EuroRV3: EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization (2016)
K. Lawonn, M. Hlawitschka, and P. Rosenthal (Editors)

# Experiences on Validation of Multi-Component System Simulations for Medical Training Applications

Yuen C. Law, Benjamin Weyers and Torsten W. Kuhlen

Visual Computing Institute, RWTH Aachen University
JARA – High-Performance Computing

**Abstract**

*In the simulation of multi-component systems, we often encounter a problem with a lack of ground-truth data. This situation makes the validation of our simulation methods and models a difficult task. In this work we present a guideline to design validation methodologies that can be applied to the validation of multi-component simulations that lack of ground-truth data. Additionally we present an example applied to an Ultrasound Image Simulation for medical training and give an overview of the considerations made and the results for each of the validation methods. With these guidelines we expect to obtain more comparable and reproducible validation results from which other similar work can benefit.*

Categories and Subject Descriptors (according to ACM CCS): I.6.4 [Simulation and Modeling]: Model Validation and Analysis —

## 1. Introduction

Validation of simulation results is important to reveal the weaknesses and limitations of the applied simulation models and method, so that this can be improved. In this work, we use validation in two categories: *method validation*, which refers to the validation of the simulation methods, models and assumptions and their accuracy with respect to the ground truth; and *use-case validation*, which refers to the assessment of the degree in which the simulation results fulfill their intended purpose.

In a wide sense, we define the ground truth of a system as the set of measurable and reproducible inputs and system intrinsic processes. Having access to the ground truth is an ideal case scenario, where the results of the simulation can be directly compared against the products of the real system; alternatively, the partial results of each simulation component can be individually validated if the partial results of the system's components are known. However, when the system's ground truth is not available, method validation of the simulation represents a challenge.This gets even worse in case of the validation of multi-component simulations which are simulations composed of various parts each comprising of an individual simulation and model that interact with each other. Here, not only the single components have to be validated but also their interaction. Figure 1 shows systems *I, II* and *III* with components *A* and *B* and output *C* (shown in purple), and their respective simulation with components *X* and *Y* and output *Z* (shown in green). For system *I*, it is assumed that we can obtain quantifiable results from all of its components (*ground truth*), which allows for a 1:1 validation of the simulation components and the simulation models used to implement them. However, in system *II*, the results of all or some of the components cannot be measured or the underlying phenomena is not well understood; some simulation components might even be missing or too simplified for a meaningful validation, which obstructs the validation of individual simulation models and thus of the simulation results. System *III* represents a hybrid case where one pair of components were identified and one pair does not map.

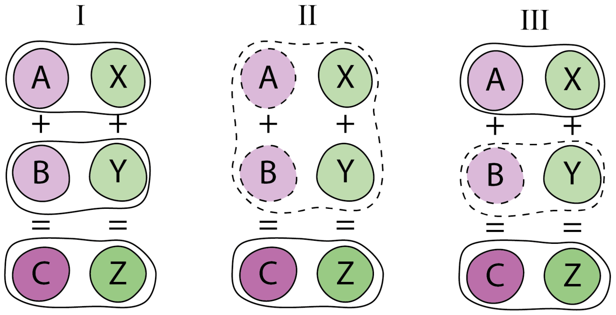

**Figure 1:** *Validating a well-known system simulation with a ground truth versus validating a system without a defined ground truth. Purple components represent the system, green ones represent the simulation. The circles around the components represent the mapping between system and simulation. The dotted lines represent uncertainty and lack of measurements that lead to a missing ground truth.*

DOI: 10.2312/eurorv3.20161113

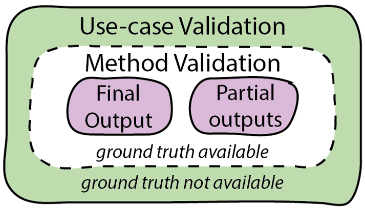

**Figure 2:** *Layers of validation for multi-component systems. Method validation is possible when ground-truth data is available, if not, we must shift to use-case validation.*

Note that in neither case, a simulation component has to necessarily be mapped to exactly one system component which leads to missing ground truth for the validation of a simulation component. Indeed, this missing mapping is often the case, rendering the validation process even more complex.

In this work, we will focus on validation methods for systems similar to System *II* and *III*, which are often the most difficult to validate since there is limited or no data and ground truth against to which objectively compare the results. In this cases, as illustrated in Figure 2, we shift the focus of the validation from evaluating the simulation method (inner white layer) to a use-case oriented validation (outer green layer). In other words, we will focus our attention on how well the simulation results fulfill their purpose and design our validation accordingly. Clearly, the validation metrics will depend on the purpose of the simulation. For use-case validation, the following options are possible:

1. Perform component-wise (partial) validation, against other simulations
2. Gather the opinion of experts in the field to validate the end results (face validation)
3. Conduct user studies in specific use-case scenarios

We will go over these options and use an Ultrasound Image Simulation (UIS), modeled as a multi-component simulation, as a case study to exemplify each of them. The UIS is a system of type *III* where no ground truth exits for the various components, such that only a use-case validation is possible in this regard. We share our different experiences —good and bad—regarding our efforts to obtain objective, quantifiable and reproducible validation methods for our simulation results following the presented conceptual classification of the presented problem domain.

## 2. Related Work

As mentioned, before starting the process of designing a validation method for a simulation, it is necessary to define the purpose of the simulation results and the according metrics by which these will be validated. In the case of a UIS for medical training, regardless of the specific training scenario, a requirement that is often mentioned is: *Adequate or enough image realism*. This alone is however not enough to define meaningful metrics.

In computer graphics, the term photo-realism is often used as a standard for image realism [Fer03]. Photo-realistic images are created taking into consideration the limitations of the human eye and the image capturing and display processes. Under this standard, a photo-realistic image needs only to be as real as a photograph of the scene and not the scene itself. Another standard for realism, is functional realism, which measures how reliable is the information that the image provides to complete a certain task. For example, an assembly instruction booklet needs only to display the information to enable readers to recognize the corresponding parts and their orientation. While evaluating photo-realism is a matter of measuring accuracy, functional realism is a matter of the perception of target users. In [RLCW01] an experiment was conducted to measure the perception of visual realism. In the experiment, real and synthetic images of scenes with simple objects were used. Here, the simulation components matched the components in reality 1:1, which allowed them to determine which components (shadowing, textures, light sources) increased the overall realism of the images. It is however in many cases not possible to isolate the effects of individual components so easily to study.

In the specific case of ultrasound image simulation, researchers in the area tend to rely on expert opinions to validate the results of their methods, and although this is in any case important and useful input, the huge amount of variability in the experience and equipment used by physicians across clinics and countries makes it difficult, if not impossible, to generalize the validation results and to compare the results of different simulation methods to one another. Furthermore, existing approaches differ depending on the goals and focus of the simulation. Solutions created for training focus on performance and are satisfied with images that look plausible. For example, Kutter et al. [KSN09] and Reichl et al. [RPAS09] present similar approaches based on information from CT data. Their focus was on performance and the presented tests and results reflect this. However, the evaluation of the photo-realism was limited to visually comparing real US images with the simulated ones. In [KWN10], Karamalis et al. present a work with focus on photo-realism and quality of the simulation that models wave propagation via the Westervelt Partial Differential Equation and solves it explicitly. In their work a set of simulated images is presented to the user for a visual evaluation of the realism.

Throughout the development of the simulation framework and its components, we have proposed, in different stages, various validation methods depending on the specific component to be tested and focused more on the functional realism, rather than photo-realism, to follow a use-case oriented validation. The next section is a recount on the experience obtained on each of the methods used applied to the Ultrasound Image Simulation approach, presented in [LKHK12].

## 3. Validation

As mentioned in the introduction, we will review the validation methods that we can use when ground-truth data is not available. We will apply these to a concrete example, namely, an ultrasound image simulation for medical training, presented in [LKHK12]. Referring back to Figure 1, for an ultrasound imaging system and its simulation we roughly obtain the following components (shown in Figure 3, and numbered accordingly):

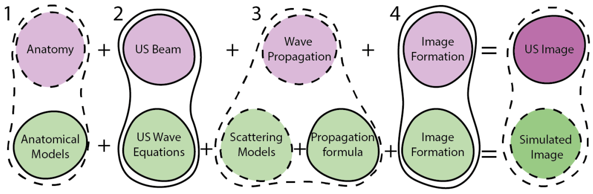

**Figure 3:** *Ultrasound imaging system and its components (purple), along with the corresponding simulation components (green).*

1. The anatomy: the shape of the structures and the acoustic properties of the various tissues
2. The ultrasound beam formation
3. The wave propagation and interaction with tissue (including scattering)
4. The image formation process (capturing, filtering and interpreting echo signals)

In this case, ground-truth data is difficult to obtain due to two main reasons. First, the characteristic noisy texture (speckle), part of component 3, present in ultrasound images is the result of a complex interaction between the ultrasound wave and particles spread throughout the tissue that scatter the wave. This interaction is not fully understood and cannot be modeled efficiently, which leads to incomplete or overly simplified simulation models. Second, anatomy and tissue properties (component 1) used to produce real ultrasound images cannot be exactly reproduced. Naturally, by removing these two components from the system and substituting them with, for example, artificial phantoms with homogeneous materials, we would obtain a simplified system for which a ground truth could be obtainable, as was done for example in [BBRH13]. This new system, however, has different outputs than the original and does not represent our target system.

In the following subsections, we will go over the three validation options mentioned in the introduction in examples applied to this concrete case, which suffers from the aforementioned lack of ground-truth data. After these, we will present a set of general guidelines to apply these validation methods in similar simulations.

### 3.1. Component-wise validation

The first validation option is to compare each of the components for which other simulation results are available. However, this comparison must be done critically, since measuring the difference between two approximation models does not tell us if one model is more accurate than the other with respect to the ground truth. Nonetheless, a qualitative comparison will reveal advantages and limitations of our simulation with respect to other approaches. We used this method for our ultrasound image simulation mainly to confirm that the models used simulate the ultrasound wave and its propagation presented the characteristics that were needed to reproduce desired effects in the resulting image, such as side lobes and focal area.

The proposed simulation approach used an analytic approximation of the beam profile combined with a geometrical acoustics approach to model the wave's propagation. We compared these to numerical FDTD simulation, which is a widely used and accepted model. As a side note, the numerical approach was not used for our simulation due to performance requirements. Figure 4 *Left* shows an example of a 2D focused beam calculated with the numerical solution. Comparing the result to the analytic solution (Figure 4 *Right*), differences are evident, for example, in the size of the cone of the main lobes, the angles in which the side lobes propagate, the intensity of the beam, and noise. However, it is also possible to observe that the profiles of both beams have similarities, too: both present a main lobe, two strong side lobes and some minor ones and an area of low intensity in the near field.

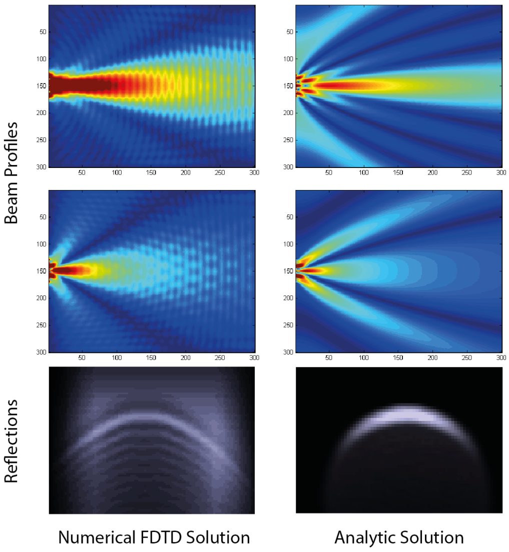

**Figure 4:** *Comparison of simulation models and results. Left: numerical approach; Right, analytic Approximation. Top: Beam Profiles; Bottom: Captured reflections.*

To model the beam's propagation and interaction with tissue, a geometrical acoustics [Vor08] approach was used, where rays are traced into the scene. Information from rays belonging to the same virtual transducer are combined to create one scanline, i.e. a vertical line of pixels scanned by one transducer. The resulting scanlines of all the transducers in the virtual probe compose the final image, emulating the actual image formation process. The results of this approach are again compared to the FDTD simulation, which includes all reflections and other propagation effects. Figure 4 *Left* shows a sequence of the 2D plot of the received echoes over time. Similarly, Figure 4 *Right* shows the intensities recorded in the geometrical acoustics approach. Differences are again clear, but a similar behavior of the main reflection is observable. Similarly, the models to produce the scattering textures was tested against known distribution models [LTJK14]. Here, we were interested in evaluating how well the histograms of the simulated textures matched against those of real images. To further validate the models, motion analysis algorithms designed for echocardiograms were applied to a sequence of simulated images using a heart phantom, with satisfactory results.

### 3.2. Face Validation

Face Validation considers the opinion of experts in the area and applies mainly for the cases when only the final output can be validated. Face validation is helpful in the initial stages of development since it can produce helpful insight on the main requirements of the simulation results. In medical simulation, face validity is often used to evaluate whether or not the simulator system behaves as expected; some examples can be found in [URK11] and [VHGJ08]. However, gathering the needed information is not an easy task. From our experience in consultations with the experts, it became evident that due to the lack of a common language, a method to more precisely communicate ideas and avoid misunderstandings was necessary.

For the specific case of the ultrasound simulation, we designed a method inspired by calibration tests used for head alignment of ink injection printers, where a series of similar images showing lines and squares are printed out on paper and the users are asked to select the best image of the sequence based on different criteria, for example, in which image are the vertical lines straighter [LUKK11]. In this case, different simulated images were generated where only one or two parameters was slightly changed, having an effect on image resolution, contrast and brightness, for example. By choosing the *most realistic* image of each set, experts indirectly calibrated simulation parameters without having to understand the more technical details. Figure 5 shows some samples of the simulated images. The top row shows some of the images that were rated as the best by the experts during the fine-tuning. The images at the bottom show the improved images after applying the suggested adjustments.

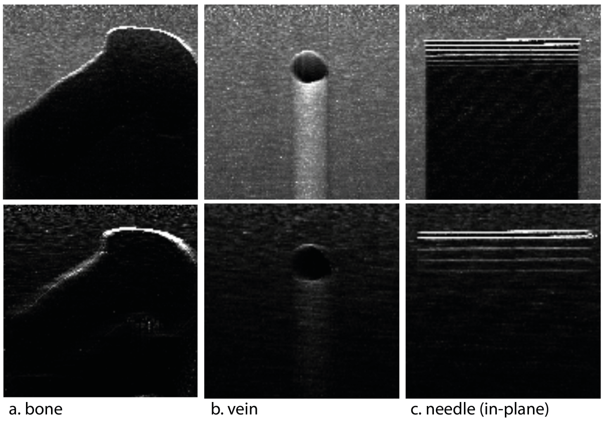

**Figure 5:** *Simulated ultrasound images:* Top: *Before calibration.* Bottom: *After calibration.*

This method requires large amounts of preparation time, however, it facilitates the communication process, can be done offline (e.g. via e-mail or online questionnaires), and can be easily tabulated and documented for later reference. Compared to an informal method, where experts gave feedback on the *best possible image*, the calibration method used yielded better results and in the long term, was less time-consuming.

### 3.3. User Studies

User studies can help assess the functional realism of simulation outputs and answer the question to whether or not the simulation is good enough to fulfill its objective, especially when photo-realism is not achievable. To apply the user study effectively, the use-case must be defined and the scope clearly delimited, thus inevitably we will lose generality in our findings. In the case of our ultrasound simulation, it was clear from the beginning that synthetic images were not going to reach the level of realism to look like real images, due to limitations in the anatomical models and performance.

A user study was performed to determine if the generated images were realistic enough to allow trainees to learn to recognize important structures in ultrasound; the results can be found in detail in [LKP*15]. The measurement of *Learning* was done twofold. First, we were interested in quantifying the knowledge the participants posses on ultrasound and anatomy before and after using the learning application. The second dimension was the users' perspective on their learning experience with the software, mainly, if they felt they were able to learn by using the application. The study was based on a 2 × 1 within-subjects design with a pre- and post-test methodology to observe two dependent variables: (a) the participants' ability to identify structures in simulated images and (b) their ability to identify structures in real ultrasound images. A test was applied to every participant at the beginning of the study. The exercises in this test were designed based on exercises found in US textbooks and on input from experts in the area who evaluated the difficulty and viability of the exercises. After the pre-test was finished, participants had a 20 minute session to use and explore the application. After this session, the post-test, which contained the same questions as the pre-test, was applied. Following the post-test, the participants were asked to fill the USE [Bro96] and SUS [Lun01] usability questionnaires.

## 4. Guidelines and Conclusions

From the case presented above, it is possible to abstract some general guidelines and recommendation to apply this methodology in similar systems. First, since we are performing use-case validation (refer to Figure 2), it is of course important to specify the exact metric by which the simulation will be validated. More specifically, the exact application and scope of the results must be clearly defined in order to design the corresponding validation tools. This will reduce the amount of variables to be tested, limit the scope and allow the definition of concrete evaluation goals. As we have seen, for our validation, each of the methods applied aimed to evaluate specific parts of the system and the validation process was designed accordingly. Second, a decomposition of the system in its smaller components will give more insight of which parts need validation and which parts cannot be validated with the available data. This will help to plan a comprehensive validation of the complete simulation, as opposed to only validating the final outputs. Finally, we must be aware that limiting the validation in the way that is suggested here, limits the results of the validation to the specific use-cases it was designed for. However, we consider this is a trade-off that must be done in order to obtain meaningful results.

## Acknowledgements

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 610425.

## References

[BBRH13] BURGER B., BETTINGHAUSEN S., RADLE M., HESSER J.: Real-Time GPU-based Ultrasound Simulation Using Deformable Mesh Models. *IEEE Transactions on Medical Imaging 32*, 3 (March 2013), 609–618. 3

[Bro96] BROOKE J.: SUS: A Quick and Dirty Usability Scale. *Usability Evaluation in Industry 189*, 194 (1996), 4–7. 4

[Fer03] FERWERDA J. A.: Three Varieties of Realism in Computer Graphics. *Electronic Imaging 2003* (2003), 290–297. 2

[KSN09] KUTTER O., SHAMS R., NAVAB N.: Visualization and GPU-accelerated Simulation of Medical Ultrasound from CT Images. *Computer Methods and Programs in Biomedicine 94*, 3 (June 2009), 250–66. 2

[KWN10] KARAMALIS A., WEIN W., NAVAB N.: Fast Ultrasound Image Simulation Using the Westervelt Equation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI) 13*, Pt 1 (Jan. 2010), 243–50. 2

[LKHK12] LAW Y. C., KNOTT T., HENTSCHEL B., KUHLEN T.: Geometrical-Acoustics-based Ultrasound Image Simulation. In *Eurographics Workshop on Visual Computing for Biology and Medicine* (Norrköping, Sweden, September 2012). 2

[LKP*15] LAW Y. C., KNOTT T., PICK S., WEYERS B., KUHLEN T. W.: Simulation-based Ultrasound Training Supported by Annotations, Haptics and Linked Multimodal Views. In *Eurographics Workshop on Visual Computing for Biology and Medicine* (2015), Bühler K., Linsen L., John N. W., (Eds.), The Eurographics Association. 4

[LTJK14] LAW Y. C., TENBRINCK D., JIANG X., KUHLEN T.: Software Phantom with Realistic Speckle Modeling for Validation of Image Analysis Methods in Echocardiography. In *SPIE Medical Imaging* (2014), International Society for Optics and Photonics, pp. 90400C–90400C. 3

[LUKK11] LAW Y. C., ULLRICH S., KNOTT T., KUHLEN T.: Ultrasound Image Simulation with GPU-based Ray Tracing. In *Virtuelle und Erweiterte Realität, 8. Workshop der GI-Fachgruppe VR/AR* (Wedel, Germany, September 2011), pp. 183–194. 4

[Lun01] LUND A.: Measuring Usability with the USE Questionnaire. STC Usability SIG Newsletter. Retrieved 5/3/2009, from http://hcibib.org/perlman/question. cgi. 4

[RLCW01] RADEMACHER P., LENGYEL J., CUTRELL E., WHITTED T.: Measuring the Perception of Visual Realism in Images. *Rendering Techniques 2001* (2001), 235–247. 2

[RPAS09] REICHL T., PASSENGER J., ACOSTA O., SALVADO O.: Ultrasound Goes GPU: Real-Time Simulation Using CUDA. *Proceedings of SPIE* (2009), 726116–726116–10. 2

[URK11] ULLRICH S., RAUSCH D., KUHLEN T.: Bimanual Haptic Simulator for Medical Training: System Architecture and Performance Measurements. *Proceedings of the 17th Eurographics Conference on Virtual Environments & Third Joint Virtual Reality* (2011), 39–46. 4

[VHGJ08] VIDAL F., HEALEY A., GOULD D., JOHN N.: Simulation of Ultrasound Guided Needle Puncture Using Patient Specific Data with 3D Textures and Volume Haptics. *Computer Animation and Virtual Worlds 19*, 2 (2008), 111–127. 4

[Vor08] VORLÄNDER M.: *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*. RWTHedition (Berlin. Print). Springer, 2008. 3