Validation for Simulators: It's All About Perspective

Alyssa Tanaka, Danielle Julian, Patricia Mattingly, Roger Smith
Florida Hospital Nicholson Center
Celebration, FL
Alyssa.Tanaka@flhosp.org, Danielle.Julian@flhosp.org,
Patricia.Mattingly@flhosp.org, Roger.Smith@flhosp.org

ABSTRACT

Does "validity" refer to the quality of an assessment, reliability of simulator outputs, or accuracy of internal simulation models? This question emerges in medical simulation and training, as educational, clinical, and engineering communities intersect. Each has developed a validation approach to meet their needs, without clear understanding of the other perspectives. Historically, validity has been assessed using a classical framework of content, criterion, and construct validity, concluding that a simulator is or is not valid. Validity has evolved into a unitary concept of construct, consisting of five distinct sources: content, response process, internal structure, relation to other variables, and consequences. Evidence for each source supports a score interpretation for a specific population, under a specific use case. This does not indicate that the assessment itself is generally valid, much less whether the simulator can be relied upon to deliver accurate results.

This unitary framework was adopted by the American Psychological Association as the standard for validating assessments and was recently endorsed as the "gold standard" for validating training tools. While this framework is effective for evaluating the appropriateness of an assessment, it may not be as robust for evaluating a *simulation device* used for assessment. This framework does not account for the physical and functional requirements of a physical system and the implications that discrepancies in those aspects may have on training and assessment.

This paper compares the classical and unitary validity methodologies with a perspective on the application to training simulators, as well as examines the inherent limitations of both. Recommendations and industry standards from other fields are also examined for applicability to surgical simulation. Finally, a recommendation for the validity classification of surgical simulators is proposed. The future of surgical certification and licensing could be reliant on simulation, however validity standards must be established to support this goal.

ABOUT THE AUTHORS

Alyssa D.S. Tanaka, Ph.D., is a Simulation Scientist at Florida Hospital Nicholson Center. Her research work focuses on robotic surgery simulation and effective surgeon training. Her current projects include rapid prototyping of surgical education devices, the validation of a robotic surgical curriculum and evaluation of robotic simulation systems. She has earned a PhD and MS in Modeling and Simulation from the University of Central Florida, as well as a Graduate Simulation Certificate in Instructional Design, and a B.S. in Psychology and Cognitive Sciences from the University of Central Florida. She holds a diploma in robotic surgery from the Department of Surgery, University of Nancy, France.

Danielle Julian, M.S., is a Research Scientist at Florida Hospital's Nicholson Center. Her current research focuses on robotic surgery simulation and effective surgeon training. . Her current projects include rapid prototyping of surgical education devices and the evaluation of robotic simulation systems. Her background includes research in Human Factors and learning and training to enhance the higher-order cognitive skills of military personnel. She is currently a Ph.D. student in Modeling and Simulation at the University of Central Florida where she previously earned an M.S. in Modeling and Simulation, Graduate Simulation Certificate in Instructional Design, and a B.S. in Psychology.

Patricia Mattingly, M.D. is a research fellow at Florida Hospital's Nicholson Center. Her research focuses on robotic surgery simulation and effective surgeon training. Her current projects include the development of a robotic surgical curriculum and assessing surgical task-induced fatigue. She is an AAGL Fellow in Minimally Invasive Gynecologic Surgery with Columbia University Medical Center. She completed her residency in obstetrics and gynecology at Carolinas Health Care System in Charlotte, NC and earned her M.D. from the Medical College of Georgia. She received her B.S. in Biochemistry and Molecular Biology at the University of Georgia.

Roger Smith, Ph.D., is an expert in the development of simulation devices and training programs. He has spent 25 years creating leading-edge simulators for the Department of Defense and Intelligence agencies, as well as accredited methods for training with these devices. He is currently the Chief Technology Officer for the Florida Hospital Nicholson Center where he is responsible for establishing the technology strategy and leading research experiments. He has served as the CTO for the U.S. Army PEO for Simulation, Training and Instrumentation (PEO-STRI); VP and CTO for training systems at Titan Corp; and Vice President of Technology at BTG Inc. He holds a Ph.D. in Computer Science, a Doctorate in Management, and an M.S. in Statistics. He has published 3 professional textbooks on simulation, 10 book chapters, and over 100 journal and conference papers. His most recent book is *A CTO Thinks About Innovation*. He has served on the editorial boards of the *Transactions on Modeling and Computer Simulation* and the *Research Technology Management* journals.

Validation for Simulators: It's All About Perspective

Alyssa Tanaka, Danielle Julian, Patricia Mattingly, Roger Smith
Florida Hospital Nicholson Center
Celebration, FL
Alyssa.Tanaka@flhosp.org, Danielle.Julian@flhosp.org,
Patricia.Mattingly@flhosp.org, Roger.Smith@flhosp.org

INTRODUCTION

In simulation, many fields converge to create the specialized training tools used to provide learners with standardized environments for the safe acquisition of skills, relying on the expertise of engineers, educators, and subject-matter experts to create valuable training tools. It is imperative that these training systems are vetted to ensure that system performance meets the expected standards, a process typically referred to as validation. The resulting measure of validity refers to the degree to which a model or system is an accurate representation of the real world concept that it is intended to replicate (Sargent, 2000; McDougall, 2007; AERA, 1997).

The underlying validation process and associated implications are often subject to the field it is being referenced for. Using a flight simulator as an example, a computer programmer may validate the model in respect to how it performs against an actual system (e.g. aerodynamic characteristics). An engineer may assess whether the controls look and feel representative to the actual aircraft platform, and an educator validates that the flight assessment and After Action Review (AAR) accurately measure and provide relevant feedback on the trainee's performance for a specific testing context.

The surgical field has adopted virtual reality (VR) simulators, similar to flight simulators, as a solution to limited training opportunities, regulated work hours, and a need for advanced training (Kuhn, 1962; Gallagher & Sullivan, 2011). Similar to the validation of a flight simulator, each stakeholder involved in the development and implementation of a surgical simulator has a specific expectation for the concept of validity. The programmers are interested in how closely the physics models of the virtual environment are representative of the real world (e.g. how tissue behaves when retracted) and the engineers verify that the controls function similarly to the actual surgical instruments. The educators and researchers are more concerned with how the benchmarks and scoring system translate to the learners.

The introduction of VR simulators coincided with a drive in the surgical field to move away from the traditional apprenticeship model and towards proficiency-based training. This has been critically important particularly in the specialized field of robotic surgery. Currently, four VR robotic surgery simulators exist: the da Vinci Skills Simulator (dVSS) by Intuitive Surgical Inc., also known as the "Backpack Simulator"; the dV-Trainer from Mimic Technologies Inc., the RoSS by Simulated Surgical Sciences Inc., and the Robotix Mentor from Simbionix (Figure 1). While all of these systems attempt to replicate the controls, visual system, and console of the actual surgical robot, each has unique qualities in regards to software, hardware, and assessment methods.

In the dVSS, the trainee sits at and operates the simulated environment using the actual da Vinci surgeon console. The simulator is a custom computer, appended to the surgical console through the actual surgical data port. Using this simulator, users can train using the actual hardware they would use during surgery. The second is a standalone system that utilizes a graphic/gaming computer, connected to a custom desktop viewing and control device that replicates the hardware of the da Vinci surgeon's console. This system shares similar software with the dVSS, but does not require the use of any actual da Vinci hardware. The third is composed of a completely customized replica of the da Vinci surgeon's console. Internally the simulator contains a graphic computer, a 3D monitor, and commercial Omni Phantom haptic controllers (Smith, Truong, & Perez, 2014). The Robotix Mentor is a standalone system that uses custom hardware for the master controllers and Sony glasses for the 3D visual system (Robotix Mentor, n.d). These variations in hardware and software have resulted in many research studies attempting to validate these systems, as illustrated in a summary of these studies in Smith et al. (2015) and Stephanidis (2015).

Figure 1. Different aspects of surgical simulation

The validation studies that have been performed over the last decade have come at a time when medical education and assessment are shifting to new standards. Therefore, the interested educational communities have called for a shift away from the methods of previous studies and towards a new standard process. This discussion has revealed a distinct difference in the perspectives of different communities that are interested in the validation of simulators and of the educational outcomes they provide. In this paper, we present three dominant models for validation which may appear to be in conflict, but which actually represent the distinct needs of different communities, at different phases in a simulator's lifecycle. This paper also provides a process for integrating multiple validation methods for effectively assessing educational technology.

VALIDATION FRAMEWORKS

Multiple professional communities have developed validation frameworks that address their own needs to insure, measure, and certify the accuracy, realism, and assessments provided by a simulator. The work of each of these communities is just beginning to be known to members of the other communities, which is triggering both mild and vehement disagreements about the meaning, purpose, and methods of validation. Cultural and intellectual clashes of these types have occurred repeatedly in other areas of science and engineering. Those cases, as in this, are often fueled by a lack of understanding of the perspectives and needs of the conflicting communities.

In surgical simulation, several frameworks for proving validity have been proposed as the standard for validating educational technology. While the American Psychological Association (APA) endorses a "unitary" framework as the gold standard for validating assessment tools, this model alone does not account for the need to validate simulators from different perspectives in other fields. A shared understanding of all of the perspectives involved may eliminate much of the friction that is being generated in this area. The most prominent validation frameworks from three different communities is shown in Figure 2 and discussed below.

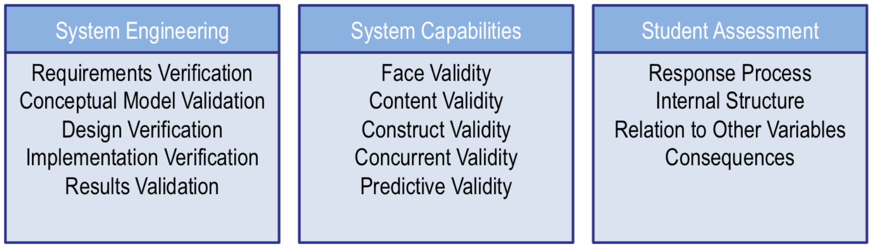


Figure 2. Summary of the validation frameworks

System Engineering Validation

The community that develops simulators and implements a formal process for validating their accuracy and usefulness has relied on Sargent's (2000) model for guidance through the engineering process, and indirectly the work of Balci (1997). In this model, the terms verification, validation, and accreditation (VV&A) are used to increase the preciseness of defining the steps in the process (Figure 3). However, this entire process is appropriately comparable to the other two frameworks that are explored in this paper.

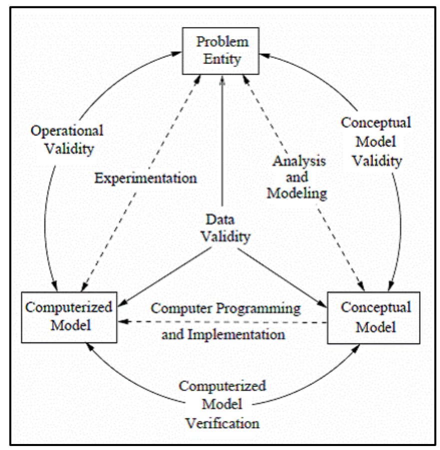


Figure 3. VV&A in Simulator Development (Sargent, 2000)

The creators and users of this framework are faced with a different set of problems than those who use of the other validation frameworks. Here, the emphasis is on guiding, controlling, modifying, and using a simulator as a hardware and software system or device. Because simulators are approximate replicas of some real world system, they can be created with dozens or hundreds of different representations of the world which may or may not be accurate and useful models of the real system and the purpose to which they are being put. This process seeks to expose the degree to which the simulator hardware, software, and data effectively represent the real world. This has to be done in the context of the expected application of the simulator. This context is essential in deciding whether compromises which have been made impact or invalidate the usefulness of the simulator in its specific application.

Sargent's framework has become the de facto validation process in the engineering and development of simulators. It is included in multiple later works which prescribe the process of simulator development and the accompanying validation of the product, such as Tolk (2012), Fishwick (2007), and others. In spite of this prevalence, the Sargent framework does not appear as a reference or an application in any of the medical simulation literature. Those communities come to simulation at a very different time in the system's lifecycle. They more typically encounter a simulator after it has been designed and manufactured for them by a device company. The users of the simulator are then more interested in the degree to which it can assist them with teaching concepts and measuring competence. So their need for validation is entirely at the user experience, educational effectiveness, and student assessment levels. In spite of the fact that the device company may have rigorously applied the VV&A methods of Sargent (2000) and Tolk (2012), the medical users will insist upon another layer of validation of the product using one of the other frameworks.

Classical Validation

To support the needs of communities using educational devices, to include simulators, the American Educational Research Association (AERA) and the American Psychological Association (APA) proposed a framework for assessing educational tools, typically referred to as the "classical" framework (AERA, 1985). The goal of this validity model is to assess educational tools to ensure that a tool is meeting the educational goals of assessing the specific abilities that it was intended to test.

Under this methodology, evidence is gathered to support a specific inference being made from test scores. For example, if a passing test score implies that a surgeon has the basic skills required to perform the removal of a prostate, then evidence would need to be gathered to support this claim. Under this framework, evidence is grouped into three categories: content related, criterion related, and construct related (Table 1).

Table 1. Summary of the Classical Framework

Validity	Meaning	Example(s)
Construct	A measure indicating the degree to which a test assesses the construct that it is intended on measuring.	What is this test supposed to measure? What is this test actually measuring?
Content	A measure of the degree to which a test's content represents a defined universe or content domain.	What is the content that needs to be tested? Is the test content representative of the actual content? Does the response type and testing format match the universe?
Criterion	A measure of the degree to which the test scores are related to one or more outcome criteria.	Can the test scores accurately predict future performance in the real world? How accurately can the test predict criterion performance?

For *construct related evidence*, information is gathered to support that the test evaluates the specific characteristics of the quality being measured (i.e. does the test evaluate what it is designed to). The construct of interest is often ingrained in the test's conceptual framework and is specific to the construct's meaning, distinguishing it from other constructs and indicating how the measure should relate to other relevant variables. Gathering evidence in this domain may also involve evaluating aspects such as test format or administration, if these circumstances affect the test meaning and interpretation.

Content evidence should demonstrate the degree to which test items, tasks, or questions are representative of a specified universe or area of content, given a proposed use of the test. Gathering evidence in this domain implies determining the content that needs to be tested and determining if the test is representative of that specific content. This also includes evaluating if the testing format and response mechanism is appropriate for the content (e.g. How is a student being assessed for a test on manual skill as opposed to critical thinking). This type of evidence often relies on expert judgment to assess the relationship between the test and the defined universe, however observation in combination with expert input is acceptable. If a test is going to be used in a way that was not originally intended, the appropriateness of original domain definition needs to be evaluated for the new use.

Criterion evidence demonstrates that test scores are systematically related to one or more relevant outcome criteria. The relationship between test scores and criterion measures may be expressed in several ways, with the goal of determining the accuracy to which the outcome criterion performance can be predicted from scores on the test. In general, there are two designs for obtaining criterion related evidence: concurrent and predictive methods. A predictive study obtains information supporting the accuracy with which test data can be used to estimate future criterion performance. A concurrent study serves the same purpose, but it obtains prediction and criterion information simultaneously.

McDougall (2007) adapted this framework for applicability to medical simulators. Under this modified framework the validation types included face, content, construct, concurrent, and predictive validity. Face validity is typically assessed informally by users and indicates whether the simulator is an accurate representation of the actual system (i.e. the realism of the simulator). Content validity is the measure of the appropriateness of the system as a teaching modality. Experts who are knowledgeable about the device typically assess this via a formal evaluation. Construct validity is the ability of a simulator to measure what it is intended to measure. Often this is characterized by the simulator's ability to differentiate between users' experience level. Concurrent validity is the extent to which the simulator correlates with the "gold standard" for training and predictive validity is the extent to which the simulator can predict a user's future surgical performance. Collectively, concurrent and predictive validity are known as criterion validity and are used as measures of the simulator's ability to correlate trainee performance with their real life performance. Face and content validity are most effective in evaluating the ability of a simulator to train a surgeon; however construct, concurrent, and predictive validity are most useful for evaluating the effectiveness of a simulator to assess a trainee. The majority of literature surrounding the validity of surgical simulators uses these categories defined by McDougall.

Unitary Validation

The AERA and APA updated the classical framework to create a new methodology for validating educational tools, referred to as the "unitary" framework because it views validity as a unitary concept of five sources of evidence: content, response process, internal structure, relations to other variables, and consequences (Table 2). The more evidence collected, the stronger the validity argument is for the test for a specific interpretation, at any given time, for a specific population. Similar to the classic framework proposed by the AERA and APA in 1997, the assessment itself is not considered completely valid or invalid, but is more or less valid.

Table 2. Summary of the Unitary Validation

Validity	Meaning	Example(s)
Test Content	A measure of the degree to which the test's content aligns with the content domain and interpretation of scores.	Are the test items assessing the content and skills that they should?
Response Process	A measure of the degree to which the response mechanisms of the test represent the skills being tested.	Are test takers demonstrating the skills being assessed?
Internal Structure	A measure of the degree to which the format and interrelatedness of the test items aligns with the construct being measured.	Is the test organized as it should be?
Relation to Other Variables	A measure of the degree to which the scores are related to variables outside of the test.	Do the scores align with a test that is currently the gold standard?
Consequences	A measure of the potential consequences of administering the test.	Are the consequences of the test scores relevant to the test's validity?

Test content evidence refers literally to the content of the test being administered. For the purpose of this measure, "content" refers to the test items, to include the wording and formatting of the test, and procedures for administration and scoring. The evidence in this domain includes either a logical or empirical analysis of the adequacy to which the test content represents the content domain and of the relevance of the content domain to the proposed interpretation of test scores. For task-based assessments, as in the case of many simulators, test evaluators create a list of tasks required by the job via observation and advisement of a subject matter expert (SME). The SME judgment assesses the criticality and frequency related to the task performance.

Response process evidence is gathered using a theoretical or empirical analysis of the response processes of test takers, which provides evidence in respect to the appropriateness of the construct and the nature of response mechanism used by the test takers. For example, if a test assesses critical analysis and reasoning, it is important to determine whether examinees are using this skill for the given material. The evidence for this domain is typically generated from an analysis of individual responses, including feedback from test takers regarding their performance strategies or reasoning of responses. In the case of scores being generated by evaluators, evidence can be gathered from the evaluators by determining the extent to which the evaluators are consistent with the interpretation of scores.

Internal structure evidence indicates the degree to which the relationships among the test items comply with the interpretation of the test score. Evidence gathered for this domain would indicate if the items on the test support the assumptions of the inter-relatedness of the items. For example if all items on a test will form a comprehensive score, then the test items should be one-dimensional. Test items may imply several aspects of a construct being tested and evidence in this domain determines the extent to which the items' relationships align with the necessity of the test framework.

Evidence gathered in regards to the *relationship to other variables* assesses the relationship of the test score to variables that are external to the test. The external variables can include measures of criteria that the test is expected to predict and relationships to other test scores that are expected to be either convergent or discriminant (i.e. measuring the same or different constructs respectively). This evidence addresses questions about the degree to which these relationships are consistent with the construct underlying the proposed test interpretation.

Lastly, evidence regarding the *consequences* does not necessarily affect the test's validity, but helps to inform the process of assessing validity. Evidence in this domain determines if there is a consequence of administering the test and if this consequence is relevant to other domains of validity. A finding in this domain of validity is relevant to the validity of the test in general if it can be directly related to another source of validity.

SYMBIOTIC FRAMEWORKS

When applying these frameworks to a simulation system being used for education, we can see that there is not one individually that meets all requirements of a system. While assessment is an essential component of a learning experience, it is not the only aspect that a user relies on for feedback when using a simulation system. Simulators are complex devices that often rely on the replicated controls and interfaces with real-world systems, including user feedback mechanisms (e.g. haptic feedback or visual stimuli). These mechanisms enhance user experience and facilitate learning by providing formative feedback and developing user expectations on how the real-world system should perform. Some simulators, including robotic surgery simulators, provide summative feedback mechanisms to the user at the end of the simulation experience, which helps to reduce the need for a proctor during the trainings. Figure 4 provides a general example of how this information is presented to the user. This feedback is often given based on specific criteria and benchmarks that are relevant to the task that the user is performing.

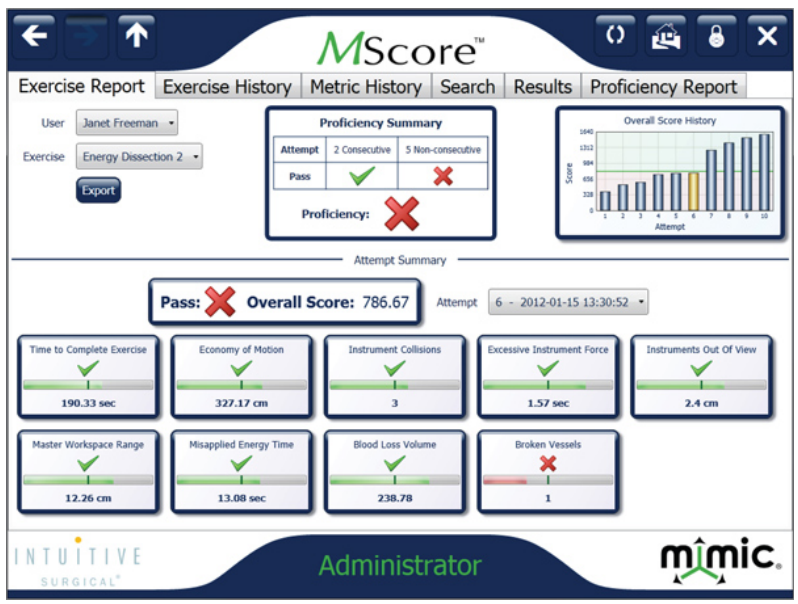


Figure 4. Robotic Surgery simulator summative feedback screen

During the simulation experience, the user makes an input into the system and receives a corresponding output from the system. For example, by moving a camera control towards a target area, the field-of-view will change to the specified location. By receiving that output the user decides what the next input will be. Using the camera example, if the user overcompensates and moves the camera past the target location, they would see this and use the camera control to adjust the field-of-view. This cycle continues until the simulation experience is complete (Figure 5).

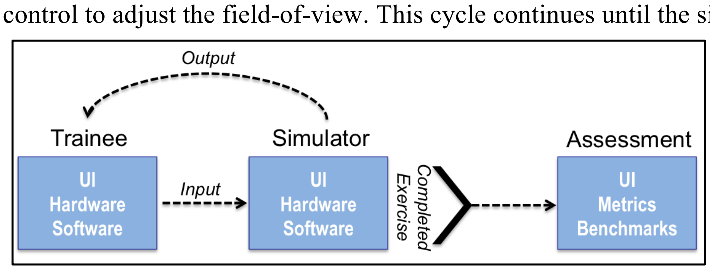
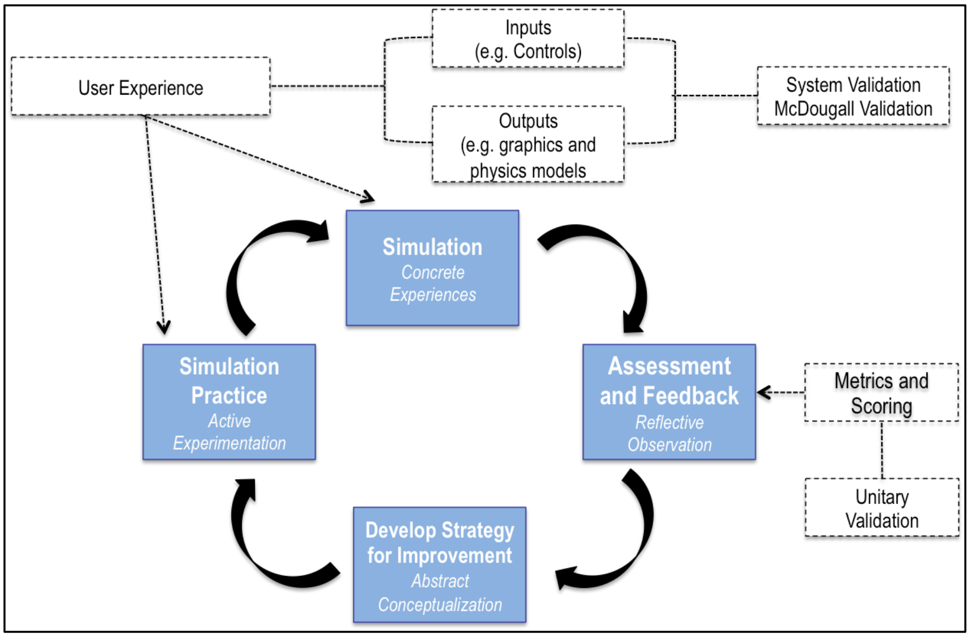


Figure 5. User interaction with simulator

The process of learning via simulation is an experiential process that can be related to the Kolb Experiential Cycle (1984) as shown in Figure 6. When looking at this model, the simulator plays a crucial role in the learning experience of the user. The user expectations are established during the *concrete experience*



with the simulator. The learner applies that experience for *reflective observation* and to form an *abstract conceptualization* of how to improve performance. Thus, the user's learning is facilitated through their interactions with the system and the formative feedback that they receive from system.

Figure 6. Image showing the relationship of the three frameworks

When looking specifically at the two educational models, the frameworks are designed for evaluating assessments and as such are focused on whether the assessment of the student was an accurate measure of the knowledge and skills that are being evaluated. If we only look at the assessment component of a simulator, then we are only looking at a small portion of the learning experience as a whole. It is possible to have a simulator that meets a high level of educational validity, but is not realistic in terms of engineering design. Conversely, we can have a simulator that almost perfectly replicates the intended system, but does not have meaningful associated metrics. In either case, the user would develop an incorrect model of their knowledge and skills during the training and assessment that would not translate to the real world system.

These frameworks cannot individually address the comprehensive needs for validation of educational simulators and thus need to be used complementarily to one another. Table 3 provides an example of different degrees of validity according to each framework which can be used to evaluate the individual simulator components and to address the needs of educators comprehensively.

Table 3. Validity Levels

	Less Validity	**Moderate Validity**	**More Validity**
Systems Engineering Framework	• Output does not match the real world measures.	• Unrealistic graphics • Pseudo-physics models.	• Highly realistic graphics • Realistic physics models.
Classical Framework (McDougall)	• Replicates real-world system to demonstrate placement of controls, but do not function the same.	• Custom hardware that is more realistic, but not exact.	• Embedded Simulator same hardware as in the real system.

Educational Framework	• Test content does not align with content domain. • Test does not measure what it is intended to.	• The content aligns with the content domain. • The users are not demonstrating the necessary skills	• Test content is relevant to the content domain. • Scores can predict future performance

CONCLUSION

This paper summarizes three prominent and valuable frameworks and demonstrates the role that each takes in the validation process. These frameworks overlap to some degree; no one framework is a complete duplication or replacement of another. Thus, the goal is to explain the rationale for the decidedly different processes that are referred to by the same term and create an awareness of these methodologies, potentially provoking adoption or adaptation. Understanding the value of different frameworks may reduce arguments and contention between communities attempting to apply their own perspective to other communities.

While valuable to specific fields, none of these validation models individually address the comprehensive needs when using simulation technologies as education and training tools. The learning experience when using a simulator encompasses components that should be evaluated distinctly to truly speak to the value of the system as an educational tool. Furthermore, disvaluing one aspect of the system during validation could have detrimental effects on the transfer of training for the user, potentially leading to negative training.

The field of simulation integrates technology, processes, and ideas from several different communities, using technology-rich learning environments to provide learners with a real-world experience for practice and assessment. To say that one method of validation alone is sufficient would be naïve. These frameworks were developed by their respective communities to address that community's specific needs, however needs of the broader simulation community require a more interdisciplinary approach.

It is imperative to critically evaluate not only about what the validation is used for, but also what the validation is evaluating and leverage the qualities of each of the validation frameworks when assessing the validity of a system. We must consider the role that each framework plays in a system and how that affects the learner.

REFERENCES

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Amer Psychological Assn.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Amer Educational Research Assn.

Balci, O. (1997). Principles of Simulation Model Validation, Verification, and Testing. *Transactions of the Society for Computer Simulation*. 14, 3-12.

Fishwick, P.A. editor. (2007). *Handbook of Dynamic System Modeling*. Boca Raton, FL: CRC Press.

Kolb, D.A. (1984). *Experiential learning: Experience as the source of learning and development* (Vol. 1). Englewood Cliffs, NJ: Prentice-Hall.

McDougall, E.M. (2007). Validation of Surgical Simulators. *Journal of Endourology*, 21(3), 244-7.

Robotix Mentor (n.d.). Simbionix website. Retrieved from http://simbionix.com/simulators/robotix-mentor/

Sargent, R.G. (2000). Verification, Validation, and Accreditation of Simulation Models. In *Proceedings of the Winter Simulation Conference*. Piscataway, NJ: IEEE Computer Press, 106-14.

Smith, R.D., Truong, M., Perez, M. (2015). Comparative analysis of the functionality of simulators of the da Vinci surgical robot. Surgical Endoscopy, 29(4), 972-83.

Stephanidis, D. et al. (2015). Simulation in Surgery: What's Needed Next? *Annals of Surgery*, 261(5), 846-53.

Tolk A. editor. (2012). *Engineering Principles of Combat Modeling and Distributed Simulation*. New York: John Wiley & Sons.