Chapter 34
Measuring the Impact of Educational Interventions: A Quantitative Approach

Jenepher A. Martin

Overview This chapter will discuss impact evaluation, an important method of measuring the effectiveness of an educational intervention. This form of evaluation represents a subset of program evaluation and focuses on outcomes and consequential events related to an educational intervention. In doing so, it incorporates several different quantitative methods and is typically reserved for stable, long-standing educational programs/curricula. Many of these methods are also used as part of program evaluation as a whole and in surgical research. Readers are directed to Chaps. 23 ("Demystifying Program Evaluation for Surgical Education", Battista et al.) and 30 ("Researching in Surgical Education: An Orientation", Ajjawi and McIllhenny) for more information on these subjects. In addition to providing a working definition of impact evaluation, this chapter will help define key concepts related to its successful use as well as aid in delineating the most useful quantitative methods to employ.

34.1 Introduction

The distinction between evaluation and research is important to reiterate in the context of this chapter. Patton [1] reminds us that evaluation research is a subset of program evaluation and more knowledge-oriented than decision and action oriented. He points out that systematic data collection for evaluation includes social science research methods and, in addition, other sources of data about programs. In the surgical education context, these may include statistics relating to training programs, assessment information and practice observation. Patton's views help us to get over our fixation on experimental method and desire for generalizability of

J. A. Martin (✉)
Medical Student Programs, Eastern Health Clinical School, Box Hill, Australia

Faculty of Medicine Nursing and Health Sciences, Monash University, Clayton, Australia

School of Medicine, Deakin University, Melbourne, Australia
e-mail: jenepher.martin@monash.edu

D. Nestel et al. (eds.), *Advancing Surgical Education*, Innovation and Change in Professional Education 17, https://doi.org/10.1007/978-981-13-3128-2_34

evaluation results and value the usefulness of evaluation in our own context. This in turn promotes a pragmatic approach of making the best judgements and decisions with the available information.

This chapter will discuss impact evaluation, and specifically quantitative methods for contemporary evaluation practice. A working definition of impact evaluation will be developed, followed by a discussion of impact evaluation design and specific applicable quantitative methods. Examples from surgical education will highlight quality of education measurement in research and evaluation. Throughout this chapter the term 'program' will be used in a generic way for any educational event, intervention or course.

34.2 What Is Impact Evaluation?

Impact evaluation focus is on outcomes and consequential effects [2], and impact evaluation is usually undertaken for an established program and with summative intent. By their very nature, impact evaluations are retrospective and assume program stability over time sufficient to have observable impacts. In the context of this chapter, the impact must also be measurable.

Impact evaluation designs are also suitable for evaluation of pilot interventions and for comparisons of two or more interventions, providing the interventions are in steady state for the period of evaluation. Thus, the findings of impact evaluation may also be useful for formative purposes in program evaluation. For example, if unintended outcomes are uncovered that are undesirable then even a stable program may be revised and improved. Attempting impact evaluation too early in program implementation, or during program development, risks unreliable and untrustworthy results, with incorrect inferences being made about the program in question and, ultimately, poor decision-making.

Impact evaluation is applicable to both large and small educational programs or interventions, when intended outcomes are clearly understood and defined. Of worldwide relevance to surgical practice, implementation of the World Health Organization (WHO) Surgical Safety Checklist from 2009 had measurable positive impacts on patient outcomes reported within 3 years [3]. On a smaller scale, Evers et al. [4] used a combined process and impact evaluation design to examine a social marketing campaign to increase asthma awareness among older adults in an Australian community. At your own local level, the immediate change in attitudes or behaviour for education participants could be the focus for impact evaluation and unintended outcomes you uncover may need to be addressed for ongoing implementation.

Your evaluation may relate to a small educational workshop you have developed and implemented, an aspect of a national surgical training program at local, regional, or national level, or the local impact of a worldwide program. Common principles apply at all levels, and the remainder of this chapter will address:

- Impact evaluation design
- Focusing impact evaluation
- Quantitative methods for impact evaluation

34.3 Designing Impact Evaluation

A practical evaluation design framework has been introduced in Chap. 23 ("Demystifying Program Evaluation for Surgical Education", Battista et al.), and the design flow diagram below (Fig. 34.1) complements the framework. When considering an impact evaluation, three key aspects require clarification:

(i) Is impact the most suitable form of evaluation?
(ii) What outcomes/impacts are of interest?
(iii) Which methods are required for the evaluation?

(i) *Is Impact the Most Suitable Form of Evaluation?*

Before launching into your impact evaluation design, determine if the program you are intending to evaluate is ready for impact evaluation and if the evaluation questions you are interested in relate to impact or another aspect of the program.

Characteristics of the program that indicate readiness for impact evaluation include full implementation, stability and a temporal duration that is sufficient for impacts of interest to have occurred [1, 2]. Clearly these criteria may be met sooner for small, local educational interventions such as a student workshop than for large and complex programs such as surgical training. Even if a program meets the criteria for impact evaluation, this may not be the preferred focus. You may need to spend some time considering this and discussing with program stakeholders just what it is they want to know about the program and for what purpose. Remember, impact evaluation can be formative, but may not be the best approach for programs in development or early implementation. On the other hand, for an established program under review, the question of impact is highly relevant.

(ii) *What Outcomes/Impacts Are of Interest?*

Once the decision to undertake an impact evaluation has been reached, the questions for evaluation are defined. In medicine, research that is valued often has an unashamedly positivist perspective, where objective reality can be quantified and defined by measurement. Tavakol and Saunders [5] remind us that in education a post-positivist approach often sits more comfortably and allows for mixed methods. To use quantitative measures in educational evaluation, however, questions related to output, outcome, or impact measures are required. In considering your evaluation questions the 'distance to target' or 'reach' of the program is a useful concept (Fig. 34.2). Is the evaluation interested in immediate effects on participants, or the longer-term outcomes and impacts on patient care for example? The impact of the implementation of the WHO Surgical Safety Checklist has been evaluated at indi-

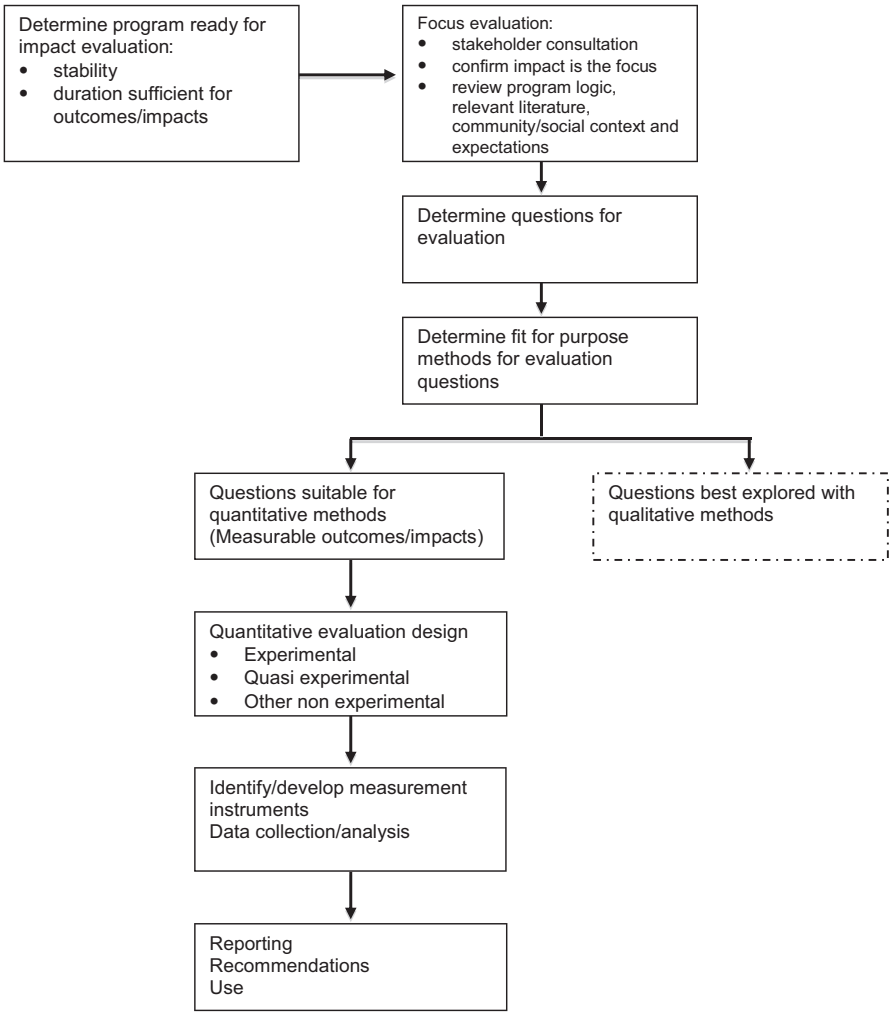

Fig. 34.1 Impact evaluation design

vidual [6] and patient outcome levels [3]. As noted in Chap. 30 ("Researching in Surgical Education: An Orientation", Ajjawi and McIllhenny), longer-term and distant impacts from educational interventions, such as patient outcomes, may be inaccessible to local researchers or evaluators. Information about more immediate outcomes for participants in the local context, such as changes in surgical team members' awareness of patient safety after checklist introduction described by Papaconstantinou et al. [6], informs the local program and supports the positive global impact objective of WHO.

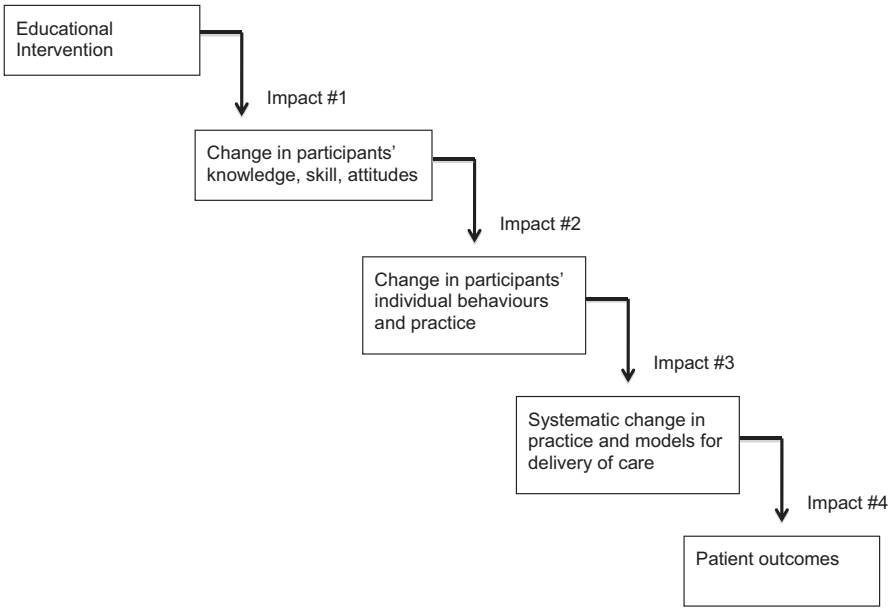

Fig. 34.2 Distance to impact

The critical step of stakeholder consultation may result in a range of evaluation questions, especially if there are competing interests for various stakeholders. For example, in multi-site specialty training programs, hospital-based supervisors may be primarily interested in consistency of implementation across sites or equity of access to learning resources and opportunities, and impact evaluation may not be the most suitable approach. On the other hand, the surgical college faced with prioritizing funding for the overall program may be seeking information about the impact of the program in terms of training outcomes. For a small local intervention such as a suturing workshop, the workshop facilitator may be interested in the immediate impact of the intervention on participants' surgical skills and/or their subsequent opportunities to put these into practice. The questions for the evaluation determined through stakeholder consultation will determine methodology, data collection and analysis.

In addition to stakeholder consultation, the 'logic' or 'theory' of the program will inform the evaluation questions, as this defines the planned outputs, outcomes and impacts [2]. As good quantitative research is driven by theory [5], so is good quantitative impact evaluation.

(iii) *Which Methods Are Required for the Evaluation?*

Ideally, methodology is considered after determining the questions for the evaluation, leading to an approach where 'fit for purpose' methods are matched with evaluation questions and there are no resource constraints. In practice, this stage of

the design is often one of compromise and reality checking. There may be existing 'good enough' sources of data available for low cost; tight time lines may preclude longitudinal data collection; or small numbers of participants limit statistical power. Your aim is to conduct rigorous evaluation within local constraints to make the best judgements and decisions about your program under the circumstances.

The remainder of this chapter will focus on quantitative methods for impact evaluation studies, with an emphasis on the development and validation of measurement instruments.

34.4 Quantitative Approaches for Impact Evaluation

Many of the concepts for the design of impact evaluation for educational interventions using quantitative methodology will be familiar to surgeons and surgical trainees as there are similarities with clinical research methods. For impact evaluation, important considerations include:

- The study design
- Sources of data, sampling and surveys
- Data analysis and reporting
- Measurement instruments

34.4.1 Quantitative Evaluation Design

Quantitative designs range from experimental to descriptive [5] and are summarized in Table 34.1. What best suits your program evaluation will be determined by multiple factors including your evaluation questions, any hypotheses you put forward, available resources, the structure of the program being evaluated, and the context of the educational program.

Experimental design for program evaluation purposes is not common in the surgical education literature. There are, however, some illustrative studies with a research purpose. Seymour et al. [7] conducted a randomized, double blinded study exploring the effect of virtual reality (VR) training on operating performance of surgical trainees. This demonstrated benefits of VR training over standard training of decreased operative time and fewer errors in the experimental group. Although this study is described as double-blinded, the participants in the education intervention were, of course, aware of their randomization status, in contrast to many therapeutic trials. The two sets of assessors were blinded to participant status in this case. Other randomized studies of interest include the examination of three different education conditions on transfer of operative skills to a cadaver model [8] and work by Moulton et al. [9] on the role of distributed practice in surgical skill development. These studies have contributed new knowledge to surgical education; however, in

Table 34.1 Designs for quantitative impact evaluation

Experimental	Random assignment to intervention (experimental) and non-intervention (control) groups
	Participants/assessors may be 'blinded' to intervention – single (participants) or double (participants and assessors) blinding
	Crossover designs are sometimes used
	Advantage: strong causal inferences
	Disadvantage: experimental conditions may not reflect the real world, design not practical for many evaluations
Quasi experimental	Non experimental design,
	Assignment to intervention/non intervention not random
	Single group design can be utilized (e.g., pre/post-intervention evaluation of participants)
	Participants/assessors may be 'blinded' to intervention – single (participants) or double (participants and assessors) blinding
	Advantage: practical design option for real world settings
	Disadvantage: complex statistical analysis often required; causal inferences made with caution
Correlational	Non experimental design
	Explores the associations between features of education programs (the variables). Can be used where assignment to a group of interest is not possible. Data can be quantitative or nominal, and exploration of relationships between two or more variables conducted
	Advantage: Suitable for most contexts. Exploration of associations can lead to further evaluation of important program aspects
	Disadvantage: Limited use when relationship is not linear, range is restricted or outliers in data. No causal inferences possible
Descriptive	Detailed documentation of program outcomes/impacts using descriptive statistics (frequencies, percentages, etc.), and graphics
	Applicable for all impact evaluation and provides overview
	Often required for reporting to stakeholders
	Advantage: Facilitates clarification of the program and explicit understanding of outputs and outcomes of interest. May identify unintended outcomes/impacts
	Disadvantage: No causal inferences possible

the context of impact evaluation of established programs, randomization of participants may be impractical, particularly so if the education is high stakes and there could be any perception of benefit or disadvantage for participants to randomization status. For 'near target' and low-stakes education programs, experimental impact evaluation design may be possible and helpful for comparing interventions under consideration.

Quasi-experimental designs [5, 10], on the other hand, are often very practical for program evaluation, although less familiar to surgeons. Quasi-experimental designs are described comprehensively in Cook and Campbell's [10] classic work on the subject, and interested evaluators are encouraged to explore further. Practical and commonly used quasi-experimental options suitable for your initial practice

include single group or control group pre-test/post-test designs, and interrupted time-series designs. These designs fall into the category of non-randomized intervention studies. Because randomization to intervention status is not used, causal inferences are not as robust as in experimental designs. However, the suitability of quasi-experimental designs for real world clinical practice and surgical education settings where randomization may be precluded due to ethical, logistic or cost considerations offsets the weaker certainty about causal inferences in these designs. Despite limitations, useful and timely information from complex and uncontrolled contexts is often acquired, facilitating decision-making.

- Pre-test/post-test designs.

The choice between single group or control group pre-test/post-test evaluation is often pragmatic. The use of a control group, even though selection bias is not managed by randomization, does enhance validity of findings. Obviously, demographic data and context details can be used to establish how closely matched control and intervention groups are. An issue you may face when using control group designs, randomized or not, is the perception of 'fairness' when some students have access to an intervention and others don't. In some situations this can be managed by delivering your intervention to the control group post hoc; in others, a single group design is the most acceptable solution. Documented change in surgical team members' perspectives before and after the introduction of the WHO Surgical Safety checklist [6] is an example of this design.

- Interrupted time series designs.

These designs are cohort studies, either cross-sectional or longitudinal, a concept familiar to many surgical educators from clinical epidemiology. Integral to these designs are multiple measurements over time, both before and after an intervention of interest. As with other quasi-experimental designs, inferences about the intervention causing observed effects must be made cautiously. Examples of this design include some studies discussed by Fudickar [3] in relation to the WHO Surgical Safety checklist and the study by Martling et al. [11] of the effect of surgeon training on rectal cancer outcome.

Non-experimental correlational designs are very common in evaluation practice. Exploring associations between variables is important; however, it does not imply causation of observed effects. In impact evaluation, correlational studies may be very useful in uncovering unintended outcomes or impacts, which may then require further study. In thinking about the use of correlational designs in your practice, the concept of 'natural experiments' may be useful, [10] where there is very little manipulation of the environment and/or no specific intervention. Exploration of the association between video game experience and laparoscopic skill is one such example [12], raising interesting questions for further research.

34.4.2 Sources of Data, Sampling and Surveys

The data required for your impact evaluation will be determined by your evaluation questions, and identifying sources of data is part of the planning process (Chap. 23, "Demystifying Program Evaluation for Surgical Education", Battista et al.). Selecting sources of data, gaining access to these and obtaining the data for analysis constitutes the assembly of evidence on which your ultimate judgements and recommendations are based [2].

Accessing the data you want is not always straightforward. For example, performance data from health services, universities and surgical training organizations may require formal application. Negotiation with third parties to distribute surveys may be required. Existing databases, while useful, may not have all the information you want, leading to modification of your plans. Ethical, logistic, financial and political considerations will also come into play. Bear in mind that when you obtain outcome/impact data from others you will be relying on the quality of measurement that generated those data without necessarily knowing how robust that measurement is.

For some impact evaluations, it will be possible to predictably obtain outcome/impact measures from all program participants. In other evaluations the population of interest may require sampling, and your approach to sampling should be determined and made explicit in the evaluation planning. The aim is to achieve a representative sample of your program participants. You will most likely use non-probability sampling methods such as convenience sampling, purposive sampling and quota sampling [13].

Data collection often involves surveys and these may include multiple forms of data including demographic information as well as embedded measurement instruments. Artino et al. [14] offer practical advice about survey design for medical education research, underpinned by sound theory [15, 16]. Underlying measurement principles are discussed further below.

34.4.3 Data Analysis and Reporting

The purpose of analysis is to make sense of the data, to construct meaning and ultimately answer your evaluation questions. Data management and analysis as described by Owen [2] involves constructing 'an organized assembly of information' or 'data display', data reduction to simplify and transform raw information and then drawing conclusions that relate to evaluation questions.

Quantitative evaluation designs require statistical analysis, and it is critical you seek advice about this in the planning stages. As evaluators, we want to make the best judgements and decisions we can with the available evidence even if the evidence would not meet clinical decision-making standards. Remember, evaluation research is a subset of program evaluation, and statisticians familiar with analysis of experimental data in biomedical research may not be familiar with some of the more

sophisticated analyses in quasi-experimental and correlational designs [10]. Clinical epidemiologists may well be able to advise about analysis of data for interrupted time-series designs. Educational measurement and associated analyses is a separate area of expertise and briefly discussed below.

Intended users of your evaluation are ideally involved during data analysis, interpretation, making judgements, and recommending consequent actions [1]. This co-construction of the evaluation outcome between evaluator and stakeholders is a distinct difference between research and evaluation and promotes use of your evaluation.

34.4.4 Identifying Measures

Measurement of outcomes, impacts and consequences is central to quantitative approaches described above, and precision of measurement underpins the robustness of the results. Precise, accurate measurement depends on reliable and valid measurement instruments.

Some outcome measures, such as mortality, numbers of errors or time are clear; however, many are more complex. The two key concepts of reliability and validity underpin your choice of measurement instrument, and these will be outlined now. For in depth information about measurement theory, further reading is recommended. An additional consideration in designing evaluations is the feasibility of implementing your measure. Reliability is a pre-requisite for measurement validity and so will be discussed first.

- Reliability

Measurement reliability refers to the consistency of scores; however, on its own, it is not sufficient to provide evidence for validity [17, 18]. Measures of reliability may quantify internal consistency of the instrument, reproducibility over time or inter-rater agreement (Table 34.2). High reliability indicates consistency with little error in the measurement, considered important when the stakes are high.

For the assessment of non-technical skills of surgeons, the Non-Technical Skills for Surgeons (NOTSS) Behaviour Rating System was developed, and reliability information is available for this relating to internal structure and inter-rater reliability [19]. In assessments of operative skill in surgical trainees, comparative reliability of global ratings and checklist scoring systems has been examined [20]. Reliability information such as this is helpful in selecting measurement instruments with the caveat that the reliability of a measure is not inherent in the instrument itself, but relates to the scores obtained, and using an instrument under different conditions (e.g. context, population or rater training status) may change the reliability. Reliability studies such as those discussed are often the first published information about measurement instruments in surgical education; however, further validity evidence is required for confident use in research and evaluation.

Table 34.2 Types of reliability

Type of reliability		Methods for reliability calculation
Internal consistency	Commonly used for tests that relate to a single construct such as 'knowledge' or 'empathy' where each item in the test should be well correlated with other items. High internal consistency supports the single construct	Split half reliability
		Kuder Richardson
		Cronbach's alpha
Stability of the measurement	These measures assess 'in-person' stability of measures either across time (test-retest) or equivalent versions of the test (parallel forms). There is an assumption that the subject remains stable with respect to the measured construct between test occasions or forms	Test-retest reliability
		Parallel forms reliability
Inter-rater reliability (IRR)	These measures assess the agreement between different raters of the same subject test performance using the same rating instrument. The most appropriate measure for IRR calculation will be determined by factors such as the form of measurement data (rank, dichotomous, continuous) and number of raters	Percent agreement
		Phi (correlation)
		Kappa
		Kendall's Tau
		Intraclass correlation
Generalizability of the measure	Assigns the variance of test scores to multiple possible sources (subjects, raters, items, etc.). Understanding where score variance is attributed is helpful in planning interventions to improve assessment such as rater training	Generalizability coefficient

APA [17, 18] and Cook and Beckman [21]

- Validity

For educational evaluations, you will often want to measure outcomes in terms of knowledge, skills and attitudes of participants in the program or of others impacted by the program. To do this, you will require a measurement instrument of some type that you are confident is actually measuring the construct of interest. The types of instruments you could consider include educational tests and examinations, observed performance ratings, attitude rating scales, psychological tests, and questionnaires. So how do you know the instrument you are considering does actually measure the construct you are interested in? After all, for many constructs, the measure is a proxy as the underlying construct is not directly measurable [21]. For example, empathy can be inferred from physicians' self-reported perceptions and behaviours [22].

Validity, as defined by the APA standards is 'the degree to which evidence and theory support the interpretations of test scores entailed by the proposed use of tests' and that validation is the 'process of constructing and evaluating arguments for and against the identified interpretation of test scores and their relevance to the proposed use' [17, 18]. This definition highlights that validation of a measurement instrument requires supporting evidence, that meaning is derived from measurements and not inherent to scores in themselves, and that validity of a specific measure is context specific. So, validation of a measurement instrument uses multiple sources of evidence, is cumulative and takes time. One contemporary view of validity is that all sources of evidence relate to construct validity, with five broad categories identified [17, 18, 21, 23] (Table 34.3).

Table 34.3 Evidence to support validity of measurement

Evidence category	Question answered	Criteria to consider
Content	How well does the content of the measurement represent the underlying construct?	Construct definition
		Intended purpose of the measurement
		Process for instrument development (blueprinting, sampling, item development, etc.)
		Item quality, wording
		Qualifications of item writers and review process
Response process	Is the response process/ behaviour of the subject to the test item(s) consistent with the underlying construct being measured?	Theoretical and /or empirical analysis of the processes test takers use in their response to item(s). (e.g. in a test of clinical reasoning, are they undertaking this process or simply applying a learned algorithm?)
	Are the judgement-making processes of the raters consistent with the intended use of the test scores?	Empirical analysis of the criteria used to arrive at judgements. (e.g. clinical performance assessment should not be influenced by unrelated student factors such as gender or race)
Internal structure	Are the relationships between test items consistent with the underlying construct?	Internal consistency as evidence for homogeneity and single construct vs multifactorial structure
		Factor structure alignment to theoretical construct(s)
	How well does test performance reflect predicted performance of particular subgroups with respect to the underlying construct?	Differential performance aligned with construct prediction. E.g. more senior trainees perform better
Relation to other variables	Is the relationship with other variables as expected based on the predicted relationship between the constructs measured?	Positive correlations between two measures that are either expected to co-vary (e.g. engagement with clinical learning and clinical performance), or are measuring the same construct (e.g. knowledge tests of common content in different formats,)
		Negative or no correlation between measures consistent with expectations based on the underlying constructs being measured (e.g. eye colour and surgical skill)
Consequences	What are the consequential effects, intended and unintended, of the assessment?	Behaviours of test takers in response to the format of the assessment. (E.g 'OSCE practice' vs authentic patient/student interactions; rote learning vs deep learning)
		Methods and criteria used to determine categorization of test takers leading to subsequent consequential outcomes for them. (E.g. pass/fail cut scores, degree of depression, level of intelligence)

APA [17, 18], Cook and Beckman [21], and Cook and Hatala [23]

Table 34.4 Example constructs and potential measurement methods in surgical education

Construct	Type of assessment	Candidate measurement methods
Knowledge (e.g. clinical sciences, disease specific information)	Written or oral tests of knowledge	Multiple choice tests, short answer or essay questions. Viva tests
	Written or oral tests of applied knowledge/problem solving (note: item construction and format must be matched to intended level of knowledge testing)	Simulation- or clinical-based objective structured clinical examination
	Performance-based assessment of applied knowledge/problem solving (note item construction matched to this assessment objective)	Observed clinical practice
Clinical skills (e.g. history, examination)	Performance-based assessment of applied knowledge/problem solving (note item construction matched to this assessment objective)	Simulation- or clinical-based objective
Communication skills		Observed clinical practice
Teamwork		
Procedural skills	Performance-based assessment of applied knowledge/problem solving	Direct observation procedural skills (DOPS),
		Objective structured assessment of technical skills (OSATS),
		Time and motion analysis,
		Error analysis
		Product quality assessment

- Selecting a measurement instrument.

Be clear about the constructs you want to measure and specify these as precisely and accurately as possible. The construct definitions will determine what measurement instruments are appropriate. For example, a written test of anatomy measures knowledge, not surgical skill (Table 34.4). The caveat here regarding validity evidence is for the interpretation of measurement for which it was established. If you are using a measurement for an alternative interpretation then validity should be established for that use [21, 23].

As surgical educators, we are often interested in student or trainee outcomes, and the following practical examples will illustrate the common options for choosing measurement instruments: (i) use existing data, (ii) use an 'off the shelf' instrument or (iii) design a new instrument.

Examination scores are one of the most frequently used sources of existing data for education outcomes/impact evaluation. If you are using these data, endeavour to assure yourself of the validity of the measurement. One disadvantage of using existing test scores is that validation studies may not have been undertaken.

An ‘off the shelf’ test may be the best choice for psychological constructs such as empathy or self-efficacy. Many of these instruments have been used in large and/or diverse populations and norms are established. Checking what validation evidence is available and in what use contexts can help you decide if an ‘off the shelf’ test is suitable. If you were interested in surgeon empathy, you might choose the Jefferson Scale of Physician Empathy [22].

When you are unable to identify a suitable measurement instrument, it may be necessary to develop one, or modify an existing one. In both cases pilot studies to validate the measure are required. The objective structured assessment of technical skills (OSATS) is an example of a new instrument developed when no suitable measure was available [24, 25]. Since its development, OSATS has become an established measure in surgical education research, evaluation and training.

34.5 Conclusion

Impact evaluation is a specific evaluation form applicable to stable programs, large or small, with defined impacts and outcomes. Quantitative methodology for impact evaluation includes experimental, quasi-experimental and non-experimental design. Measurement of outcomes/impact for evaluation must be reliable and valid for credible judgments and well-founded decision-making.

References

1. Patton, M. Q. (1997). *Utilization-focused evaluation* (4th ed.). Thousand Oaks: Sage Publications.
2. Owen, J. M. (2006). *Program evaluation: Forms and approaches* (3rd ed.). Crows Nest: Allen and Unwin.
3. Fudickar, A., et al. (2012). The effect of the WHO Surgical Safety Checklist on complication rate and communication. *Deutsches Ärzteblatt International, 109*(42), 695–701.
4. Evers, U., et al. (2013). ‘Get your life back’: Process and impact evaluation of an asthma social marketing campaign targeting older adults. *BMC Public Health, 13*, 759–768.
5. Tavakol, M., & Sanders, J. (2014). Quantitative and qualitative methods in medical education research: AMEE Guide No 90: Part I. *Medical Teacher, 36*(9), 746–756.
6. Papaconstantinou, H. T., et al. (2013). Implementation of a surgical safety checklist: Impact on surgical team perspectives. *The Oschner Journal, 13*, 299–309.
7. Seymour, N. E., et al. (2002). Virtual reality training improves operating room performance. Results of a randomized, double-blinded study. *Annals of Surgery, 236*(4), 458–464.
8. Anastakis, D. J., et al. (1999). Assessment of technical skills transfer from the bench training model to the human model. *American Journal of Surgery, 177*(2), 167–170.
9. Moulton, C. E., et al. (2006). Teaching surgical skills: What kind of practice makes perfect? A randomized, controlled trial. *Annals of Surgery, 244*(3), 400–409.
10. Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings* (1st ed.). Chicago: Rand McNally.

11. Martling, A. L., et al. (2000). Effect of a surgical training programme on outcome of rectal cancer in the County of Stockholm. *Lancet, 356*(9224), 93–96.
12. Rosser, J. C., et al. (2007). The impact of video games on training surgeons in the 21st century. *Archives of Surgery, 142*, 181–186.
13. Tavakol, M., & Sanders, J. (2014). Quantitative and qualitative methods in medical education research: AMEE Guide No 90: Part II. *Medical Teacher, 36*(10), 838–848.
14. Artino, A. R., et al. (2014). Developing questionnaires for educational research: AMEE Guide No.87. *Medical Teacher, 36*(6), 463–474.
15. DeVellis, R. F. (2014). *Scale development: Theory and applications* (2nd ed.). Newbury Park: Sage Publications.
16. Dillman, D., et al. (2009). *Internet, mail and mixed-mode surveys: The tailored design method* (3rd ed.). Hoboken: Wiley.
17. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
18. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
19. Yule, S., et al. (2008). Surgeons' non-technical skills in the operating room: Reliability testing of the NOTSS behaviour rating system. *World Journal of Surgery, 32*, 548–556.
20. Regher, G., et al. (1998). Comparing psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine, 73*(9), 993–997.
21. Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine, 119*, 166. e7–166.e16.
22. Hojat, M., et al. (2002). Physician empathy: Definition, components, measurement and relationship to gender and speciality. *American Journal of Psychiatry, 159*(90), 1563–1569.
23. Cook, D. A., & Hatala, R. (2016). Validation of educational assessments: A primer for simulation and beyond. *Advances in Simulation, 1*, 31.
24. Martin, J. A., et al. (1997). Objective structured assessment of technical skill (OSATS) for surgical residents. *British Journal of Surgery, 84*, 273–278.
25. Reznick, R., et al. (1997). Testing technical skill via an innovative 'bench station' examination. *American Journal of Surgery, 173*(3), 226–230.