A Meta-Analysis of the Training Effectiveness of Virtual Reality Surgical Simulators

Syed Haque, *Member, IEEE,* and Shankar Srinivasan

***Abstract*—The increasing use of virtual reality (VR) simulators in surgical training makes it imperative that definitive studies be performed to assess their training effectiveness. Indeed, in this paper we report the meta-analysis of the efficacy of virtual reality simulators in: 1) the transference of skills from the simulator training environment to the operating room, and 2) their ability to discriminate between the experience levels of their users. The task completion time and the error score were the two study outcomes collated and analyzed in this meta-analysis. Sixteen studies were identified from a computer-based literature search (1996–2004). The meta-analysis of the random effects model (because of the heterogeneity of the data) revealed that training on virtual reality simulators did lessen the time taken to complete a given surgical task as well as clearly differentiate between the experienced and the novice trainees. Meta-analytic studies such as the one reported here would be very helpful in the planning and setting up of surgical training programs and for the establishment of reference 'learning curves' for a specific simulator and surgical task. If any such programs already exist, they can then indicate the improvements to be made in the simulator used, such as providing for more variety in their case scenarios based on the state and/or rate of learning of the trainee.**

***Index Terms*—Meta-analysis, surgical simulators, training performance, virtual reality (VR).**

I. Introduction

SURGICAL education, which traditionally depended on the apprentice-mentor relationship, is slowly changing with the introduction of virtual reality simulators for residency and student training programs [1]–[4]. Just as training with flight simulators has been shown to reduce costs and improve the expertise of pilots, so medical training on VR surgical simulators promises similar results. Using a combination of visual and haptic interfaces, the purported aim of VR surgical simulators is to help train surgical students and residents in complex surgical procedures even before they enter the operating room [5], [6]. Such procedures include neurosurgery, orthopedic surgery, plastic surgery, and even coronary artery bypass surgery. The representation of information in three dimensions as opposed to two provides medical students and practitioners with the tools for more accurate diagnosis and planning of surgical procedures. They enable the trainees to practice complex procedures over and over again, and allow training for uncommon emergency scenarios that they would not otherwise encounter for the first time without the experience necessary to guide them.

Manuscript received August 17, 2004, May 4, 2005; revised November 24, 2004, May 4, 2005.

The authors are with the Department of Health Informatics, University of Medicine and Dentistry of New Jersey, Newark, NJ 07107-3001 USA (e-mail: haque@umdnj.edu; srinivsh@UMDNJ.edu).

Digital Object Identifier 10.1109/TITB.2005.855529

Besides being an ethically suitable alternative to the use of animals and cadavers, the ability to measure technical competence is another attractive and unique feature of VR training [7]–[9]. In the traditional training technique, the assessment of skill acquisition and actual technical competence is subjective. The surgical educator directly observes the trainee and must then judge whether he/she is technically competent. There are obvious inefficiencies with this one-on-one evaluation system. Current VR simulators, on the other hand, allow users to compare their performance with that of their peers [10], [11]. They provide objective assessments such as the time taken to complete a task, the errors made in the process, and also the efficiency with which the movements were made in the accomplishment of the task. Despite such advantages, statistically appropriate experiments testing the efficacy of VR simulators are few and have very small sample sizes. The limiting factors are mainly the highly specialized subjects required to be used in such experiments and also the ethical factors involved in placing novice trainees in the operating room. Furthermore, the cost of removing specialized clinical personnel (such as surgeons) from high revenue producing work to undergo experimental tests and training should not be ignored. But it is still imperative that definitive experiments to improve our understanding of the effects of training on VR simulators is much needed, as it will allow them to be used more intelligently to improve provider performance, reduce errors, and, ultimately, promote patient safety. Although, as mentioned previously, few such training performance experiments exist, this can be obviated with the use of meta-analytic procedures. Meta-analyses purport to statistically combine and validate the results of individual studies with small sample sizes as is the case in surgical training. Indeed, in this paper we report the meta-analysis of the efficacy of VR simulators in surgical training. Specifically, the objectives of this study were primarily: 1) to determine the effectiveness of the VR training in the operating room (patient/animal surgery or examination) as opposed to training by traditional methods, and 2) to determine whether VR simulators can discriminate between the levels of experience among its users.

II. Methodology

A comprehensive literature search (OVID with MEDLINE data base) under the headings of virtual reality, simulators, and training was first done electronically and then expanded by manual search of current contents and bibliographies from recent journals and reports. Abstracts from surgical and endoscopy society meetings were also screened. The search was augmented by reports and papers located on the Internet.

Sixteen studies, published between 1996 and 2004, fulfilled the minimum criteria of prospective studies of the training performance of surgical residents (experienced and inexperienced) and students, with virtual reality simulators, reporting the results with statistical data. These were divided into two groups based on the two objectives of this study. Inclusion into either of these groups was solely based on whether proper statistically relevant data were available or not.

For the first study, which was to validate the effectiveness of transference of skills from the simulation environment to the operating room, the papers or reports chosen were solely those that had experienced surgeons or physicians operating on real patients or animal subjects. The control group was the one trained using traditional methods. This criterion of course limited the number of studies both because very few such experiments exist, and also it was not considered appropriate to include those which used artificial (such as foam models) or abstract setups (within the simulator environment) [12], [13]. For the second study, however, the criterion for inclusion was more relaxed, since we wanted to see how effective the VR simulator is in differentiating between the experienced (but not in VR simulators) and the novices based on their training performance. Here the term "novice" is used for those trainees who have done less than ten procedures of the tested task and the term "experienced" for those who have performed more than 50 such procedures. The ability of differentiation was considered to be better if there were a large difference between the training results of the two groups. This will enable the simultaneous testing of the surgical realism of the simulator and the skill measurement criteria. If the surgical realism is good and the skill measurement criteria are adequate, then it would be reasonable to see that the more experienced trainees (surgeons or final year residents) would perform much better than the less experienced (students, for example). If either the realism effects or the criteria were inappropriate or insufficient, then the results would not be clearly differentiable.

Besides the aforementioned, there were no other exclusions or exceptions to the inclusion criteria. On the whole, the studies chosen were sufficiently randomized and especially for the first objective, both the groups were reported to be equally qualified and experienced in their surgical skills. Tables I & II below list the papers chosen for both of the study objectives.

For the purpose of meta-analysis we chose the *task completion time* and the *error score* as the study outcomes to be collated and analyzed, since they were the only ones consistently reported across all the studies chosen in this report. Task completion time is the amount of time (measured in seconds or minutes) taken by the trainee to complete the task. Within reasonable limits (such as not causing traumatizing injuries), a lesser time taken for completion was deemed suitable. The outcome error score was either reported as the number of "wall strikes" or based on rating by external "expert" observers and could then be a subjective evaluation of inefficiency of movements, incurrence of injuries, and requests for assistance. Accordingly, an error score, which was less than that scored by the control group, was deemed desirable. Where the external observer ratings were for the efficacy of the performance, the error score was imputed as the difference between the maximum score possible and the given efficacy score. In all of the papers chosen, the data on both of the outcomes was either given in terms of raw means and standard deviations, or in terms of t-test probabilities. For both the cases, the effect size d [30] was calculated using the expressions given as follow:

$$\text{EffectSize} \quad d = \frac{\mu_{\text{ex}} - \mu_{\text{ctrl}}}{\text{SD}_{\text{pooled}}} \tag{1}$$

TABLE I
LIST OF THE STUDIES, THE VR SIMULATORS USED AND THE TASKS EVALUATED

STUDY I			
Authors	**Publication Year**	**VR Simulator used**	**Task**
Ost *et al.* [14]	2001	AccuTouch Flexible Bronchoscopy Simulator	Bronchoscopy (patients)
Gerson & Van Dam [15]	2003	Sigmoidoscopy Simulator	Sigmoidoscopy (patients)
Rowe & Cohen [16]	2002	AccuTouch Flexible Bronchoscopy Simulator	Fiberoptic Intubation (patients)
Ahlberg *et al.* [17]	2002	MIST-VR	Laparoscopic Appendectomy (pig)
Hyltander *et al.* [18]	2002	LapSim	Laparoscopic Cholecystectomy (pigs)
Seymour *et al.* [19]	2002	MIST-VR	Laparoscopic Cholecystectomy (patients)
Grantcharov *et al.*[20]	2004	MIST-VR	Laparoscopic Cholecystectomy (patients)

TABLE II
LISTS OF THE STUDIES, THE VR SIMULATORS USED AND THE TASKS EVALUATED

STUDY II			
Authors	**Publication Year**	**VR Simulator**	**Task**
Datta *et al.* [21]	2001	AccuTouch Endoscopy Simulator	GI Endoscopy
Jacomides *et al.*[22]	2004	Uromentor	Ureteroscopy
Prystowsky *et al.*[23]	1999	Prototype by MusculoGraphics, Inc.	IV Catheter Placement
Gallagher *et al.*[24]	2001	MIST-VR	Laparoscopic Cholecystectomy
Gallagher *et al.*[25]	2004	MIST-VR	Laparoscopic Cholecystectomy
Ritter *et al.*[26]	2003	GI Mentor II	GI Endoscopy
O'Toole *et al.*[27]	1999	Simulator by Boston Dynamics, Inc.	Anastomosis
Colt *et al.* [28]	2001	AccuTouch Endoscopy Simulator	Flexible Fiberoptic Bronchoscopy
Grantcharov *et al.*[29]	2003	MIST-VR	Laparoscopic Cholecystectomy

where μ_{ex} is the group mean of the experimental group; i.e., VR simulator trained for objective 1, or more experienced simulator trainees for objective 2, and μ_{ctrl} is the group mean for the control group; i.e., traditionally trained surgeons for objective 1, or novice trainees for objective 2. $\mathrm{SD}_{\mathrm{pooled}}$ is the pooled sample standard deviation, which is calculated as

$$\mathrm{SD}_{\mathrm{pooled}} = \sqrt{\frac{(n_{\mathrm{ex}}-1)\mathrm{SD}_{\mathrm{ex}}^2 + (n_{\mathrm{ctrl}}-1)\mathrm{SD}_{\mathrm{ctrl}}^2}{n_{\mathrm{ex}}+n_{\mathrm{ctrl}}}} \quad (2)$$

where n_{ctrl} and n_{ex} are the sizes of the control and experimental group and $\mathrm{SD}_{\mathrm{ctrl}}$ and $\mathrm{SD}_{\mathrm{ex}}$ are their corresponding standard deviations.

In studies where the probabilities were given, the effect size was calculated as

$$d = t\sqrt{\left(\frac{(n_{\mathrm{ex}}+n_{\mathrm{ctrl}})}{n_{\mathrm{ex}}n_{\mathrm{ctrl}}}\right)\left(\frac{(n_{\mathrm{ex}}+n_{\mathrm{ctrl}})}{n_{\mathrm{ex}}+n_{\mathrm{ctrl}}-2}\right)} \quad (3)$$

where t is the value of the students'-test statistic for a two-sided test, obtained from the P values (either the exact value or the significance level). The sign of the effect size was then assigned based on whether the difference was in the expected direction or otherwise. Negative effect sizes indicate that the mean values for experimental group are smaller than those for control group. Thus, for the first study objective, a negative effect size indicated that the VR trained group took less time in completing the task, or had a lesser error score than the traditionally trained control group.

Since in our collection of studies the sample sizes are very unequal, accordingly, we have to take into account that larger sample size studies will have less sampling error than smaller size studies, and accordingly their respective effect sizes should be weighted appropriately to yield a sample weighted mean value of d, and also its variance, σ_d^2. This is done as follows:

$$\bar{d} = \frac{\sum(N_i * d_i)}{\sum N_i}$$

and

$$\sigma_d^2 = \frac{\sum(N_i * (d_i - \bar{d})^2}{\sum N_i} \quad (4)$$

where N_i is the sample size for each study, and d_i is its effect size.

After calculation of effect sizes for each study, the next decision is the choice of models to estimate overall effect size and its variability. One model assumes that all effect sizes found in the trials arise from a single distribution (homogeneous effect sizes), and is known as a *fixed effect* model. Another model, known as a *random effects* model, assumes that the values potentially arise from distributions with different mean values, with significant variation among these mean values (heterogeneous). It might very well indicate the presence or absence of moderator or intervening variables affecting the variance observed. To select the appropriate model, we used a chi-square test (the Q statistic) to determine if the effect sizes are heterogeneous or homogeneous. The Q statistic is computed as follows:

$$Q_{(k-1)} = k * \left[\frac{\sigma_d^2}{\sigma_e^2}\right]. \quad (5)$$

Where k is the number of studies whose effect size, d has been calculated, σ_d^2 is the variance of the sample weighted d, and σ_e^2 is the sampling error variance calculated as

$$\sigma_e^2 = \left[\frac{\bar{N}-1}{\bar{N}-3}\right] * \left[\frac{4}{\bar{N}}\right] * \left[\frac{\bar{d}}{8}\right]. \quad (6)$$

The sampling error variance, once computed, allows us to determine how much of the variance in the effect sizes we calculated was attributable to the sampling error alone, and thus yielding a more accurate estimate of the true variance across the effect sizes. In this work, all effect sizes were found to be heterogeneous. Despite close scrutiny for subcategories that would account for the heterogeneity, none was found in any of the outcome measures. The effect sizes were then computed by the random effects model of DerSimonian and Laird [31] to take into account heterogeneity. This will be further discussed in Section III.

III. Results

The effect size calculations and the test results on task completion time and performance or error score for both of the study objectives are illustrated in Figs. 1–4. The corresponding forest plots for each of the results provide an easy way to visualize the range of the effect size values and compare their variations. The variables N1 and N2 are the sample sizes of the experimental and the control groups, respectively. Also, note that a negative value of the effect size termed standardized mean difference (SMD) indicates that the experimental group got a lesser value than the control group. This translates to either lesser task completion time or a lesser error score. Along with the effect size, values their ranges are given in terms of the lower and upper 95% confidence interval values. The result of the chi-square test for heterogeneity is given by the value of Q, and the appropriate degrees of freedom (DF) as also the significance level.

A. The Task Completion Time

Calculated from the effect sizes, the VR trained group took much less time than the traditionally trained group. Using the fixed effect model, a significant effect size was obtained, but the inadequacy of analysis was indicated by lack of homogeneity ($P < 0.0001$). Since we could find no covariates to explain heterogeneity (besides the different simulators employed and the tasks themselves being different), reanalysis using the random effects model was undertaken, which yielded an effect size of $-2.175(95\%\mathrm{CI} - 3.865, -0.485)$, a statistically significant finding.

B. Error Score

The effect sizes are highly heterogeneous ($P < 0.0001$), and hence a reanalysis with a random effects model was done, yielding an effect size of $-1.565(95\%\mathrm{CI} - 3.445, 0.314)$. As

Study – Effectiveness of transference of skills from virtual reality environment to the operating room

Study	N1	N2	Total	SMD	95% CI	
Ost et al.	3	3	6	-5.415	-13.161	2.331
Gerson et al.	9	7	16	0.000	-1.081	1.081
Hyltander et al.	12	12	24	-5.299	-7.230	-3.368
Rowe&Cohen	12	8	20	-4.011	-5.780	-2.241
Seymour et al.	8	8	16	-0.821	-1.952	0.310
Grantcharov et al.	8	8	16	-0.321	-1.403	0.760
Total (fixed effects)	52	46	98	-1.178	-1.706	-0.651
Total (random effects)	52	46	98	-2.175	-3.865	-0.485

Test for heterogeneity
Q = 42.0283, DF = 5, P < 0.0001

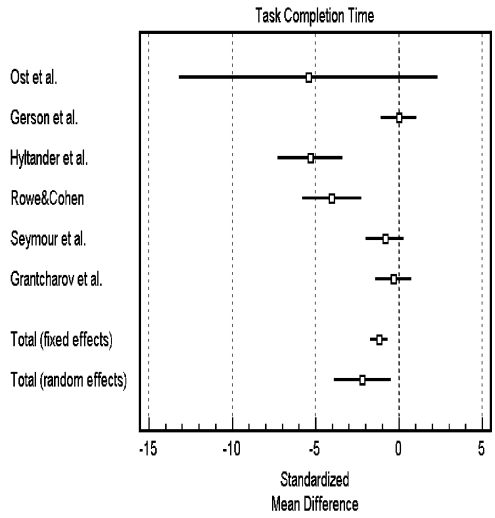

Fig. 1. Effect size values and their confidence interval estimates for the task completion time (study I).

Study	N1	N2	Total	SMD	95% CI	
Ost et al.	3	3	6	-1.316	-4.210	1.579
Gerson et al.	9	7	16	4.255	2.105	6.404
Hyltander et al.	12	12	24	-6.645	-8.980	-4.311
Rowe&Cohen	12	8	20	-2.377	-3.680	-1.075
Seymour et al.	8	8	16	-4.021	-6.079	-1.964
Grantcharov et al.	8	8	16	-1.266	-2.472	-0.059
Ahlberg et al.	14	15	29	0.057	-0.705	0.820
Total (fixed effects)	66	61	127	-0.974	-1.457	-0.491
Total (random effects)	66	61	127	-1.565	-3.445	0.314

Test for heterogeneity
Q = 75.8808, DF = 6, P < 0.0001

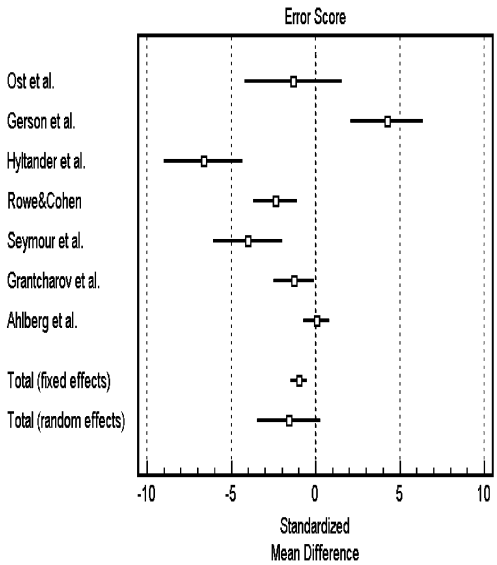

Fig. 2. Effect size values and their confidence interval estimates for the error score (study i).

the confidence interval spans the value of zero, we interpret the data to show that the errors committed by VR trained users are comparable to those of the traditionally trained group. It also indicates the ability of the VR simulators to reveal ahead of time and in a nonthreatening (to the patient's health) environment the required improvements to be made before transference to the operating room. A similar observation has been made in a recent study of the performance of residents with differing years of training on a cadaver [32].

A possible explanation for the heterogeneity of the effect sizes could be the result of the differences of the simulators employed, and also the differing nature of the tasks to be accomplished. Since the total number of available studies is very small at this point in time, dividing them into various subgroups based on the specific simulator used would have resulted in very small sized (one or two studies) subgroups. Hence, it was considered worthwhile to wait until more studies were available to make a more detailed study of the simulator-based subgroups. However, these results still indicate the ability of training under VR simulators to provide for surgical skills comparable (and even perhaps better) to those of the well-validated traditional techniques [33], [34].

C. Task Completion Time

Using the fixed effects model, a significant effect size $-1.059(95\%\text{CI} - 1.331, -0.786)$, was obtained and the adequacy of analysis was indicated by the homogeneity $(P = 0.1305)$. To confirm this finding, reanalysis using the random effects model was undertaken which yielded an effect size very close to that of the fixed effects model. Thus, no new information resulted. Both of the effect size values reveal that there is a difference between the time taken to complete a task between the experienced users and the novices. The chi-square test not being significant indicates that the findings come from the same population, and that there are no moderator or intervening variables operating on the outcome. The results also reveal that the simulator's realism effects are very sufficient and the measurement criteria appropriate, as otherwise the groups would not be so clearly differentiable.

D. Error Score

The effect sizes are highly heterogeneous $(P < 0.0001)$. Since no obvious subcategories could be found in the data, reanalysis with a random effects model yielded an effect size of $-1.325(95\%\text{CI} - 2.125, -0.525)$. As the confidence interval is significantly different from zero, we interpret the data to

Study – Effectiveness of transference of skills from virtual reality environment to the operating room

Study	N1	N2	Total	SMD	95% CI	
Datta et al	15	15	30	-0.714	-1.489	0.061
Jacomides et al	16	16	32	-1.182	-1.973	-0.391
Prystowsky et al	9	37	46	-1.975	-2.840	-1.109
Gallag1 et al	12	100	112	-0.331	-0.938	0.276
Gallag2 et al	12	12	24	-1.522	-2.505	-0.540
Johnston et al	15	10	25	-0.885	-1.776	0.005
Ritter et al	5	6	11	-1.066	-2.579	0.447
OToole et al	8	12	20	-1.524	-2.637	-0.411
Colt et al.	4	5	9	-1.788	-3.858	0.281
Grantcharov et al.	8	25	33	-1.174	-2.060	-0.288
Total (fixed effects)	104	238	342	-1.059	-1.331	-0.786
Total (random effects)	104	238	342	-1.117	-1.465	-0.769

Test for heterogeneity
Q = 13.7757, DF = 9, P = 0.1305

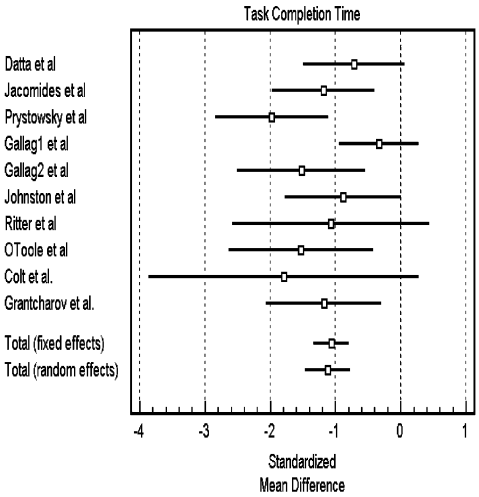

Fig. 3. Effect size values and their confidence interval estimates for the task completion time (study II).

Study	N1	N2	Total	SMD	95% CI	
Datta et al	15	15	30	-2.255	-3.239	-1.270
Jacomides et al	16	16	32	-6.165	-7.994	-4.336
Prystowsky et al	9	37	46	-0.306	-1.058	0.446
Gallag1 et al	12	100	112	-0.261	-0.867	0.346
Gallag2 et al	12	12	24	-0.539	-1.403	0.326
Ritter et al	5	6	11	-1.556	-3.216	0.103
OToole et al	8	12	20	-0.504	-1.481	0.473
Colt et al.	4	5	9	-0.914	-2.639	0.812
Grantcharov et al.	8	25	33	-1.014	-1.886	-0.143
Total (fixed effects)	89	228	317	-0.884	-1.184	-0.584
Total (random effects)	89	228	317	-1.325	-2.125	-0.525

Test for heterogeneity
Q = 51.7335, DF = 8, P < 0.0001

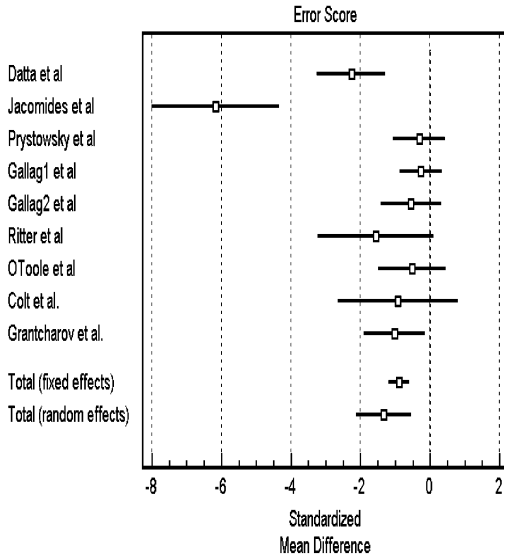

Fig. 4. Effect size values and their confidence interval estimates for the error score (study II).

show that experienced users did commit far fewer errors than the novices. It also indicates the discriminative ability of the VR simulator to separate out the two groups, and also provide for a quantitative assessment of the learning curve followed by a trainee during his/her training sessions. A possible explanation for the heterogeneity of the effect sizes in this case, which was not found with the task completion time, could be the result of the differences between the assignment of error scores, some being given as wall strikes, and some others such as inefficient movements or requests for assistance. Since dividing the studies into various subgroups based on their error score assignment techniques would have resulted in very small sized (one or two studies) subgroups the quest for searching for moderator variables or subpopulations was accordingly abandoned. When more studies are forthcoming, then it would be more appropriate to categorize them according to the specific VR simulators that they employ, and thus subgrouping according to their individual error score measurement techniques would yield a more uniform effect size values and the heterogeneity (if any) more easily explained.

IV. Discussion

Thus far, VR simulators have been mostly explored as a means of enhancing minimally invasive techniques such as laparoscopic surgery and endoscopy. Currently, there is no evidence that VR simulation-based training leads to improved patient outcome because it may be difficult to conduct such patient outcome studies. For example, it would require large cohorts of patients to be followed during and after surgery by surgeons and physicians who were randomized and have undergone different cumulative amounts of simulation training. Also, there being a large number of patient-based and system-based factors that contribute to negative outcomes, any such study would have to be massive and prolonged. To obviate all this, the surgeon's training performance, with its known limitations, has been used as the best alternative outcome. The benefits in using such a path is that in training for procedures, simulators have high face validity because they ease trainees' transition to actual patients, which seems inherently beneficial as a means to avoid adverse events. Further, it is quite reasonable to assume that a successful patient outcome is related to the training and experience of the surgeon. To that effect, this meta-analytic study has proven

that VR simulation based training is highly effective in both the transference of skills from the simulation environment to the operating room, and to help discriminate between the experienced and the inexperienced trainees. The latter would prove valuable in the design and implementation of intelligent student models into the simulation software, both for the training procedure as well as the assessment. Based on background experience, the simulation software can accordingly provide for more difficult tasks or those with intricate maneuvers, thus, in essence, providing a graded training environment. In this regard, establishing a set of learning curves specific to a VR simulator would be beneficial to assess the trainee's state and rate of learning. The term learning curve as related to surgical training has been aptly defined by Subramonian and Muir [35] as "the time taken and/or the number of procedures an average surgeon needs to be able to perform a procedure independently with a reasonable outcome." According to this definition, which is indeed a feasible one, there need to be studies done with a large sample of experienced surgeons to establish reference levels for the task completion time, the error scores, and also the rates of improvement over a specific task (such as laparoscopy or ureteroscopy on the simulator). Gallagher *et al*. [36] have done one such representative study with nearly 200 surgeons, and the results of their study and of other similar ones, once available, could be suitably employed for the establishment of the reference levels and reference learning curves. One method to establish such characteristics is to employ the meta-analysis technique that we have employed to evaluate the sample weighted effect sizes for the task completion time and the error scores over specific training periods (e.g., after 15 minutes of training) or trials. Differences between effect sizes over a specified duration could provide an estimate of the rate of skill acquisition. Trainees or students who embark on the VR simulation training program can then be assessed accordingly by determining how far or how slow they differ either in terms of percentage difference or standard deviations from the reference levels already established for the mean and rate of skill acquisition of the experienced surgeons as noted previously. Although there are some studies, such as Strom *et al*. [37], which have correctly noted the learning curves in terms of the reduction in task completion time and error scores until they reached saturation levels, most of the other studies listed in this paper (such as [24] and [9]) have just chosen two or three trials arbitrarily, and it is not clear as to their choice of those numbers. Thus, it cannot be determined whether the trainees' performance would have improved over time (in terms of trials) or not. Also, in the context of this study, the small sample sizes, the disparate assessment procedures, and the nonstandardized case scenarios could have diluted the magnitude of the effect-size and significantly contributed to the heterogeneity observed in the findings. It would therefore require a statistically significant number of studies with the same research design to establish a meta-analysis based reference learning curve for a specific VR simulator, and to make comparative assessments of the trainees' rate and state of progress in their training. Ideally, therefore, the experimental studies should be multicenter trials with a statistically sufficient number of trainees, and the skills being evaluated part of a standard surgical training course.

Also, once such studies become available, it would be worthwhile considering the incorporation of intelligent tutoring techniques to setup a student model to provide a training schedule varied in terms of training time and/or case scenarios customized to the individual user's background and learning rate. In this regard, the work done by Ota *et al*. [38] in the use of fuzzy logic is commendable, and may lead to very sophisticated VR simulators in the near future.

Currently, the limiting factors for the widespread inclusion of VR simulator based surgical training are in their high cost outlays for both personnel and setup. As in other simulation training, the dominant cost in VR simulator training also is that of instructor time and the indirect costs of removing clinical personnel (such as surgeons) from high revenue producing work to undergo experimental tests and training. The healthcare setup currently does not fully embed time or costs of training into the system, but instead often leaves these costs for the individual clinicians to bear. Furthermore, the costs of VR simulators are quite steep and vary widely. They could range from $10 000 ("home-made" ones) to $100 000, and establishing a dedicated simulation center can cost up to $1 000 000 depending on the amount of space, the type of clinical equipment to be used, the extent of renovations needed, and the sophistication of the audio-visual equipment. However, such capital costs can be amortized over a long period of time, since such centers typically are used for a wide variety of training curricula for diverse target populations. With decreasing costs of computer hardware and software, this will probably ease up the situation with time. On the whole, training with virtual reality simulators (especially in the very near future) may still be cost-effective compared to current training (resources and human power) costs [39]. Recommendations for their inclusion in the surgical training program are already put forth by the Royal Society of Obstetricians and Gynecologists in the United Kingdom [40].

Given the high costs of VR simulator training as mentioned above, it is therefore not surprising that there are very few studies pertaining to the effects of training with such simulators using surgeons and physicians. Indeed, the largest market for VR training simulators are the medical and dental schools and colleges. However, the results of our meta-analysis study support the growing evidence that the incorporation of virtual simulation technology will engender a revolution in surgical procedural training of residents and students. Compared to the traditional system of developing competence by seeing and doing procedures on real patients or cadavers, the desired skills can be readily acquired with virtual simulation. Such an environment includes opportunities for repetition, feedback, and correction of errors, and the ability for customized individual training, which are the components of any educational process. At the very least, they will enhance supplemental learning from an expert instructor on a one-to-one basis. Also, as simulators become more advanced, especially with appropriate sensory feedback, they may be reasonable substitutes to improve proficiency of medical students and residents, and also provide for exercises in dexterity for experienced surgeons and physicians. This increase in proficiency may have an important impact in patient outcomes.

Although the results from our meta-analytic study are indeed very encouraging for the inclusion of VR surgical simulators into any surgical training program, it should be noted that there are potential risks in simulation-based training. Where the simulator cannot properly replicate the tasks or task environment of operating or caring for patients, there is a risk that clinicians might acquire inappropriate behaviors (negative training) or develop a false sense of security in their skills that could theoretically lead to harm. Although there are no data to suggest that this currently happens, such risks will have to be weighed and evaluated as simulators become more commonly used. In conclusion, despite the small sample sizes and disparate measures used for the error assessment of the trainees, a more optimistic interpretation of the findings of this meta-analytic study is that VR simulator training allows for the acceleration of the training curve as also gives exposure to novel procedures and technologies.

V. Conclusion

In this paper, we have reported the results of the meta-analysis of the efficacy of surgical virtual reality simulators in the transference of skills to the operating room and also their discriminative abilities. The available observational data provides evidence that training with virtual reality simulators does lessen the time and the errors in the performance of a given surgical task, and, furthermore, it can clearly differentiate between the less experienced and the experienced trainees among its users. However, many more future studies on the criterion and content validity of the virtual reality simulators and the link between simulator-based training and performance on actual patients need to be performed to better assess the appropriate role of virtual reality simulators in surgical training and patient healthcare. Also, once the efficacy of VR simulator training has been firmly established (using studies such as ours) their cost-benefit analyses need to be evaluated. With possible incorporation of artificial intelligence components (currently none or very minimal) and more sensory feedback, it is our belief that virtual reality simulators will prove to be very effective tools for training in the skills necessary for surgery and endoscopies.

References

[1] R. S. Haluck and T. M. Krummel, "Computers and virtual reality for surgical education in the 21st century," *Arch. Surgery*, vol. 135, pp. 786–92, 2000.

[2] R. M. Satava, "Medical virtual reality, the current status of the future," in *Health Care in the Information Age*, S. Weghorst, H. Siegurg, and H. K. Morgan, Eds. Amsterdam, The Netherlands: IOS Press, 1996, pp. 542–545.

[3] R. M. Satava, "Virtual reality surgical simulator: The first steps," *Surgical Endoscopy*, vol. 7, pp. 203–205, 1993.

[4] S. B. Issenberg, W. C. McGeghie, and I. R. Hart, "Simulation technology for health care professional skills training and assessment," *J. Amer. Medical Assoc*., vol. 282, pp. 861–867, 1999.

[5] T. Hikichi, A. Yoshida, S. Igarashi, N. Mukai, M. Harada, K. Muroi, and T. Terada, "Vitreous surgery simulator," *Arch. Ophthalmology*, vol. 118, pp. 1679–1681, 2000.

[6] J. Torkington, S. G. T. Smith, B. I. Rees, and A. Darzi, "Skill transfer from virtual reality to a real laparoscopic task," *Surgical Endoscopy*, vol. 15, pp. 1076–1079, 2001.

[7] K. Moorthy, Y. Munz, S. K. Sarker, and A. Darzi, "Objective assessment of technical skills in surgery," *British Medical J*., vol. 327, pp. 1032–1037, 2003.

[8] N. Taffinder, C. Sutton, R. J. Fishwick, I. C. McManus, and A. Darzi, "Validation of virtual reality to teach and assess psychomotor skills in laparoscopic surgery: Results from randomized controlled studies using the MIST VR laparoscopic simulator," *Stud. Health Technol. Infor.*, vol. 50, pp. 124–130, 1998.

[9] D. M. Wilhelm, K. Ogan, C. G. Roehrborn, J. A. Cadeddu, and M. S. Pearle, "Assessment of basic endoscopic performance using a virtual reality simulator," *J. Amer. College Surgeons*, vol. 195, pp. 675–681, 2002.

[10] A. M. Pearson, A. G. Gallagher, J. C. Rosser, and R. M. Satava, "Evaluation of structured and quantitative training methods for teaching intracorporeal knot tying," *Surgical Endoscopy*, vol. 16, pp. 130–137, 2002.

[11] J. D Watterson, D. T. Beiko, J. K. Kuan, and J. D. Denstedt, "Randomized prospective blinded study validating acquisition of ureteroscopy skills using computer based virtual reality endourological simulator," *J. Urology*, vol. 168, pp. 1928–1932, 2002.

[12] M. Gor, R. McCloy, R. Stone, and A. Smith, "Virtual reality laparoscopic simulator for assessment in gynaecology," *BJOG: Int. J. of Obstetrics and Gynaecology*, vol. 110, pp. 181–187, 2003.

[13] J. V. Rossi, D. Verma, G. Y. Fujii, R. R. Lakhanpal, M. S. Humayun, and E. De Juan, "Virtual vitroretinal surgical simulator as a training tool," *Retina*, vol. 24, pp. 231–236, 2004.

[14] D. Ost, A. DeRosiers, E. J. Britt, A. M. Fein, M. L. Lesser, and A. C. Mehta, "Assessment of a bronchoscopy simulator," *Amer. J. Respiratory Critical Care Med.*, vol. 164, pp. 2248–2255, 2001.

[15] L. B. Gerson and J. Van Dam, "A prospective randomized trial comparing a virtual reality simulator to bedside teaching for training in sigmoidoscopy," *Endoscopy*, vol. 35, no. 7, pp. 569–575, 2003.

[16] R. Rowe and R. A. Cohen, "An evaluation of a virtual reality airway simulator," *Anesthesia and Analgesia*, vol. 95, no. 1, pp. 62–66, 2002.

[17] G. Ahlberg, T. Heikkinen, L. Iselius, C. E. Leijonmarck, J. Rutqvist, and D. Arvidsson, "Does training in a virtual reality simulator improve surgical performance?," *Surgical Endoscopy*, vol. 16, pp. 126–129, 2002.

[18] A. Hyltander, E. Liljegren, P. H. Rhodin, and H. Lonroth, "The transfer of basic skills learned in a laparoscopic simulator to the operating room," *Surgical Endoscopy*, vol. 16, pp. 1324–1328, 2002.

[19] N. E. Seymour, A. G. Gallagher, S. A. Roman, M. K. O'Brien, V. K. Bansal, D. K. Andersen, and R. M. Satava, "Virtual reality training improves operating room performance results of a randomized, double-blinded study," *Annals of Surgery*, vol. 236, pp. 458–464, 2002.

[20] T. P. Grantcharov, V. B. Kristiansen, J. Bendix, L. Bardram, J. Rosenberg, and P. Funch-Jensen, "Randomized clinical trial of virtual reality simulation for laparoscopic skills training," *British J. Surgery*, vol. 91, pp. 146–150, 2004.

[21] V. Datta, M. Mandalia, S. Mackay, and A. Darzi, "The preop flexible sigmoidoscopy trainer, validation and early evaluation of a virtual reality based system," *Surgical Endoscopy*, vol. 16, pp. 1459–1463, 2002.

[22] L. Jacomides, K. Ogan, J. A. Cadeddu, and M. S. Pearle, "Use of a virtual reality simulator for ureteroscopy training," *J. Urology*, vol. 171, pp. 329–323, 2004.

[23] J. B. Prystowsky, G. Regehr, D. A. Rogers, J. P. Loan, L. L. Hiemenz, and K. M. Smith, "A virtual reality module for intravenous catheter placement," *American J. Surgery*, vol. 177, pp. 171–175, 1999.

[24] A. G. Gallagher, A. B. Lederman, K. McGlade, R. M. Satava, and C. D. Smith, "Discriminative validity of the minimally invasive surgical trainer in virtual reality (MIST VR) using criteria levels based on expert performance," *Surgical Endoscopy*, vol. 18, pp. 660–665, 2004.

[25] A. G. Gallagher, K. Ritchie, N. McClure, and J. McGuigan, "Objective psychomotor assessment of senior, junior and novice laparoscopists with virtual reality," *World J. Surgery*, vol. 25, pp. 1478–1483, 2001.

[26] E. M. Ritter, D. A. McCluskey, A. B. Lederman, A. G. Gallagher, and D. Smith, "Objective psychomotor skills assessment of experienced and novice flexible endoscopists with a virtual reality simulator," *J. Gastrointestinal Surgery*, vol. 7, pp. 871–878, 2003.

[27] R. V. O'Toole, R. R. Playter, T. M. Krummel, W. C. Blank, N. H. Cornelius, W. R. Roberts, W. J. Bell, and M. Raibert, "Measuring and developing suturing technique with a virtual reality surgical simulator," *J. Amer. College Surgeons*, vol. 189, pp. 114–127, 1999.

[28] H. G. Colt, S. W. Crawford, and O. Galbraith, "Virtual reality bronchoscopy simulation," *Chest*, vol. 120, pp. 1333–1339, 2001.

[29] T. P. Grantcharov, L. Bardram, P. Funch-Jensen, and J. Rosenberg, "Learning curves and impact of previous operative experience on performance

on a virtual reality simulator to test laparoscopic skills," *Amer. J. Surgery*, vol. 185, pp. 146–149, 2003.
[30] L. V. Hedges and I. Olkin, *Statistical Methods for Meta Analysis*. San Diego, CA: Academic, 1985.
[31] R. DerSimonian and N. Laird, "Meta analysis in clinical trials," *Controlled Clinical Trials*, vol. 7, pp. 177–188, 1986.
[32] K. Ogan, L. Jacomides, M. J. Shulman, C. G. Roehrborn, J. A. Cadeddu, and M. S. Pearle, "Virtual ureteroscopy predicts ureteroscopic proficiency of medical students on a cadaver," *J. Urology*, vol. 172, pp. 667–671, 2004.
[33] S. N. Kothari, B. J. Kaplan, E. J. DeMaria, T. J. Broderick, and R. C. Merrell, "Training in laparoscopic suturing skills using a new computer-based virtual reality simulator (MIST-VR) provides results comparable to those with an established pelvic trainer system," *J. Laparoendoscopic and Advanced Surgical Techniques*, vol. 12, pp. 167–173, 2002.
[34] E. C. Hamilton, D. J. Scott, J. B. Fleming, R. V. Rege, R. Laycock, P. C. Bergen, S. T. Tesfay, and D. B. Jones, "Comparison of video trainer and virtual reality training systems on acquisition of laparoscopic skills," *Surgical Endoscopy*, vol. 16, pp. 406–411, 2002.
[35] K. Subramonian and G. Muir, "The learning curve in surgery: What is it, how do we measure it and can we influence it?," *BJU Int.*, vol. 93, no. 9, pp. 1173–1174, 2004.
[36] A. G. Gallagher, C. D. Smith, S. P. Bowers, N. E. Seymour, A. Pearson, S. McNatt, D. Hananel, and R. M. Satava, "Psychomotor skills assessment in practising surgeons experienced in performing advanced laparoscopic procedures," *J. Amer. College Surgeons*, vol. 197, pp. 479–488, 2003.
[37] P. Strom, A. Kjellin, L. Hedman, E. Johnson, T. Wredmark, and L. Fellander-Tsai, "Validation and learning in the procedicus ksa virtual reality surgical simulator," *Surgical Endoscopy*, vol. 17, pp. 227–231, 2003.
[38] D. Ota, B. Loftin, T. Saito, R. Lea, and J. Keller, "Virtual reality in surgical education," *Comput. Med. Bio.*, vol. 25, pp. 127–137, 1995.
[39] M. Bridges and D. L. Diamond, "The financial impact of teaching surgical residents in the operating room," *Amer. J. Surgery*, vol. 177, pp. 28–32, 1999.
[40] P. Bowen-Simpkins *et al*. "Discussion Document on Further Training for Doctors in Difficulty," Royal College of Obstetricians and Gynaecologists, Apr. 2002.

Syed Haque (M'90) received the Ph.D. degree in educational systems development from Michigan State University, East Lansing, MI.

Currently, he is a Professor and Chair of the Department of Health Informatics at the University of Medicine and Dentistry of New Jersey, Newark, NJ. His research interests include health care outcomes measurement, and research and health care data mining. He has published extensively in journals and in several conferences, and served as Chair in several international conferences.

Dr. Haque received an Excellence in Teaching Award and the Process Improvement and Process Innovation (2002) Award from University Hospital. He has been recognized/certified as an Internationally Registered Technology Specialist (IR7), by the Secretariat to the IRTS Council, International Technology Institute, San Diego, CA. He is also on the Editorial Board of the *Journal on Information Technology in Healthcare*. He is a Member of AMIA.

Shankar Srinivasan received the M.Sc. and Ph.D. degrees in biomedical engineering from the University of Saskatchewan, Saskatoon, Canada.

He is an Assistant Professor in the Department of Health Informatics at the University of Medicine and Dentistry of New Jersey (UMDNJ), Newark, NJ. Prior to his appointment at UMDNJ, he taught various topics related to biomedical and electrical engineering at universities in Singapore and Australia. His research interests are mainly in the areas of health care data analysis, neuro-fuzzy systems, and enterprise-wide health care web services. He has published in various international conference proceedings and journals, and has served as a Chair for sessions on biomedical engineering in two international conferences.