Proceedings of the ASME 2011 World Conference on Innovative Virtual Reality
WINVR2011
June 27-29, 2011, Milan, Italy

WINVR2011-5554

A GLOBAL APPROACH TO THE DESIGN AND EVALUATION OF VIRTUAL REALITY MEDICAL SIMULATORS

Sofia Bayona
Universidad Rey Juan Carlos
Mostoles, Madrid, Spain

Jose Manuel Fernandez-Arroyo
Hospital Severo Ochoa
Leganes, Madrid, Spain

Pilar Bayona
Instituto de Psicologia
Integral, Salud y Coaching
Avanza
Madrid, Spain

Isaac Martin
Universidad Rey Juan
Carlos
Mostoles, Madrid, Spain

ABSTRACT

VR Simulators are a powerful alternative to traditional educational techniques in many domains; and in particular, in surgery. Although they offer new possibilities for learning, training and assessment, they still found difficult to be accepted and integrated into hospitals.

In this paper, we explain what we consider the key issues to create successful VR simulators, and we present two methodologies: the guidelines for the simulator design and the evaluation of their validity.

Research on VR surgical simulators should be interdisciplinary. It involves medicine, educational psychology, computer science, and engineering. Optimal interdisciplinary communication is difficult, and most projects in surgical simulation are strongly influenced by the engineering perspective, with little or no contributions from the others. This unbalance often leads to a premature end of the project or to simulators which are less practical for surgeons.

A design methodology should be used as a guide in the process of creating VR simulators. A thorough description of the problem, the simulator's role, and an exhaustive task analysis will lead to the identification of the requirements. For the technical implementation, decisions will be taken related to the hardware interface and the interaction that users will have with the virtual world; which will determine collision detection and response algorithms, and the behaviour of the 3D models. In addition to the technical testing, it is necessary to prove the validity of the simulator and design procedures to measure the user performance.

We explain a methodology to evaluate the validity (face, content, and criterion-related validity), reliability and transfer of skills from a VR simulator to the real environment in a structured and rigorous way. Following this methodology, an evaluation experiment involving 19 orthopaedic doctors using a VR arthroscopy simulator was carried out. Results prove face and content validities, and inform about the factors and measures that are considered important for arthroscopic surgery.

In order to consolidate the research results, we encourage the establishment of an intersectorial consortium with agents from the academic, healthcare and industrial sectors to ensure the long-term sustainability of research lines, additional funding, and to guarantee that simulators, once validated, can be widely available in hospitals.

This paper presents a global approach including relevant guidelines and methodologies for designing and evaluating VR simulators. It can provide a solid structure for other researchers when facing those processes and contribute to the successful integration of VR simulators within the educational curriculum.

KEY WORDS

Virtual Reality, Surgical Simulation, System Design, System Evaluation, Healthcare Education, Assessment, Experimental Research, Interactive Technology, Innovative Education

INTRODUCTION AND MOTIVATION

Virtual Reality Simulators are a powerful alternative to traditional educational techniques in many domains. In particular, simulators are necessary when learning involves risks, high costs or both.

Virtual reality (VR) can facilitate learning by reducing the stress for the trainee, since consequences cease to be critical, and the training session is no more a single-opportunity operation. As the trainee progresses, if required, stressful stimuli can be gradually incorporated.

For those tasks that involve high risks, failure in virtual environment does not imply high risk. Apprentices can learn from their own mistakes. They start to recognize hazardous situations, learn how to recovery from errors, and how to remediate them in case they could not be avoided.

Virtual Reality allows the creation of learner-centred scenarios designed to optimize the learning process. In this way, the learning process can be adapted to the level and features of the learners, trying to stay within the Zone of Proximal Development [1]. Consequently, the VR system can adjust the difficulty according to the level of each apprentice so that learning continue to be both challenging and motivating, but always within the learner's grasp limits.

Virtual Reality can offer a non-degradable realistic environment, in which novices may learn and practice as much as they want. With VR simulators, learning can be asynchronous, adapting to the learners' needs and leaving time for reflection. It overcomes the limitations of needing to be in the real environment under the constant supervision of an expert.

In surgical education, many procedures are complex to perform. They have a long learning curve and require specialized training in order to obtain proficiency. Patient safety is at risk, so surgeons should be assessed to guarantee their competence.

Benefits of VR new solutions for healthcare are obvious, then, why simulators are not completely integrated as part of the surgical curriculum yet?

Creating VR simulators is a high complex task concerning aspects such as: task analysis, learning methods, hardware interface, realistic feedback, 3D modeling, user interaction, collision detection and response algorithms, physical simulation, simulation of the behavior of the model, efficiency, performance assessment, evaluation processes, or economical costs. The limitations of the technology, the lack of knowledge about the tasks and aims, and the conflict between the different aspects involved can make it really arduous.

In some cases, simulators evolve as research projects in the academic sector, and they never manage to get to the market. Sometimes failure is due to a too specific and rigid design that does not allow extensibility. In other cases simulators become commercial, but they do not succeed in proving their instructional power and transferability of skills to the real settings [2]. Therefore, hospital managers need argument to support and justify the investment.

Decision making during the design, development and evaluation of VR simulators is critical and can lead to failure if the awareness about the overall direction is not acquired. This paper provides a global approach, elucidating key points so that these decisions can take into account the whole big picture.

The next section highlights the importance of forming an interdisciplinary teamwork. Then, we describe in detail the steps of the proposed design methodology. Afterwards, an evaluation methodology of VR simulators is explained. As an example of applying this evaluation methodology, we present an evaluation study involving 19 orthopaedists using a VR arthroscopy simulator. Subsequently, we show the benefit of forming an intersectorial consortium to reach the objective successfully. In the conclusion, we present the impact of undertaking this global approach in designing VR medical simulator.

INTERDISCIPLINARY RESEARCH

As mentioned in the introduction, wide-ranging knowledge is involved in the creation of VR simulators. Each aspect belongs to its particular discipline, but choices made according to one discipline can affect the others. This is why interdisciplinary team is essential.

Multiple disciplines need to work together harmoniously towards the ultimate objective. This is more complicated that it seems. Not only people from different disciplines need to communicate into a way that makes possible to understand each other, but also, different disciplines need to learn from each other to cooperate. Competition for resources or recognition needs to be put apart so that interdisciplinary working will make possible to attain the ultimate objective. Disciplines working harmoniously will share the satisfaction of their mutual success and accomplishments.

The overall goal needs to be clearly established from the beginning. It can be to create a VR simulator for learning and training; for assessing surgeons; for preoperative rehearsal with patient-specific data; or creating a platform to test new procedures and techniques with no risk to the patient. All the different disciplines: surgery, educational psychology, and computer science and engineering need to embrace and understand the common goal.

For example, the involved disciplines in the case of a VR surgical simulator whose purpose is to be a complementary tool for learning and training will be:

- Surgery: Surgical techniques and skills must be thoroughly studied to understand how they are acquired so that learning and training deficiencies can be prevented,
- Educational psychology: Task analysis needs to be done during the design process to define the educational outcomes and a complete learning programme that facilitates learning towards competence. Then, they can follow the evaluation methodology explained in this paper to design experiments to validate the simulator. Feedback is essential to direct learning. The simulator will need to be able to provide feedback to the users about their performance. Surgeons, educationalists and engineers will have to collaborate to define this feedback and the assessment measures that the simulator will calculate.
- Computer Science and Engineering: Engineers know the advantages and limitations of technology, and they will put technology to serve the global objectives. Communication

needs to be done so that the rest of disciplines will be able to understand the possibilities and resources of VR. All disciples will share their vision and suggest ideas. Having economical and efficiency costs into account, simplifications to achieve real time response will be carried out while assuring an adequate degree of realism depending on the aim and the role of each particular VR simulator. The design methodology proposed in this article can guide this process and help in the decision making.

As it can be seen, a harmonious collaboration between the different disciplines will be one of the key factors to success. If the final parts are going to work together, they must be developed by groups that share a common picture of what each part must accomplish. This article and the methodologies proposed intends to provide guidelines that will support this process and help in the decision making.

DESIGN METHODOLOGY

Designing a VR simulator is a complex task. This article intends to provide guidance to structure this process and to help the interdisciplinary team to create a global vision since the beginning of the project, so that better decisions can be made.

Description of the problem and the role of the VR simulator

The research team needs to define the role of the simulator to be designed. It could be a VR simulator for pre-operative planning of a particular intervention, or a simulator for assessing surgeons.

For example, if the task consists of designing a simulator that can be used as a complementary tool for training surgeons in a particular procedure, it is necessary to answer the following questions: What are the difficulties that surgeons face when learning this procedure? What are the current ways of training surgeons in this procedure? How they have been learning it so far? What are the advantages and disadvantages? What would be the relevance of improving training in this particular procedure?

Answering these questions will help defining the problem. Doctors and educationalist have to study the advantages and disadvantages of current methods for training that procedure, and the possibilities of building a VR simulator as a complementary training tool.

Applying didactic VR resources: aims, feedback and performance measurement

A good starting point to better defining the aims of the simulator is to think about the VR didactic resources. According to Lamata et al. [3], these didactic resources can be categorised as:

- Fidelity resources: Simulators create environments that approximate reality, and the different aspects of the fidelity employed in this reconstruction of the real world are the first category of didactic resources. This engineering or physical fidelity is the degree to which the training device or environment replicates the real task’s physical characteristics. This contrasts with psychological or functional fidelity, or the degree to which the skill or skills in the real task are captured in the simulated task.
- Teaching resources: VR simulators also offer features unique to a computer-simulated environment that can enhance training. These include cues and instructions given to the user to guide a task, or features to manage a training program.
- Assessment resources: Assessment resources offer evaluation metrics to assess performance and follow up progress, and ways to deliver constructive feedback to the user.

The interdisciplinary team should specify the objectives and the kind of didactic resources considered for the VR simulator.

The team have to discuss about the **degree of realism** required. Clinical people know what they want (often the more realistic, the better) but they often ignore the costs and the extent of the system’s feasibility. Computer scientists and engineers will have to study if current technology is able to provide that level realism and, in that case, what would be the associated costs.

From the beginning the cost and the required realism must be taken into account because these issues will determine the cost-benefit **i**nterplay of designing and implementing a VR simulator. This estimation will help to a make the right decision about investing in the development of the VR simulator. The team of psychologists and educationalists will carry out the task analysis. This analysis consists of decomposing the task to into elements called subtasks. It is a strategy based on the curriculum of competences. Literature can be found about on how to perform task analysis in [4]. The most common options for doing task analysis are hierarchical decomposition and the technique of critical incidents (in which key events that can lead to success or failure are identified). They will define the teaching resources so that the VR simulator can offer an adequate level of guidance and help to facilitate learning.

It is important to avoid that apprentices acquire bad habits when using the simulator, for this, the learning programme should include a didactic sequence to guide the apprentice through the different stages, starting with how to handle surgical instruments in an ergonomic way, explaining their functioning and start performing easy tasks that will become more difficult as the apprentice progresses.

Based on educational theories [5] they will design the exercises and decide on the kind of feedback about the performance that the apprentice will receive and when this feedback will be provided. Depending on the purpose and the moment in which this feedback is provided, we can classify feedback as follows:

- **Formative feedback** occurs during the process of a course or a training programme. It is a process that intends to enhance student's learning and to promote student's accomplishment. Normally it is designed to provide the learner with information on their progress but it does not contribute to the final assessment. This feedback is useful to the learner to advise them about their progress compared with competency standards, to recognise mastered areas or tasks that the student does not know or is unable to do. It helps the student to make a self-evaluation, to provoke questions, to guide the learning process and to identify strengths and weaknesses.

- **Summative feedback**, on the contrary, is usually done at the end of a course or of some larger educational programme to determine success. It is a formal test to determine what has been learned. It summarises the development of learners at a particular time, providing a snapshot of a student's level in relation to the end-of-year expected level. Summative assessment is designed to determine grades or marks according to the intended learning outcomes. The outcome of this summative assessment can be an award of qualification, a demonstration of the level of competence, a barrier to progression or it can be used to generate a relative measure of attainment, by establishing ranking, or awarding merits or distinctions.

Quantitative objective data for measuring dexterity and performance is essential both for learning and for assessing purposes. The simulator could count with assessing resources and register metrics such as time for completing the task, number of errors, economy of movement, smoothness in handling anatomical tissue, knowledge of the procedure, movement of the surgical instruments, anatomical knowledge or ambidexterity. We also discuss the importance of these assessing measures in the section about evaluation of VR simulators.

3D models

Decisions will be made about the 3D models that will be used for the VR simulation. They could be synthesised generic models; models with specific pathologies; or even patient-specific models, obtained from medical image data of real patients. Again, the degree of realism will need to be chosen according to the purpose of the simulator. For example, for a simulator of basic surgical skills a generic model could be the best option, whereas for a pre-operative planning VR simulator, we will probably need patient-specific 3D models. The surgical instruments and their functionality will also need to be modelled.

Since volumetric representations require more resources, unless it is strictly necessary, most VR simulators will use superficial representations of 3D models.

Interface

According to Burdea, Virtual reality is: *"a high-end user-computer interface that involves real-time simulation and interactions through multiple sensorial channels. These sensorial modalities are visual, auditory, tactile, smell, and taste"* [6].

During the design, in order to decide about the interaction between the user and the VR simulator and what kind of sensorial feedback provide, the team can answer to the following questions: Which virtual objects can be manipulated and how? How the surgical instruments will be simulated? How many degrees of freedom do they have? If the position of the user needs to be known, which tracking device could be more appropriate?

Before defining the required hardware devices, first we need to know what feedback will be given to the user. Most frequently, the VR simulator provides visual feedback (for example, in a 2D monitor or in a 3D stereo display), haptic feedback to provide touch and/or force feedback, and audible stimuli (with stereo sound or 3D sound).

Depending on the particularities of the task, hardware modifications or new designs may be necessary. Again, decisions can be made to increase realism or to simplify the feedback provided. Designers will pay special attention to ergonomics.

Interactions: collision detection, collision response, behaviour and deformation of virtual objects

Virtual scenes will contain a number of 3D objects. We need to simulate the interactions between those virtual objects.

During the simulation, the objects will move and contacts will be produced. We will use collision detection algorithms to detect those contacts. There are different collision detection algorithms: some of them are based in hierarchical structures of bounding volumes, others use techniques based on distance fields, stochastic models or spatial partition methods. Some good surveys about collision detection can be found in [7,8].

On the other hand, we have to simulate the behavior of the virtual objects, which will respond to the collisions and forces applied on them. These methods could be classified in mesh-less methods and mesh-based methods. For a detailed description, refer to [9].

In ordr to select the algorithms to be used, the following questions should be answered: What could be the colliding objects? What will happen when they collide? Are the objects rigid or deformable? Is it possible for an object to collide with itself (self-collisions)? Do we need to exploit temporal coherence? How do the objects behave? Are there big deformations? Do we need to cut or to topologically change the virtual models? What is the degree of realism and accuracy required?

Here again, we will have to look for simplifications to find a balance between accuracy and speed. We could have hybrid scenarios and use different techniques for the different type of objects (rigid and deformable). In some cases, we may not need physically realistic behaviour, and a plausible reaction will be enough. In other cases, accuracy might be necessary.

Software architecture

Regarding the software architecture of a VR simulator, it could be implemented as suggested in [10] using three threads:

- A thread in charge of the graphical output. It should run at rates between 25Hz and 60Hz and it will have low priority.
- A thread in charge of reading the position of the tracking systems or haptic devices (if applicable), calculating the collision detection, and the deformations of the 3D virtual objects (if any). This thread could run at rates between 80Hz and 150Hz.
- A thread that computes the output forces that the haptic devices will give to the user. Usually, haptic devices need to be refreshed at 1000Hz, so this thread needs to be optimized and have the highest priority assigned.

As in every software development, designers should take care of efficiency, modularity, reusability, robustness, reliability, maintenance, extensibility, scalability and usability.

If the design solutions are based on the deep knowledge about the real needs of the tasks and procedures, the probability of failure will decrease. Once the VR simulator is designed, implemented and tested from the technological point of view, it will still need to be evaluated to prove its benefits as an educational tool, as it is detailed in the next section. It is important to note that it is useful to detail the evaluation process at the design stage, since it will influence some of the decisions. Particularky, the design decisions affecting the assessment of the available resources or user performance metrics.

METHODOLOGY FOR THE EVALUATION OF VR SIMULATORS

A complete review of evaluation studies on surgical simulation is presented by Sutherland et al. [11]. They point out that most studies in surgical simulation have been of poor experimental design.

As a basis for the evaluation of VR surgical simulators, we propose the following methodology. More detail and a practical application can be found in [12].

In order to evaluate a VR simulator we need to know its purpose. Most frequently, simulators are intended for training, assessment or both:

- For evaluation of their training potential, we would like to test the *instructional effectiveness* of surgical simulation [13] that is to say that repeated use improves performance. The final aim is to test if the skills acquired and trained with the simulator are transferable to the operating room.
- On the other hand, before the simulator can be used for assessment purposes, it needs to prove that it is both reliable, showing different assessments on the same individual are similar under different circumstances, and valid, i.e. seeing if the assessment measures what is intended to measure [14].

The first stage is the definition of the evaluation objectives and the formulation of hypothesis. The objectives should be relevant, concrete, consistent, assessable, and feasible.

We can start by posing questions that we would like to be answered. By instance, we could ask “Will the simulator be useful to learn how to handle a particular surgical instrument?” or “Will the simulator be able to measure the correctness of an apprentice performing a procedure?” The first question is about using the simulator for learning, while the second one will be focused on the assessment of the correctness of a particular task. Posing these questions will help us to better define our desired evaluation outcomes. To transform these questions into objectives, we can try to relate or classify these questions according to the different types of validities.

Next, a classification of the different types of validities, with their definitions gathered and synthesised from different sources [15-19] is shown.

- **Face validity** is the extent to which the examination resembles real life situations. In the virtual simulation domain, it could suggest how the simulator reflects the real operation. Face validity can be understood as the degree to which a measure appears to be valid to the observed subject. It is important because it can influence the degree of cooperation of the person being observed. This type of validity is subjective.
- **Content validity** is the degree to which elements of an assessment instrument are relevant, adequate and representative of the target for a particular assessment purpose. It is the extent to which the domain that is being measured is measured by the tool. Experts examine in detail the contents to establish if they are appropriate and specific to a particular situation. As Moorthy exemplifies, it could happen, for example, that while trying to measure technical skills, we may actually be testing knowledge [15]. Determining content validity is subjective and it is based on the experts’ judgments.
- **Construct validity** tries to ensure the existence of a psychological construct which underlies and gives meaning and significance to the test scores. One inference for evaluating the construct validity of a surgical simulator would be to study whether the surgical skills improve with practice in the simulator. Another inference is the extent to which a test discriminates between various levels of that construct. A usual way of evaluating construct validity is to study the ability of an assessment tool to differentiate between experts and novices performing a task. For example, if our construct is anatomy skills, the test should be able to discriminate between anatomy experts and novices.
- **Criterion related validity** is the degree to which the assessment tool predicts other objective measures of performance. To determine criterion validity, we need at least two different measure situations: one is the predictor (X) and the other is the criterion (Y). The same subjects are measured and two observations are

obtained as a result: one with X and the other with Y. Criterion validity is often divided into concurrent validity and predictive validity:

  - **Concurrent validity** is the extent to which the results of the tool correlate with the gold standard for that domain. It can also be defined as the relationship between the test scores and the scores on another external instrument purporting to measure the same target. For example, if the gold standard for evaluating Digital Rectal Examination (DRE) procedure was to evaluate surgeons with a plastic model, we could compare a VR simulator for assessing DRE with the traditional plastic model to look for correlation.
  - **Predictive validity** tries to determine the effectiveness of the test to predict a variable of interest. If the scores that apprentices get in a VR surgical simulator are a good predictor of the level of performance that these apprentices will have in the real setting, that is, in operating room, the simulator will have proven predictive validity. Once a simulator is able to achieve predictive validity, we could consider the possibility of using it as an assessment tool.

We can classify the initial evaluation questions according to the different types of validities. This will help in defining the evaluation objectives and assigning them a priority depending on their feasibility and relevance. With the definition of the objectives and hypothesis, the researcher can start designing the evaluation study, which will consider the ethical and legal regulations in force.

When preparing an ethics application, the ethical aspects (based on the basic principles of respect, beneficence and justice) covered in the Declaration of Helsinki, and the Belmont Report have to be taken into account:

- The design of the experiment will include the informed consent, and specify the data to be collected and how to deal with the correspondent data protection issues.
- The study should assure fairness in the selection, distribution and monitoring of subjects.
- If the experiment requires the handling of hazardous substances, safe facilities and trained personnel should be ensured.
- The relevant risks, benefits and uncertainties related to the experiment must be clearly exposed so that the Ethical Committee can study if the Risk-Benefit Ratio is favourable.

A thorough classification of different studies according to different criteria can be found in [12]. Here, we summarize them in Table 1.

- Depending on its purpose: descriptive or analytical studies
- Depending on the temporal direction: transverse or longitudinal studies
- Depending on the start time of the study: prospective or retrospective studies
- Depending on the allocation factor study, they can be divided into:
  - Observational studies: cohort or case-control.
  - Experimental studies: the definition of the presence or absence of a control group and the method of allocating subjects to groups will lead to studies without control group, randomized controlled clinical trials, or not randomized controlled clinical trials.

Table 1. Types of evaluation studies

If the study is experimental, the researcher needs to specify the explicative variables, which can be divided into dependent and independent:

- The independent variable/s is/are the one/s whose value is intentionally manipulated or changed by the researcher because s/he expects it to influence into the dependant variables.
- The dependent variables receive the effect of the independent variables.

For example, we could want to study the effect of training with a VR simulator into the suture skills of surgeons. For that, we could divide subjects into two groups. The independent variable in this case would be “having done training with the VR simulator”. One of the groups will do training and the other will not. The dependant variables could be measures about the suture skills, such as time for completing the task, manual precision, correctness in the procedure or the final result of the stitches. We want to analyse, if there is a relation between the independent and the dependent variables.

Within the design of the study, the researcher should also try to minimize the effect of the extraneous variables. Extraneous variables are those variables that are not controlled by the researcher and that can influence the dependent variables. Extraneous variables can affect the results introducing a bias.

Sometimes, a habituation study to habituate the subjects to the use of the tool may be required before the final evaluation procedure.

The next step will be to clearly describe all the steps of the intervention protocol, so that every subject will receive the same information and go through the same steps. This will include specifying for example, how many training sessions will be, of what length, the resting intervals between sessions, and what each session will consist of.

Analogously the observation protocol for the experiment needs to be designed. If we defined dependant variables, we need to describe how they will be measured. For example, we may want to observe the subject's performance. For that, we can gather some objective data based on variables provided by the VR simulator such as time, and some subjective data by making an observer give a score to the subject.

Researchers will decide which statistical analysis they will apply to the data and if the subject identity will remain anonymous.

If possible, the study must be blind, so that both the observers and the researchers that will analyze the data ignore to which group each subject belongs to. This will prevent them from introducing a bias. Sometimes we can also make subjects ignore to which group they belong to. For example, the experiment could be to study two particular haptic responses (independent variable) to see which one allows a better handling of the surgical instruments (dependent variable). We could assign subjects to different groups, each group with a particular haptic response, and subject could ignore even fail to notice that the existence of different groups.

A feasibility study will be done taking into account the available resources, infrastructures, the time to get the ethical approval and the costs involved.

It is advisable to conduct a pilot study before conducting large experiments. The pilot study will involve fewer resources and will have a smaller sample size. It can be very useful to identify potential problems regarding both the intervention and the observation protocols.

The information and data obtained in the pilot study will lead to corrections or modifications in the design and the protocols, and will be used as an orientation of what might be found in the final study. The pilot study, if appropriate, will allow the adjustment of the final study.

Then, the final large study will be conducted. Data will be recorded and analyzed, and conclusions will be drawn.

EVALUATION STUDY

As an example of how the presented evaluation methodology can help in the evaluation of a VR simulator, we designed and carried out an evaluation study.

The present study is the first step in the evaluation of a virtual reality commercial simulator for training arthroscopic skills [20].

The arthroscopic simulator consists of a platform composed by two devices that provide force feedback and that simulate the arthroscope (an optical instrument inserted through a small incision that allows doctors to view the interior of a joint) and the instrumental. A monitor shows the arthroscopic view consequently with the simulated arthroscope's movements.

According to the methodology, the first step is to raise the questions we want to answer and to relate them with the correspondent type of validities in order to define our hypothesis and objectives.

The questions we want to answer are: Will surgeons consider the simulator as a useful learning method for acquiring arthroscopic skills? Will it be considered equally useful for all arthroscopists or will it be considered more useful for novice surgeons? Will surgeons favourably value its quality and degree of realism? Will they validate the exercises included in the simulator? What aspects will surgeons consider important when assessing a trainee?

If we examine our questions, we see that they want to evaluate the subjective opinion of orthopaedic surgeons. The raised questions will be related to face and content validity. Therefore, the evaluation study will focus on face and content validities researching on the surgeons' opinions.

With this aim we posed two hypotheses:

- Surgeons will consider the haptic simulator as a useful learning method for acquiring arthroscopic skills
- Arthroscopists will favourably value the simulator as well as its quality, its realism and the exercises that are included

This evaluation study has a descriptive purpose (so the guidelines specific to experimental studies do not apply to this case).

The intervention protocol was as follows:

- All subjects received an explanation of the purpose and characteristics of the study and signed the informed consent.
- We assigned a number to each subject to keep their identity anonymous.
- We showed the virtual reality arthroscopy simulator and we explained its operation, functionality, instructions for use, and the exercises contained.
- Participants used the simulator doing exercises in which they had to handle the simulated surgical instruments. They had to perform a diagnostic arthroscopy, feeling the haptic feedback when palpating different anatomical targets.
- After using the simulator, we asked the arthroscopists to answer an anonymous questionnaire with questions about face and content validities.

The observation protocol in this study was just to help with doubts, to supervise how the subjects experienced the virtual reality simulation, and to gather the questionnaires with the variables we wanted to analyze.

Considering face validity, we asked the surgeons about the utility of the simulator for surgeons with and without previous experience. We inquired to what extent the surgeons think that the simulator will be useful for learning skills. They valued as well the quality, organization of the platform, realism and the variety of the exercises. The surgeons gave a global mark to the virtual reality simulator. We asked them if they would recommend the arthroscopy simulator to other colleagues and how long would they practice in case the simulator was at their disposal all the time.

Regarding content validity, we asked about the exercises included. In order to gather information for future evaluation experiments that will test other types of validities; we asked which measures the surgeons considered important when evaluating arthroscopic skills. We asked them to value different metrics and their importance, and to suggest new measures that they would consider relevant.

We carried out our experiment during a National Conference on Orthopaedic Surgery and Traumatology, in which 19 specialists participated.

The characteristics of the subjects are shown in Table 2. It is remarkable the high proportion of ambidextrous people. A future study could be carried out to determine if there is a correlation between ambidextrous subjects and their experience in arthroscopy.

Variable	Description
Sex	Female: 26.3 %; Male: 73.7 %
Age	32.3 (Std. Dev. 6.95)
Ambidextrous	15.8 %
Surgery	Yes: 63.2 %
	More than five years of surgery: 27.3 %

Table 2. Characteristics of the participants into the evaluation study

In order to test face validity, the surgeons answered about the utility of the simulator as a tool for learning skills. Surgeons considered that the simulator was useful for learning arthroscopic skills. A mean qualification of 8.67 out of 10 with a 95% confidence interval (7.81 ; 9.52) was obtained. In fact, 73.8% of the surgeons gave a mark equal or higher than 7. From the results regarding the utility for learning skills, we can affirm that the simulator showed face validity.

We compared the simulator utility for non-expert and expert surgeons, obtaining a mean difference of 2.53, with a 95% confidence interval (1.32 ; 3.75). This difference is statistically significant (P-value < 0.001), which means that, in general, the simulator is considered more useful for inexperienced than for expert surgeons (see Table 3).

Variable	Mean (Std. Dev.)	95 % C.I.
Utility of the simulator	8.67 (1.54)	(7.81 ; 9.52)
For beginner surgeons	8.80 (1.37)	(8.04 ; 9.56)
For surgeons with experience	6.27 (1.91)	(5.21 ; 7.32)
Means difference	2.53 (2.20)	(1.32 ; 3.75)
Global score	8.47 (0.83)	(8.00 ; 8.93)

Table 3. Utility of the simulator (measures out of 10, 10 being the highest mark). The simulator is considered more useful for beginner surgeons (P-value < 0.001).

Surgeons replied about how much time they would practise with the simulator if it were at their disposal full time. While only one subject marked less than one hour per week, the 26.3% of the surgeons said they would train more than five hours per week.

Regarding if the participants would recommend the simulator, there were 5 subjects that did not answer the question. The rest of the surgeons (73.7%) affirmed that they would recommend the simulator to their colleagues. Once again, these results support the face validity of the simulator.

Surgeons were asked to measure the quality and organization of the platform. They could answer: *poor, mediocre, normal, good, or excellent*. The lowest mark was *good* (57.9% of the surgeons), while 21.1% answered *excellent*.

Regarding content validity, we asked about the quantity, variety and quality of the exercises included in the simulator. The results are favourable for content validity: 21.1% answered *normal*, while 47.4% choose *good* and 10.5% *excellent*. Nobody marked *poor* or *mediocre*.

Concerning the time for training with the simulator, 42.1% considered it *appropriate* whereas 36.9% thought that it was *short* or *insufficient*.

The simulator obtained a very good global mark. It was valued 8.47 out of 10, with a standard deviation of 0.83 and a closed 95% confidence interval (8.00 ; 8.93). No surgeon qualified the simulator lower than 7 out of 10.

As mentioned, this study also wanted to elucidate which aspects are important when evaluating a surgeon in formation. The results about metrics to measure competence will be used for subsequent studies.

We observed how the surgeons value different aspects for arthroscopic skills learning. Features were marked according to their relevancy. Results are gathered on Table 4.

Variable	Mean (Std. Dev.)	95 % C.I.
Anatomical knowledge	9.36 (1.34)	(8.59 ; 10.13)
Pathological knowledge	8.36 (2.34)	(7.01 ; 9.71)
Having performed complete procedures	8.43 (2.31)	(7.09 ; 9.76)
Knot	7.93 (2.54)	(6.45 ; 9.40)
Surgical dexterity	8.15 (1.46)	(7.27 ; 9.04)
Hand-eye Coordination	9.15 (0.80)	(8.67 ; 9.64)
Manual precision	8.43 (1.56)	(7.53 ; 9.33)
Steady hand	7.69 (1.80)	(6.61 ; 8.78)
To overcome failures and difficulties	8.23 (2.83)	(6.52 ; 9.94)

Table 4. Different measures of competence and their importance (measures out of 10, 10 being the highest importance).

Mean comparison was made between the different pairs of skills. There was 1,69 of statistically significant mean difference (P-value = 0.004) between *Anatomical knowledge* and *Steady hand.* Also 1.23 statistically significant mean difference (P-value = 0.04) between *Anatomical knowledge* and *Surgical dexterity.* There was 1.46 of statistically significant mean difference (P-value = 0.006) between *Hand-eye coordination* and *Steady hand.* Also 1.00 statistically significant mean difference (P-value = 0.009) between *Hand-eye coordination* and *Surgical dexterity.* There was 0.69 of statistically significant mean difference (P-value = 0.044) between *Manual precision* and *Steady hand.*

From the obtained results we conclude that the two most important skills from the surgeons' point of view are anatomical knowledge and hand-eye coordination.

<u>Discussion of the results of the evaluation study</u>

The evaluation methodology proposed includes the design of analytical experimental longitudinal prospective studies. Following this methodology, the first steps are to be able to establish face and content validities, which has been the purpose of the descriptive study.

Regarding the first hypothesis "Surgeons will consider the haptic simulator as a useful learning method for acquiring arthroscopic skills", the mean value was 8.67 out of 10 (7.81 ; 9.52). Then, we can conclude that the collected data supports the first hypothesis.

Concerning the second hypothesis: "Arthroscopists will favourably value the simulator as well as its quality, its realism and the exercises that are included", nobody marked *poor* or *mediocre;* most of the subjects (57.9%) qualified them as good or excellent. The simulator's global mark was 8.47 out of 10 (8.00 ; 8.93). These results support our second hypothesis.

Future research will take profit of these results in order to design new evaluation studies. Construct, concurrent and predictive validities should also be demonstrated before considering the simulator as a fully tested validated tool capable of evaluating the surgeons' proficiency level.

INTERSECTORIAL CONSORTIUM

So far we have highlighted the necessity of building an interdisciplinary team, following a design methodology, and designing evaluation studies in a structured way to test all the types of validities so that it can prove validity, reliability and transfer of skills to the real settings.

However, there is still a key aspect that will highly influence the success and lifetime of a VR simulator. We recommend building an intersectorial consortium with agents from the academic, healthcare and industrial sectors.

The presence of industrial partners is essential for ensuring consolidation; additional funding; the long-term sustainability of research lines, and for guaranteeing that simulators once validated can be widely available in clinics and hospitals.

CONCLUSIONS

This paper offers a global approach to the process of designing, implementing and evaluating a VR simulator. It highlights essential aspects to increase the successfulness of creating a VR simulator. Forming an interdisciplinary teamwork is essential. During the design and evaluation processes, disciplines have to communicate and collaborate harmoniously.

The article presents two methodologies to provide structured guidelines and facilitate understanding the issues specific to each discipline, so that the whole team contribute to the ultimate objective. The design methodology is presented taking into account the participation of the different disciplines, which will facilitate the communication and the specification of the different roles. The evaluation methodology offers a framework to design evaluation studies so that a VR simulator can prove its validity, reliability and transfer of skills to the real settings.

Since surgical simulation is a niche area clearly identified, the aim is to develop complete surgical simulators that are integrated into the surgeons' educational curriculum; thus we consider essential to close the intersectorial triangle: Academy – Medical practitioner – Industry, fostering interaction between the sectors for a common success.

FUTURE WORK

Further qualitative and quantitative research still needs to be carried out on the design and implementation and performance measures to make VR simulators considered as a reliable and beneficial tool for training and assessment of surgical skills.

With regards to the evaluation methodology, we are conducting further experimental designs to assess the construct, concurrent and predictive validity of the mentioned VR arthroscopy surgical simulator.

ACKNOWLEDGMENTS

This work was supported by grants from the Spanish Ministry of Education, Science and Innovation (TIN2010-21289 and the Cajal Blue Brain Project). S.B. thanks the FP7 Marie Curie Actions for the Intra European Fellowship (grant PIEF-GA-2009-236642).

REFERENCES

[1] Vygotsky, L., and Cole, M., 1978 "Mind in society: the development of higher psychological processes". Cambridge: Harvard University Press

[2] Aggarwal, R., Crochet, P., Dias, A., Misra, A., Ziprin, P., and Darzi A., 2009, "Development of a virtual reality

training curriculum for laparoscopic cholecystectomy" The British journal of surgery, 96(9): pp. 1086-1093.

[3] Lamata, P., Gomez, E., Bello, F., Kneebone, R., Aggarwal, R., and Lamata, F., 2006, "Conceptual Framework for Laparoscopic VR Simulators", IEEE Computer Graphics and Applications, 26(6): pp. 69-79.

[4] Jonassen, D.H., Hannum, W.H., and Tessmer, M., 1989, "Handbook of task analysis procedures", New York: Praeger.

[5] Gipps, C., 1994, "Beyond testing: towards a theory of educational assessment" London ; Falmer Press.

[6] Burdea, G., and Coiffet, P., 2003, "Virtual reality technology", Hoboken, N.J.: Wiley-Interscience.

[7] Kockara, S., Halic, T., Iqbal, K., Bayrak, C., and Rowe, R., 2007, "Collision detection: A survey", IEEE International Conference on Systems, Man and Cybernetics, pp. 4046-4051.

[8] Jiménez, P., 2001, "3D collision detection: a survey", Computers & Graphics;25(2): pp. 269-285.

[9] Nealen, A., Müller, M., Keiser, R., Boxerman, E., and Carlson, M., 2006, "Physically Based Deformable Models in Computer Graphics", Computer Graphics Forum, 25(4): pp. 809-836.

[10] Bayona, S., Espadero, J.M., Fernandez, J.M., Pastor, L., and Rodriguez, A., 2010, "Implementing Virtual Reality in the Healthcare Sector". In: Rao R. (ed.) Virtual Technologies for Business and Industrial Applications: Innovative and Synergistic Approaches. Business Science Reference, pp. 350-138-162.

[11] Sutherland, L.M., Middleton, P.F., Anthony, A., Hamdorf, J., Cregan, P., Scott, D. et al., 2006, "Surgical simulation: a systematic review", Annals of Surgery, 243(3): pp. 291-300.

[12] Bayona, S., Fernandez-Arroyo, J.M., Bayona, P., and Pastor, L., 2009, "A new assessment methodology for virtual reality surgical simulators", Computer Animation and Virtual Worlds, 20(1): pp. 39 - 52.

[13] Berg, D., Raugi, G., Gladstone, H., Berkley, J., Weghorst, S., Ganter, M. et al., 2001, "Virtual reality simulators for dermatologic surgery: measuring their validity as a teaching tool" Dermatologic surgery, 27(4): pp. 370-374.

[14] Paisley, A.M., Baldwin, P.J., and Paterson-Brown , S., 2001, "Validity of surgical simulation for the assessment of operative skill", The British journal of surgery , 88(11): pp. 1525-1532.

[15] Moorthy, K., Munz, Y., Sarker, S.K., and Darzi, A., 2003, "Objective assessment of technical skills in surgery" British Medical Journal, 327(7422): pp. 1032-1037.

[16] Feldman, L.S., Sherman, V., and Fried, G.M., 2004, "Using simulators to assess laparoscopic competence: ready for widespread use?", Surgery, 135(1): pp. 28-42.

[17] Sevdalis, N., Lyons, M., Healey, A.N., Undre, S., Darzi, A., and Vincent, C.A., 2009, "Observational teamwork assessment for surgery: construct validation with expert versus novice raters", Annals of Surgery, 249(6): pp. 1047-1051.

[18] Anastasi, A., and Urbina, S.P., 1997, "Psychological testing", Prentice Hall.

[19] Michelson, J.D., 2006, "Simulation in Orthopaedic Education: An Overview of Theory and Practice", The Journal of Bone and Joint Surgery, 88(6): pp. 1405-1411.

[20] GMV, S.A., 2010, "*insightArthroVR*®. Virtual Reality Arthroscopy Simulator", http://www.insightarthrovr.com/index_en.htm.