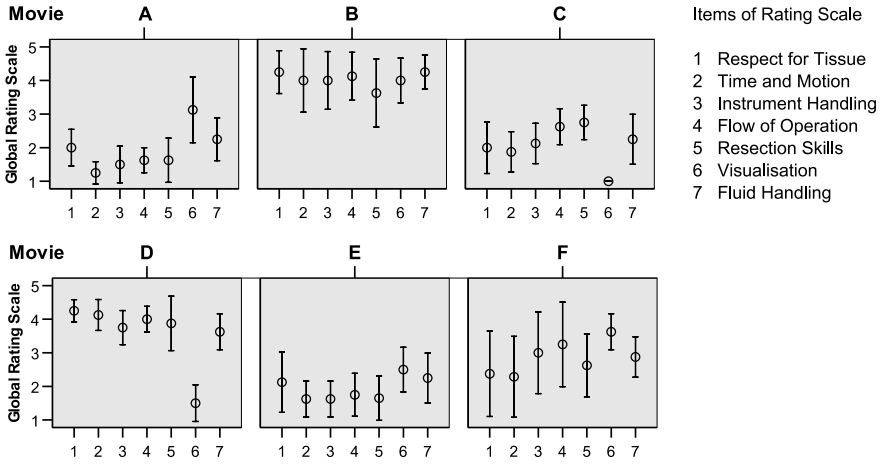



**Fig. 4.** Results of the global rating scale (GRS) for the 8 different expert raters. The mean scores for all items are shown with 95% confidence intervals represented by the bars. Scores are based on a 5-point Likert scale.

is significantly higher (p=0.006, Mann-Whitney U-test, two sided, exact). This trend is also visible for the individual experts, where simulator agreement with the consensus opinion is higher than that of the experts in 7 out of 8 cases. Due to the small number of movies for each expert this is, however, not statistically significant on the p¡0.05 level.

Finally, we also investigated how well the overall score of the simulator feedback correlated with that of the OSATS. The average ratings of the eight experts on the GRS were rounded to the nearest Likert value from 1 to 5. Direct comparison of the score numbers is not straightforward since the simulator does not provide categories matching the GRS. Moreover, the simulator feedback is set on a different scale. Therefore, we examined the ranked order of the movies according to the scores. We computed the Spearman-Rho correlation of the overall scores for Movies A - E. This was done between experts, as well as between simulator and experts. We excluded Movie F in this analysis since the simulator does not provide an overall score for a perforation, but labels it as an emergency scenario. Interrater correlation between experts was in the range from 0.4 to 1.0 (mean 0.77). In contrast to this, the agreement between consensus opinion and simulator reached 1.0. Thus, the simulator ranked the movies in exactly the same order as the expert panel opinion.

## 4 Discussion and Conclusions

We have introduced an adapted OSATS for VR hysteroscopic myomectomy and investigated expert opinion coherence in the assessment of six movies of simulated interventions. Comparing ratings of the checklist items showed that the simulator provided more consistent ratings than 7 out of the 8 experts. It is also interesting to note that a large part of the disagreement between simulator and expert consensus was due to a single item ("Only starts cutting when view is clear"). Thus, there is potential to further increase rating reliability by adjusting simulator metrics.

The challenge of comparing simulator with OSATS assessments is the very different nature of the two rating systems. While the OSATS has been specifically created for efficient and reliable scoring by a human rater, the multi-metric scoring system of the simulator was designed for optimal feedback for performance improvement [19]. Nevertheless, the ranked order of the overall scores was the same for both simulator and the consolidated expert opinion based on OSATS.

A major limiting factor of integrating OSATS into a training and assessment program is the limited availability of staff surgeons to observe the performance of trainees [1]. A simulator providing OSATS with the same reliability as an expert panel consensus for each individual procedure has the potential to make competency-based training possible, regardless of busy operating room schedules or the availability of cadavers or patients.

Our current plan is to extend the simulator feedback system to directly generate the full OSATS report as a complement to the multi-metrics feedback report. While items of the GRS such as "Respect for Tissue", "Instrument Handling", or "Flow of Operation" will be more difficult to implement with high specificity than checklist items, the high correlation between the ranked order of the overall scores indicates that such an endeavour could be successful. In addition, the availability of the full OSATS on the simulator would also facilitate meaningful transfer studies between VR and operation room.

In general, tests are considered reliable with Cronbach $\alpha > 0.7$, however important decisions about an individual's future should not be made unless $\alpha$ is $>0.9$ [21]. Current OSATS studies with human raters seldom reach this reliability level. Based on the initial results presented in this paper, we believe that the automatic generation of OSATS on a virtual reality simulator with adequate reliability is feasible. This would be especially important for high-stake assessment which may decide about a surgical career. While the past decades have seen a significant shift towards evidence-based clinical medicine, such development would support and accelerate a similar shift in education, training, and assessment of surgeons in the very near future.

*Acknowledgment:* This work has been performed within the NCCR Co-Me supported by the Swiss National Science Foundation.

## References


1. Moorthy, K., Munz, Y., Sarker, S., Darzi, A.: Objective assessment of technical skills in surgery. BMJ **327**(7422) (2003) 1032–1037
2. Gallagher, A., Ritter, E., Champion, H., G.Higgins, Fried, M., Moses, G., Smith, C.D., Satava, R.M.: Virtual reality simulation for the operating room: proficiency-based training as a paradigm shift in surgical skills training. Ann Surg **241**(2) (2005) 364–372
3. Reznick, R., Regehr, G., MacRae, H., Martin, J., McCulloch, W.: Testing technical skill via an innovative "bench station" examination. Am J Surg **173**(3) (1997) 226–230
4. Grantcharov, T.P., Kristiansen, V.B., Bendix, J., Bardram, L., Rosenberg, J., Funch-Jensen, P.: Randomized clinical trial of virtual reality simulation for laparoscopic skills training. Br J Surg **91**(2) (2004) 146–150
5. Matsumoto, E.D., Hamstra, S.J., Radomski, S.B., Cusimano, M.D.: The effect of bench model fidelity on endourological skills: a randomized controlled study. J Urol **167**(3) (2002) 1243–1247
6. Cremers, S.L., Lora, A.N., Ferrufino-Ponce, Z.K.: Global rating assessment of skills in intraocular surgery (GRASIS). Ophthalmology **112**(10) (2005) 1655–1660
7. Mandel, L.P., Lentz, G.M., Goff, B.A.: Teaching and evaluating surgical skills. Obstet Gynecol **95**(5) (2000) 783–785
8. VanBlaricom, A.L., Goff, B.A., Chinn, M., Icasiano, M.M., Nielsen, P., Mandel, L.: A new curriculum for hysteroscopy training as demonstrated by an objective structured assessment of technical skills (OSATS). Am J Obstet Gynecol **193**(5) (2005) 1856–1865
9. Darzi, A., Smith, S., Taffinder, N.: Assessing operative skill. Needs to become more objective. BMJ **318**(7188) (1999) 887–888
10. Rosen, J., Hannaford, B., Richards, C.G., Sinanan, M.N.: Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills. IEEE Trans Biomed Eng **48**(5) (2001) 579–591
11. Moorthy, K., Munz, Y., Dosis, A., Bello, F., Darzi, A.: Motion analysis in the training and assessment of minimally invasive surgery. Minim Invasive Ther Allied Technol **12**(3) (2003) 137–142
12. Stylopoulos, N., Cotin, S., Maithel, S.K., Ottensmeye, M., Jackson, P.G., Bardsley, R.S., Neumann, P.F., Rattner, D.W., Dawson, S.L.: Computer-enhanced laparoscopic training system (celts): bridging the gap. Surg Endosc **18**(5) (2004) 782–789
13. Ritter, E.M., McClusky, D.A., Gallagher, A.G., Smith, C.D.: Real-time objective assessment of knot quality with a portable tensiometer is superior to execution time for assessment of laparoscopic knot-tying performance. Surg Innov **12**(3) (2005) 233–237
14. Sewell, C.: Automatic Performance Evaluation in Surgical Simulation. PhD thesis, Stanford University (2007)
15. Harders, M., Bachofen, D., Bajka, M., Grassi, M., Heidelberger, B., Sierra, R., Spaelter, U., Steinemann, D., Teschner, M., Tuchschmid, S., Zatonyi, J., Szekely, G.: Virtual reality based simulation of hysteroscopic interventions. Presence: Teleoperators and Virtual Environments **17(5)** (2008) 441–462
16. Website: VirtaMed HystSim. http://www.simbionix.com/HystSim/HystSim.html (2009)
17. Bajka, M., Tuchschmid, S., Streich, M., Fink, D., Szekely, G., Harders, M.: Evaluation of a new virtual-reality training simulator for hysteroscopy. Surg Endosc **23(9)** (2009) 2026–33 (Epub Apr 24, 2008)

18. Tuchschmid, S., Bajka, M., Bachofen, D., Szekely, G., Harders, M.: Objective surgical performance assessment for virtual hysteroscopy. Stud Health Technol Inform **125** (2007) 473–478
19. Bajka, M., Tuchschmid, S., Fink, D., Szekely, G., Harders, M.: Establishing construct validity of a virtual reality training simulator for hysteroscopy via a multi metric scoring system. Surg Endosc (2009) (Epub ahead of print)
20. Dath, D., Regehr, G., Birch, D., Schlachta, C., Poulin, E., Mamazza, J., Reznick, R., MacRae, H.M.: Toward reliable operative assessment: the reliability and feasibility of videotaped assessment of laparoscopic technical skills. Surg Endosc **18**(12) (2004) 1800–1804
21. Gallagher, A.G., Ritter, E.M., Satava, R.M.: Fundamental principles of validation, and reliability: rigorous science for the assessment of surgical education and training. Surg Endosc **17**(10) (2003) 1525–1529