Theory on Simulator Validation

8

Jamie Y. Ferguson, Abtin Alvand, Andrew J. Price, and Jonathan L. Rees

Not everything that counts can be measured, not everything that can be measured counts. – Albert Einstein

Take-Home Messages

- Validation of surgical simulators is a key prerequisite for developing simulation-based surgical education and ensures that teaching and assessment methods are scientifically robust.
- Validation is not a binary concept but involves gathering evidence that a preconceived "construct" holds true in a given context.
- All assessment involves compromise. It is important to understand where these compromises can be made and where they should not.
- The most important aspect of validity is the hardest to measure; that simulation impacts clinical performance and results in improved patient outcomes.

J.Y. Ferguson MB ChB (Hons), MRCS (Ed), MEd (✉)
J.L. Rees, MD
Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK
e-mail: jamieferguson@doctors.org.uk; jonathan.rees@ndorms.ox.ac.uk

A. Alvand, BSc(Hons), MBBS, MRCS(Eng)
A.J. Price, MD
Nuffield Department of Orthopaedic Surgery, Nuffield Orthopaedic Centre, Oxford, UK
e-mail: abtin.alvand@ndorms.ox.ac.uk; andrew.price@ndorms.ox.ac.uk

8.1 Importance of Developing Simulation-Based Surgical Education

It is argued that traditional apprenticeship training lacks objectivity in the assessment of operative ability (Darzi et al. 1999). The implementation of standardized curricula aims to ensure all trainees achieve critical competencies and so the role of simulation is becoming more important (Motola et al. 2013). It is imperative, however, that any simulation models developed provide a fair reflection of the tasks that they are designed to replicate and their use genuinely improves the relevant skill domains they are aimed at improving.

In this chapter, we will introduce some of the important concepts in ensuring simulation is valid and useful. We will look at some of the theory underpinning how new simulation technologies can be evaluated to ensure they deliver in their intended applications specifically within arthroscopic training.

Given the high costs associated with introducing simulation-based technologies into training curricula, the process of validation is an important one as it attempts to establish whether or not the intended simulators are able to deliver on some of their claims. Several concepts need to be understood including validity and reliability.

M. Karahan et al. (eds.), *Effective Training of Arthroscopic Skills*,
DOI 10.1007/978-3-662-44943-1_8, © ESSKA 2015

8.2 Validity

Validity is a fundamental property of a test or assessment tool and is concerned with whether or not it measures what it purports to measure (Gallagher et al. 2003). Validity is not a characteristic of the simulation model itself but of the theoretical framework (otherwise known as the "construct") used in the model's application (Aucar et al. 2005). In other words, validity is related to the way in which the simulation model is used, rather than being an inherent property of the simulator itself. Simulation-based training models could be used in many different ways. Examples include an aid to training, a way of assessing progress or as a high-stakes competency assessment. In all of these applications, the simulation tool may remain the same, but the construct is different because the way that the simulation is applied and interpreted varies (Clauser et al. 2008; Scalese and Hatala 2013).

It is a common misconception that once validity is proven for a simulation model, it acts as a blanket term, applying to all other possible applications of that simulator. Instead each particular application of the simulation model has a specific construct that relates to that particular application. Any changing in the way the simulation model is used will result in a change in the construct and as a result may not be supported by the previous validation process. When designing a simulation tool, it is important that a clear decision is made regarding its intended role or purpose. If a simulation tool is used within a different context or in a different way to that which it was first conceived, its validity must again be demonstrated with further testing (Sedlack 2011). Validity should not be thought of as a binary concept but as a spectrum. Rather than being merely present or absent, there are degrees of validity, determined by the weight of supporting evidence available for that test. Proving perfect validity for any test is probably unachievable in the real world. Validation studies aim to provide sufficient evidence to support the construct as providing a true measure of what is tested within a specific context.

Any confusion surrounding the concept of validity may be related to the many different definitions discussed in the literature. Despite the various terms described for validity, it is in fact a singular entity. The various types described refer to slightly different facets of the same single concept (Garden 2008).

In the classical model of validity, three principle components of validity were described, namely, content validity, criterion-orientated validity, and construct validity (Cronbach and Meehl 1955). Various other facets of validity are grouped under these three principle headings (as outlined in Table 8.1) (Carter et al. 2005; Garden 2008; Michelson 2006).

Table 8.1 Forms of validity

(1) *Content validity*	
Evidence that the items of the simulation reflect the domain being tested. Each content area that is related to the construct should be included	
(a) Face	Subjective impression by non-experts of how closely the simulation replicates the real environment
(b) Content	Ensuring the simulation covers all the important components of a task as determined by expert opinion
(2) *Criterion-orientated validity*	
The relationship of performance in the new simulation compared to other independent established measures of the ability in the domain of interest	
(a) Concurrent	Correlation with an independent measure of ability performed at the same time as the simulation
(b) Predictive	Ability of the simulation to predict future performance by correlation with future test score
(3) *Construct validity*	
The overarching concept supported by all other forms of validity. It is the degree to which the simulation measures the theoretical construct. In other words, does the simulation measure arthroscopic ability, or does it merely measure the ability to perform the simulated task	
(a) Discriminant	The ability of the simulation to discriminate between those with differing abilities (such as junior and senior trainees)
(b) Convergent	The ability of the simulation to not differentiate between individuals of similar ability

More recently, there has been growing dissatisfaction with these categories which some feel make arbitrary distinctions between different forms of validity that do not really exist. The more modern view of validity is that it is a unitary concept without differing forms. Contemporary authors have proposed that in psychometric testing, these three distinct themes should be subsumed into the more comprehensive overarching theme of *construct validity* (American Educational Research Association et al. 1999).

8.3 Face Validity

Face validity is increasingly sidelined within validation processes. It is a subjective measure of how closely a simulation resembles real life and is usually measured through questioning experts. This is often a basic prerequisite of designing simulation-based studies or tasks, and is not really a part of validity testing. As Downing and Haladyna note, "*the appearance of validity is not validity*," (Downing and Haladyna 2004). However, a high degree of face validity can positively influence the *acceptance* of simulation-based tasks by end users – especially among trainee surgeons.

8.4 Content Validity

This looks at the components of a test or simulation and ensures that all the appropriate areas are covered effectively and are relevant to the test. It ensures the steps within the task are thought out and linked. Often during a simulation's design phase, this process is performed using cognitive task analysis when an expert is asked to talk through a task so that the various steps can be noted down by the developers with the ultimate aim of their incorporation into the simulation scenario. This form of validity is also relatively subjective, often relying on expert opinion.

8.5 Construct Validity

This is the ability of a test to identify and measure the attributes of performance it is designed to measure such that it is able to differentiate between novices and experts. There can be no argument that a simulation task for knee arthroscopy that cannot distinguish between expert surgeons and junior trainees possesses little validity as an assessment tool. Furthermore, construct validity must be reassessed as new/further skill metrics are discovered, in order to ensure the model is a fair representation of what is being tested.

8.6 Concurrent Validity

This is achieved by using other measurements of ability and correlating them with the simulation. This process is often employed when introducing a new assessment tool so that it can be compared to the current gold standard assessment. An example might be linking motion analysis movement data (e.g., hand path length) with global rating scales (Alvand et al. 2013). High correlation between different assessment tools indicates good concurrent validity. This process of using multiple data to establish validity is often termed triangulation.

8.7 Discriminate Validity

This involves ensuring that there is no correlation between aspects of the test that should not correlate. In other words it confirms that unrelated parts of a test are in actual fact unrelated. In the context of simulation, it means that the parameters are able to differentiate between established experts and novices.

8.8 Convergent Validity

This is the counterpart to discriminate validity. This is the ability of a test to demonstrate that elements that should be related are related.

An example in simulation is the ability of a test to show that individuals of a similar skill level are grouped together appropriately.

8.9 Predictive Validity

This is the ability of a simulation/simulated task/simulator to predict actual performance in the real clinical setting from the simulated performance. This is probably the most important aspect of validity testing, but in reality little literature has looked at this in arthroscopic simulation and it is one of the most challenging aspects to prove (Hodgins and Veillette 2013; Slade Shantz et al. 2014). Long-term transferability studies are necessary for predictive validity to be established. Furthermore, a reliable way of assessing operative ability is required so as to compare performance in real-life settings with simulation performance. In addition, it is important to remember that technical ability is only one of a number of influences on patient outcome. Spencer stated that surgery is 75 % decision-making and 25 % dexterity (Spencer 1978). Daley and coworkers identified several other factors that contribute to the quality of surgical care including, leadership, which technology is used, the interface with other services and institutions, the level of the coordination of work, and how quality of care is monitored (Daley et al. 1997). Therefore, although it is highly desirable to link technical skill scores and clinical outcome, the large number of *nontechnical* factors that influence patient outcome make identifying a correlation very challenging.

8.10 Sources of validity evidence

As previously stated, validity is a unitary concept and the various aspects of validity discussed are not distinct types but different forms of evidence accumulated to support the intended interpretation of performance for the proposed purpose (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 1999). Evidence for a construct's validity can be gathered in five different domains outlined in Table 8.2 (American Educational Research Association, American Psychological

Table 8.2 The five elements of construct validity as outlined by Messick (Messick 1989; Messick 1995)

Types of evidence for validity (Messick 1989)	Description
Content	This is a measure of the extent to which a test's content assesses the skill domain that it purports to assess/measure. This involves ensuring that all the relevant aspects of the task assessed are included to avoid the problem of "underrepresentation" as well as avoiding the risk of "construct-irrelevance" (a situation where factors irrelevant to the construct are measured). This is usually achieved through expert opinion on the test contents
Response process	This ensures the fit between the construct and the performance of the test. For example, scores in a mathematical test of higher-order thinking should be different between those who actually use higher-order thinking and those who have simply memorized the answers. Ensuring this may involve asking test takers to "show their working" or demonstrate their thought process. It also encapsulates rater scoring, ensuring judgments are not made based on irrelevant factors, such as how the candidate is dressed
Internal structure	Scores that are intended to measure a single construct should deliver homogenous results where individuals with varying ability should attain scores that can allow discrimination between them. This is also used as a test to ensure reliability by testing internal consistency
Relation to other variables	Correlation with other instruments where observed relationships match with predicted relationships or a lack of correlation where it is not expected would support this. The instruments used, such as motion analysis or global rating scales, would also need to have been previously validated for use in this way

Table 8.2 (continued)

Types of evidence for validity (Messick 1989)	Description
Consequences of testing	These are the intended and unintended consequences of testing. For example, trainees may only concentrate on elements of the curriculum that are tested while neglecting other topics. Another example might be using a simulator for selection from an unrelated domain. If a flight simulator was used for selection into higher surgical training, this process may have questionable validity

Association, and National Council on Measurement in Education 1999; Cook and Beckman 2006; Downing 2003; Messick 1989).

8.11 Threats to Validity

There are two principle threats to validity that must be avoided, namely,

> *construct underrepresentation* and *construct-irrelevant variance*. (Messick 1995).

Construct underrepresentation refers to the degree to which the assessment fails to capture important aspects of the construct. This will have an impact on the score interpretations, as the evidence they are based on will be weak if important aspects of the construct are not tested. An example of this might be trying to use an isolated plastic synthetic bone model without soft tissue cover to test competence at performing open reduction internal fixation of a tibial plateau fracture. Although this model would be good at assessing procedural knowledge, not simulating the soft tissues overlying the bone would greatly reduce the validity of the task as the sole test of competence for this complex procedure.

Construct-irrelevant variance refers to the degree to which extraneous or irrelevant factors impact upon the test score. This may be systematic, such as from bias, or a result of the testing scenario being so broad that it incorporates elements irrelevant to the tested construct. This generates "noise" making the interpretation of the results more difficult. Poor design of the simulation instrument can make this problem worse if the performance of some users is improved by extraneous clues or prompts in the test format that are irrelevant to the construct or if some are disadvantaged for reasons outside the construct of interest.

8.12 Reliability

Reliability refers to the consistency or stability of measurement in a test (Kazdin 2003). It is the measure of the reproducibility of test scores obtained from an assessment given multiple times under the same conditions. All measurement has inherent variability, and the difference between a single measurement and the "true" measurement is termed the measurement error (Boulet and Murray 2012). All assessment involves taking a sample of an individual's knowledge or performance and making inferences about that data to reach a conclusion about the individual's true ability. The greater the difference between the assessment result and the individual's true ability, the less reliable the assessment. A reliable test gives a fair reflection of an individual's true ability.

The concepts of reliability and validity are intrinsically linked, and their relationship can be illustrated using the analogy of hitting archery targets (Fig. 8.1). Reliability is a necessary, but not sufficient component of validity (Cook and Beckman 2006). If the components of a test are unreliable, then conclusions cannot be drawn from the results, and the test is no longer valid. For example, if a new simulator is used to assess an experienced surgeon's operative ability and of four repetitions it rates his performance as "average," "very poor," "excellent," and "good," the test can be seen to lack reliability. Conversely, if the simulation result was consistently "poor," then although the test could be called reliable (due to the consistent results over multiple tests) it would lack validity, assuming that there was sufficient objective evidence that the surgeon really possessed expert surgical skills. Only when the test consistently rates his performance as excellent could the simulator be said to be both reliable and valid.

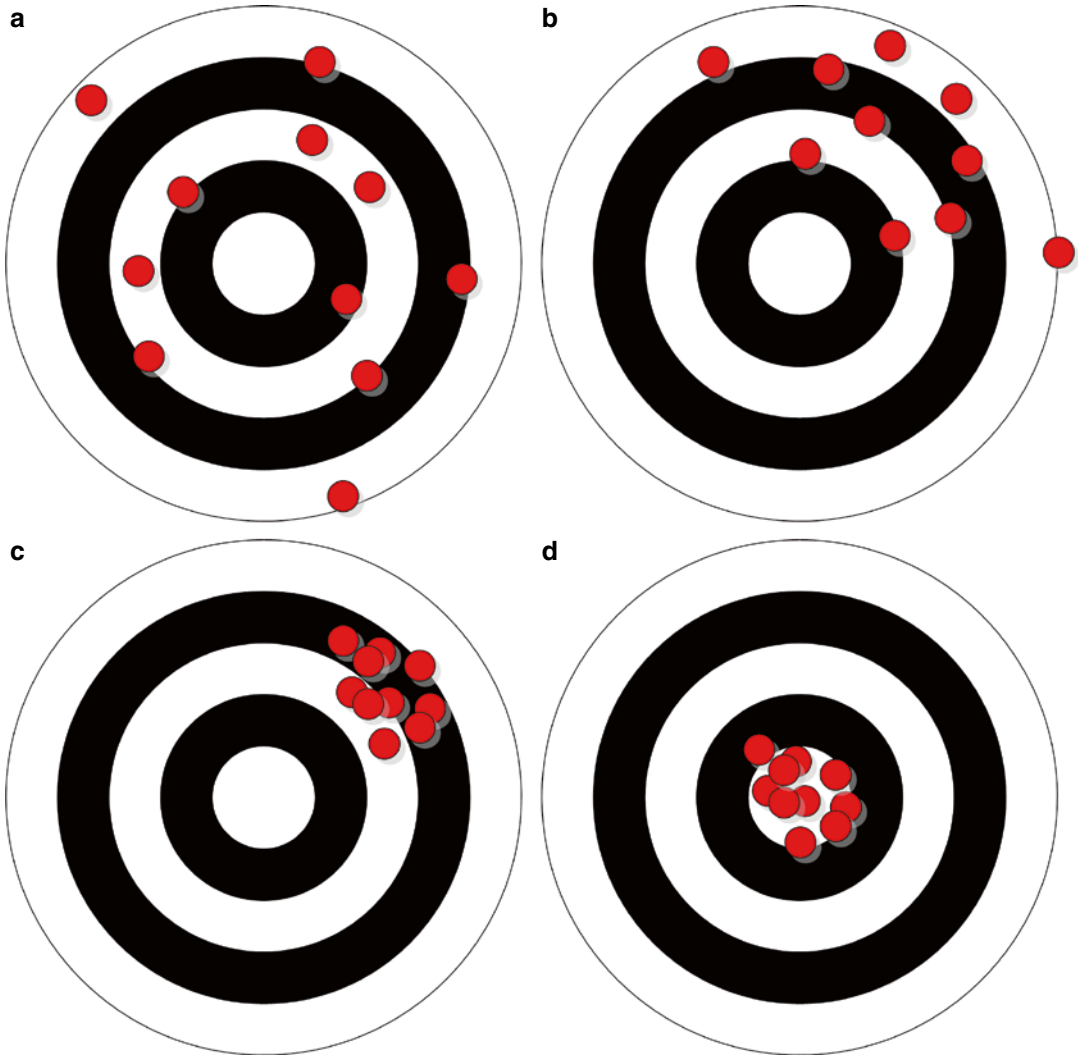

Fig. 8.1 Validity and reliability are intrinsically linked. Imagine each shot represents a test score and the bull's eye represents a candidate's true ability. (**a**) Shots centred around bull's eye but spread out, therefore valid but not reliable. (**b**) Shots not centred around Bull's eye and spread out, therefore not valid or reliable. (**c**) Reliable but not valid. (**d**) Valid and reliable

Assuming the measurement error is equally distributed, the reliability of an assessment should be improved by increasing the sampling (Downing 2004). This is because with sufficient repetition of assessment, the error should average towards zero. This can be achieved by making the test longer, increasing the number of different assessment parameters or by increasing the number of raters. The degree of measurement error impacts on how long a test must be to achieve adequate reliability and will therefore determine the values of any single measurement (Garden 2008).

In psychometric testing, it is often not practical to obtain multiple measurements of an individual to correct for high measurement errors. Therefore designing simulators with good reliability is important, particularly if they are to be used for assessment. This is especially true for high-stakes assessment (such as for licensing and certification assessment which are designed to protect real patients from incompetence) where the consequences of a false positive result may cause patient harm. Reliability can be measured in several ways as outlined in Table 8.3.

Table 8.3 The various aspects of reliability testing

Type of reliability evidence	Description
Test-retest	Otherwise known as intrasubject reliability, this measures if trainees achieve similar scores on two different occasions
Internal consistency	This is assessed by comparing the relationship between different elements of the test or simulation. Correlations can be measured between each item of the test, known as inter-item correlation, or by dividing the test into two parts and comparing them, known as split-half correlation. Poor correlation may suggest that more than one construct is being measured
Parallel forms	If the test items for the content of interest are randomly divided into two separate tests and administered to subjects at the same time, there should be strong correlation
Inter-rater	This test ensures that there is good agreement between assessors of a trainee's performance. Two forms exist: interobserver reliability measures the agreement between different assessors for a given test, whereas intraobserver reliability determines the variability of a single assessor's marks for the same test on different occasions

8.13 Statistically Measuring Reliability

8.13.1 Cronbach Alpha

The most common method of determining the reliability of an assessment tool is by use of the Cronbach Alpha (Cronbach 1951). This is a test of internal consistency, and it calculates the correlation between all the test items in all possible combinations. It can be expressed as

$$\alpha = \frac{n}{n-1}\left(1-\frac{\Sigma Vi}{V_{test}}\right)$$

where n is the number of test elements, Vi is a measure of the variance of the score on each test element, and V_{test} is the total variance of all scores on the whole assessment.

A shortcut for estabilising the degree of variance for each test item can be calculated using the following formula:

$$Vi = Pi \times (1 - Pi)$$

where Pi is the percentage of candidates who correctly perform the test element (expressed as a decimal). This will always give a number between 0-0.25.

Cronbach alpha generates a score between 0 and 1 to give a coefficient of internal consistency. The figure required will depend on the context of the assessment. For high-stakes tests, such as licensing exams, a figure of 0.9 or above is preferred, but for other forms of assessment, values of 0.7–0.8 may be acceptable (Downing 2004).

One of the strongest methods of improving reliability of a test is to lengthen the assessment by including more test items. This can be seen from the formula where the biggest impact on reliability is the V_{test} item because the larger the value, the higher the α score. For example, if we were to double the length of the assessment, the V_{test} will increase by a power of four because variance involves a squared term. In contrast the ΣVi will only double because each Vi is just a number between 0 and 0.25. As V_{test} increases faster than ΣVi, the alpha score will increase by virtue of lengthening the test. Therefore, it is important that any simulated task is of sufficient length to ensure reliability.

8.13.2 Standard Error of Measurement (SEM)

This is another less commonly used measure of reliability. It scores the degree of variance in candidate scores by the following formula (Harvill 1991):

$$\text{SEM} = \text{Standard Deviation} \times \sqrt{(1 - \text{reliability})}$$

It represents the standard deviation of an individual's scores (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 1999) and gives an indication of

the degree of certainty of the true score from the observed score. A confidence interval can be generated for the candidate's true score such that 95 % of an individual's retest scores should fall within 2 SEM of the true score (Harvill 1991).

8.13.3 Intrasubject Reliability

When using test-retest methods, the correlation between results is usually arrived upon using Pearson's R correlation test. This is generally a more conservative estimate of reliability than Cronbach alpha. However, the practicalities of using this technique are more challenging as it requires two separate sittings of the assessment. Other confounding factors resulting from changes in the conditions of the two assessments must be anticipated and controlled for, such as the potential learning effect from taking the initial test.

8.13.4 Inter-rater Reliability

Obtaining multiple scores is an important component of reliability. All too often scoring of performance is made based on single assessments or by single raters. It has been stated that "a person with one watch knows what time it is, a person with two watches is never quite sure" (Brennan 2010). This illustrates the potential difficulty of using multiple raters. However, increasing the number of assessors is one way on increasing reliability. The correlation between different raters can be measured in several different ways. The simplest is by assessing the percentage agreement between raters. The criticism of this method is that it does not take into account the possibility of agreement through chance alone. Cohen's kappa coefficient is a method for measuring agreement between two observers using a categorical assessment scale (Cohen 1960). It generates a value between -1 and 1 (although negative values are rarely generated and are of little significance in assessment validation). A value of 0 denotes no agreement, and 1 denotes perfect agreement. Figures above 0.6 suggest moderate agreement and above 0.8 suggest strong agreement (McHugh 2012). When comparing performance using an ordinal scale such as a Likert scale, a variation called the weighted kappa is used, which penalizes wider differences in scores between raters than narrower disagreements (Cohen 1968). If more than two raters are used, Fleiss' kappa should be employed (Fleiss 1971). The nonparametric Kendall tau test can be used if assessors use an assessment that involves ranking candidates or data (Cook and Beckman 2006; Sullivan 2011).

8.13.5 Generalizability Theory (G-Theory)

This is another more modern method of estimating reliability using factorial analysis of variance (Brennan 2010; Cook and Beckman 2006). It is able to look at the many sources of error within testing (termed facets) that influence the reliability of performance assessments. The impact of these various facets (such as item variance, rater variance, or subject variance) can be quantified, and the source and magnitude of the variability can be measured (Cook and Beckman 2006). This allows researchers to ask what factors have the greatest impact on reliability as well as helping to determine how to improve reliability though altering various error effects. For example, it may show that the greatest impact on reliability is the variation in inter-rater scoring. In this situation, this would tell us that the generalizability of the test across more observers is likely to be reduced.

8.14 Simulation Utility Involves Compromise

Van de Vleuten proposed that rather than thinking of factors such as reliability and validity in isolation, the most important overall measure of an instrument is its "utility" (Van der Vleuten 1996). This is a product of several different elements that all contribute to how useful it is in practice. As well as reliability and validity, these factors include educational impact, acceptability, and cost. In the real world, it is impossible to produce the perfect simulation due to the limitation of

resources such as time and cost. Consequently, all constructs require compromise to be feasible. Understanding this reality allows careful consideration of where the greatest compromise can be made which will depend upon what the main purpose of the simulation is envisaged to be. If the simulation is designed for a high-stakes assessment of competency, then reliability cannot be compromised to ensure no unsafe trainee is allowed to progress incorrectly. However, if the simulation is designated as a training tool that gives feedback for learning, reliability is less important, with efforts made to limit compromising validity, so that feedback is ensured to be of relevance to the task in question.

Test utility is therefore a function of all these factors and can be expressed in the conceptual model as follows (Van der Vleuten 1996):

$$\text{Utility} = \text{Reliability} \times \text{Validity} \times \text{Educational impact} \times \text{Acceptability} \times \text{Cost}$$

8.15 How to Practically Ensure Validity

Kane proposed a framework for evaluating the validity of a construct. This involves a chain of inferences to develop a validity argument (Kane 1992; Kane 2001; Kane 2006).

First, the proposed interpretive argument for a construct should be stated as clearly and explicitly as possible. Next, all available evidence for and against the validity argument can be investigated, and a coherent argument for the proposed interpretation of scores can be developed, as well as arguments against plausible alternate explanations. As a result of these evaluations, the interpretive argument may be rejected, or it may be improved by adapting the interpretation or measurement techniques to correct any problems identified. If the interpretive argument survives all reasonable challenges, it can be accepted provisionally, with the caveat that further factors may come to light in the future that challenge this argument.

This chain of inferences has four principle links that extend from simulation implementation to result interpretation. These are *scoring*, *generalization*, *extrapolation*, and *decision*.

8.15.1 Scoring

This concerns how observations on a participant's performance are made and how this performance is converted into a score. It evaluates if the simulation is reproducibly administered under standard conditions and includes scrutinizing the scoring rubrics, ensuring that they are applied constantly to all candidates and safeguarding security of the assessment so that no candidates gain an unfair advantage. One of the strengths of simulation assessment is that it can provide a standardized testing environment to all candidates. However, potential threats to this first inference can occur, including such things as simulation malfunction or vague scoring criteria. Validity evidence that addresses these issues might include regular checks and calibration of simulators and appropriate design and scrutiny of marking sheets by experts to ensure marking is homogenous.

8.15.2 Generalization

This concerns the inference that the performance tested is representative of the "universe" of scores that could be obtained in similar tasks. In other words, are the scores sufficiently representative of all other possible observations? The main threat to this is construct underrepresentation. Most simulations contain a relatively small number of items, which means making inferences about performance in the real world from simulation can be risky. Ensuring the simulation is constructed suitably and that appropriate sampling of the construct is undertaken will limit this issue. This inference also encompasses issues of reliability, internal consistency, and sources of

measurement error. One of the ways of strengthening the generalization inference is to increase the number of items tested. One of the strengths of simulation is that additional targeted models can be developed to ensure the breadth of surgical performance is covered. This could be achieved by generating simulation lists, when trainees can perform several different simulations is one sitting, much like a regular operating list.

8.15.3 Extrapolation

This inference is principally concerned with the extrapolation of simulation performance to real-world performance. This can be gauged by looking at the correlation between simulation scores and measures of real-life clinical performance. For example, there is a more robust argument that a knee arthroscopy simulation is able to predict real-life ability if experienced knee surgeons performed better than trainees. This represents construct validity (by demonstrating an ability to differentiate between surgeons of differing experience levels), and it is a key component of the extrapolation inference. Through a process termed "triangulation," other direct or indirect markers of ability can also be used in combination to strengthen this inference. Such an example is the use of motion tracking systems. Other measures that could be selected might include the results of in-training exams, OSCE scores, seniority, or other similar studies (Sullivan 2011).

8.15.4 Decision Making

When judgments are made about technical ability from simulation performance, cut scores are required to determine if individuals meet the required standard. It is important that the setting of these standards of pass and fail are robust and defensible (Boulet et al. 2003). Moreover, the wishes of other stakeholders impacted by these decisions must also be considered. Even if strong evidence exists of a simulator's validity from the three other inferences outlined already, if those to whom the results are important do not believe them to be credible or meaningful, then they are not valid (Scalese and Hatala 2013). For example, the general public would probably dismiss the credibility of a simulation assessment that allowed poorly performing surgeons to pass through without being identified and call its validity into question.

8.16 Discussion

Simulation training is an exciting area, with much potential for use in training orthopedic surgeons of the future. However, for its potential to be realized, it must be feasible, and its implementation must ensure simulated tasks and assessment systems have adequate reliability and validity.

In this chapter, we have discussed the various elements of validity desirable in simulation. Validity is a broad concept with many facets and should involve the accumulation of a variety of evidence to construct a strong validation argument. Careful thought is needed prior to this process to identify the simulation's application. It is important that future developers aim to coordinate their efforts with policy makers, those writing the curricula and simulation model manufacturers so that alignment is achieved between simulation and critical learning objectives. This would ensure that future training programs have a common theme and simulation is delivered with clear aims and in an effective manner (Table 8.4).

Table 8.4 Checklist for simulation validation

Checklist for validating simulation
1. Determine the construct
State the aims of the simulator. The construct's form will depend of several factors, such as:
(a) *What will its purpose be*?
e.g. introducing junior trainees to arthroscopy or high-stakes certification exams at the end of training
(b) *How will it be applied*?
(c) *Under what conditions will the simulation take place*?
(d) *What will it measure*? *What are the outcome parameters*?
Performance metrics e.g. time taken, motion analysis
Rater scoring e.g. checklists, rating scales or subjective assessment
End product
(e) *What group of people will it be used with*?
If various groups are to use the simulator, validation must include these groups
(f) *What type of model will be used*?
Phantom model/benchtop model
Cadaveric
Virtual reality
Simulated patient actors
(g) *What evidence is there within the literature for this simulation modality*?
2. Content evidence
(a) Expert panel
Was there expert consensus on the construct design including formal task analysis?
(b) Instrument validation
Are new instruments based on previously validated instruments?
(c) Pilot testing
Have the simulation instruments been developed and revised through piloting and modified as appropriate?
(d) Score framework
What evidence was used to determine scoring methods and can a scoring blueprint be prepared?
(e) Test blueprinting
Is a blueprint used to develop test instruments?
(f) Evidence of content-construct mismatch
Is there any discrepancy between alignment of test content and the construct?
3. Reliability tests
(a) Test/retest
(b) Internal consistency
(c) Inter-/intra-rater reliability
4. Test consequences
(a) *How will test thresholds be established*?
e.g. Angoff method, modified borderline group method, Markov modeling, ROC curve
(b) *Have unanticipated test consequences been considered*?
5. Feasibility
(a) Ethical considerations and institutional approval
(b) Consideration of cost implication for local unit
6. Educational issues
Establish how learner feedback is to be delivered:
Metrics such as time taken, instrument path length, etc.
Video
Performance score e.g. check list, scoring rubric, GRS, etc.
One-to-one debriefing with experienced surgeon
7. Predictive validity
Establish correlation of performance in real-world environment with simulation performance

Bibliography

Alvand A, Logishetty K, Middleton R, Khan T, Jackson WF, Price AJ, Rees JL (2013) Validating a global rating scale to monitor individual resident learning curves during arthroscopic knee meniscal repair. Arthroscopy 29(5):906–912

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

Aucar JA, Groch NR, Troxel SA, Eubanks SW (2005) A review of surgical simulation with attention to validation methodology. Surg Laparosc Endosc Percutan Tech 15(2):82–89

Boulet JR, De Champlain AF, McKinley DW (2003) Setting defensible performance standards on osces and standardized patient examinations. Med Teach 25(3): 245–249

Brennan RL (2010) Generalizability theory. Springer, New York

Carter FJ, Schijven MP, Aggarwal R, Grantcharov T, Francis NK, Hanna GB, Jakimowicz JJ (2005) Consensus guidelines for validation of virtual reality surgical simulators. Surg Endosc 19(12):1523–1532

Chikwe J, De Souza AC, Pepper JR (2004) No time to train the surgeons. Br Med J 328(7437):418–419

Clauser BE, Margolis MJ, Swanson DB (2008) Issues of validity and reliability for assessments in medical education. In: Practical guide to the evaluation of clinical competence. p 10–23

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46

Cohen J (1968) Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull 70(4):213–220

Cook DA, Beckman TJ (2006) Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med 119(2):166.e7–116.e16

Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. Psychometrika 16(3):297–334

Cronbach LJ, Meehl PE (1955) Construct validity in psychological tests. Psychol Bull 52(4):281–302

Daley J, Forbes MG, Young GJ, Charns MP, Gibbs JO, Hur K et al (1997) Validating risk-adjusted surgical outcomes: site visit assessment of process and structure. J Am Coll Surg 185(4):341–351

Darzi A, Smith S, Taffinder N (1999) Assessing operative skill: needs to become more objective. Br Med J 318(7188):887–888

Downing SM (2003) Validity: on the meaningful interpretation of assessment data. Med Educ 37(9):830–837

Downing SM (2004) Reliability: on the reproducibility of assessment data. Med Educ 38(9):1006–1012

Downing SM, Haladyna TM (2004) Validity threats: overcoming interference with proposed interpretations of assessment data. Med Educ 38(3):327–333

Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76(5):378–382

Gallagher AG, Ritter EM, Satava RM (2003) Fundamental principles of validation, and reliability: rigorous science for the assessment of surgical education and training. Surg Endosc 17(10):1525–1529

Garden A (2008) Research in simulation. In: Riley RH (ed) Manual of simulation in healthcare. Oxford University Press, New York

Harvill LM (1991) Standard error of measurement. Educ Measure Issues Pract 10(2):33–41

Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M (1999) OSCE checklists do not capture increasing levels of expertise. Acad Med 74(10):1129–1134

Hodges B, McNaughton N, Regehr G, Tiberius R, Hanson M (2002) The challenge of creating new OSCE measures to capture the characteristics of expertise. Med Educ 36(8):742–748

Hodgins JL, Veillette C (2013) Arthroscopic proficiency: methods in evaluating competency. BMC Med Educ 13(1):61–69

Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR (2010) The role of assessment in competency-based medical education. Med Teach 32(8):676–682

Jarvis-Selinger S, Pratt DD, Regehr G (2012) Competency is not enough: integrating identity formation into the medical education discourse. Acad Med 87(9):1185–1190

Kane MT (1992) The assessment of professional competence. Eval Health Prof 15(2):163–182

Kane MT (2001) Current concerns in validity theory. J Educ Measure 38(4):319–342

Kane M (2006) Validation. In: Brennan RL (ed) Educational measurement. ACE/Praeger, Westport, pp 7–64

Kassab E, Tun JK, Kneebone RL (2012) A novel approach to contextualized surgical simulation training. Simul Healthc 7(3):155–161

Kazdin AE (2003) Research design in clinical psychology, 4th edn. Pearson

Kneebone R (2009) Perspective: simulation and transformational change: the paradox of expertise. Acad Med 84(7):954–957

Kunkler K (2006) The role of medical simulation: an overview. Int J Med Robot 2(3):203–210

Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, Brown M (1997) Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg 84(2):273–278

McHugh ML (2012) Interrater reliability: the kappa statistic. Biochem Med 22(3):276–282

Mehay R, Burns R (2009) Available from: http://www.bradfordvts.co.uk/wp-content/onlineresources/0002mrcgp/mrcgp%20in%20a%20nutshell.ppt. Accessed 20 Jan 2014

Messick S (1989) Validity. In: Linn RL (ed) Educational measurement. American Council on Education/ Macmillan, New York, pp 13–103

Messick S (1995) Validity of psychological assessment: validation of inferences from persons' responses and

performances as scientific inquiry into score meaning. Am Psychol 50(9):741–749
Michelson JD (2006) Simulation in orthopaedic education: an overview of theory and practice. J Bone Joint Surg 88(6):1405–1411
Miller GE (1990) The assessment of clinical skills/competence/performance. Acad Med 65(9):S63–S67
Motola I, Devine LA, Chung HS, Sullivan JE, Issenberg SB (2013) Simulation in healthcare education: a best evidence practical guide. AMEE guide no. 82. Med Teach 35(10):e1511–e1530
Murray D (2012) Review article: assessment in anesthesiology education. Can J Anesth 59(2):182–192
Regehr G, MacRae H, Reznick RK, Szalay D (1998) Comparing the psychometric properties of checklists and global rating scales for assessing performance on an osce-format examination. Acad Med 73(9):993–997
Sadideen H, Hamaoui K, Saadeddin M, Kneebone R (2012) Simulators and the simulation environment: getting the balance right in simulation-based surgical education. Int J Surg 10:458–462
Scalese RJ, Hatala R (2013) Competency assessment. In: Levine AI, DeMaria S, Schwartz AD, Sim AJ (eds) The comprehensive textbook of healthcare simulation. Springer, New York
Schuwirth LW, Van Der Vleuten CP (2004) Changing education, changing assessment, changing research? Med Educ 38(8):805–812
Sedlack RE (2011) Validation process for new endoscopy teaching tools. Tech Gastrointest Endosc 13(2):151–154
Shantz JAS, Leiter JR, Gottschalk T, MacDonald PB (2014) The internal validity of arthroscopic simulators and their effectiveness in arthroscopic education. Knee Surg Sports Traumatol Arthrosc 22:33–40
Spencer F (1978) Teaching and measuring surgical techniques: the technical evaluation of competence. Bull Am Coll Surg 63(3):9–12
Sullivan GM (2011) A primer on the validity of assessment instruments. J Grad Med Educ 3(2):119–120
Van Der Vleuten CP (1996) The assessment of professional competence: developments, research and practical implications. Adv Health Sci Educ 1(1):41–67
Wass V, Vleuten CVD, Shatzer J, Jones R (2001) Assessment of clinical competence. Lancet 357(9260): 945–949