Use of Competency-Based Metrics to Determine Effectiveness of a Postgraduate Thoracoscopy Course

Henri G. Colt^a Mohsen Davoudi^a Silvia Quadrelli^b Nazanin Zamanian Rohani^a

^aDivision of Pulmonary and Critical Care Medicine, University of California, Irvine, Calif., USA; ^bBritish Hospital, Buenos Aires, Argentina

Key Words

Competency-based metrics • Education • Minimally invasive surgery • Thoracoscopy

Abstract

Background: Despite the paradigm shift from process to competency-based education, no study has explored how competency-based metrics might be used to assess short-term effectiveness of thoracoscopy-related postgraduate medical education. **Objectives:** To assess the use of a single-group, pre-/post-test model comprised of multiple-choice questions (MCQ) and psychomotor skill measures to ascertain the effectiveness of a postgraduate thoracoscopy program. **Methods:** A 37-item MCQ test of cognitive knowledge was administered to 17 chest physicians before and after a 2-day continued medical education-approved program. Pre- and post-course technical skills were assessed using rigid videothoracoscopy simulation stations. Competency-based metrics (mean relative gain, mean absolute gain, and *class-average normalized gain* <g>) were calculated. A <g> >30% was used to determine curricular effectiveness. **Results:** Mean cognitive knowledge score improved significantly from 20.9 to 28.7 (7.8 ± 1.3 points, $p < 0.001$), representing a relative gain of 37% and an absolute gain of 21%.

Mean technical skill score improved significantly from 5.20 to 7.82 (2.62 ± 0.33 points, $p < 0.001$), representing a relative gain of 50% and an absolute gain of 33%. Non-parametric testing confirmed t test results ($p < 0.001$). *Class-average normalized gains* were 48 and 92%, respectively. **Conclusion:** Competency-based metrics, including *class-average normalized gain*, can be used to assess course effectiveness and to determine if a program meets predesignated objectives of knowledge acquisition and psychomotor technical skill.

Copyright © 2010 S. Karger AG, Basel

Introduction

As new procedures (and devices) become increasingly available, practitioners need to become familiar with their use and potential applications. Many physicians will desire to incorporate some of them into their clinical practices. Procedural competence is traditionally gained through the apprenticeship model of subspecialty training, while enhanced knowledge and technical skill are presumed to be acquired through participation in short postgraduate courses usually comprised of didactic lectures and simulation-based hands-on procedural training.

In accordance with most postgraduate education guidelines, specific learner objectives are identified in advance, and programs are evaluated by asking participants to provide written feedback on the quality of the lectures and perceived value of the educational program. With the traditional process-based approach, course participants are presumed to acquire knowledge and technical skills by attending such programs, documented by the number of hours of course material delivered that qualify for postgraduate education credits.

From the perspective of competency-based education, skills should be teachable, learnable, and measurable [1, 2]. As opposed to process-based education, the competency-based paradigm requires that the effectiveness of education be assessed by the use of objective metrics to measure acquired skills. Although it may not be expected that all participants in a 1- or 2-day program develop competence in all aspects of a newly learned procedure, the majority of participants should, at least in the short term, significantly improve their cognitive knowledge and technical skill if the program is judged to be effective [3, 4].

To assess the effectiveness of a procedure-related postgraduate program, competency-based assessment tools can include low-stakes tests of cognitive, psychomotor (technical), and experiential knowledge. Contrary to high-stakes assessments that feature a clearly defined pass-fail threshold with significant direct consequences for the test taker, results from low-stakes assessments are used by learners to document the gradual acquisition of knowledge and skill, as well as to identify areas where improvements are desirable [5–8]. Program organizers can also use results to ascertain the strengths, weaknesses, and educational value of their curricula and to measure their success in achieving predetermined course objectives.

Diverse opinions regarding methodologies, curricular structure and measures of effectiveness, however, prompt debate among educators on topics such as the value of pre-/post-test models and the application of metrics, such as calculated percent change, also referred to as ‘gain’ [9–12], to measure student learning and determine course effectiveness. One generally accepted method for evaluating the effectiveness of an educational intervention is to calculate the $\langle g \rangle$ metric (*class-average normalized gain*) [11, 13] based on a group’s performance. Essentially, $\langle g \rangle$ is a measure of how much positive change is achieved as a result of an educational intervention, compared to the maximum possible gain if the intervention has the best-expected outcome. This is represented by the ratio of av-

erage actual gain to average maximum gain possible, and is calculated according to the formula:

$$\langle g \rangle = \text{post-test (\%)} - \text{pre-test (\%)} / 100\% - \text{pre-test (\%)}.$$

Class-average normalized gain is much less biased by a group’s pre-test score level than gain or simple comparisons of pre-/post-test results. Thus, $\langle g \rangle$ is considered to be a stronger indicator of the extent to which education is effective; when $\langle g \rangle$ is $>30\%$, the robustness of the educational intervention is inferred, regardless of the group’s baseline pre-test scores [14, 15].

The purpose of this prospective study of a 2-day postgraduate course designed to introduce thoracoscopy to pulmonary and critical-care physicians was to determine whether the short-term acquisition of cognitive knowledge and technical skills could be objectively demonstrated using a single-group, pre-/post-test model comprised of multiple-choice questions (MCQ) and psychomotor skill measures assessing procedural dexterity, speed, and accuracy. Three competency-based metrics (absolute gain, relative gain, and *class-average normalized gain*) were calculated to ascertain curricular effectiveness. A predesignated $\langle g \rangle > 30\%$ for both cognitive and technical components of the education intervention was used to declare robustness of this educational intervention.

Methods

Course Curriculum

Didactic lectures and hands-on training were offered to 17 participants as part of a 2-day structured introductory thoracoscopy course conducted at the University of California, Irvine (table 1). Participants were a self-selected group of pulmonary specialists who had responded to advertisements for this continued medical education-accredited event. None had formal training in thoracoscopy; the group’s unanimous objective was to explore the possibility of incorporating this procedure into their practices.

Day 1 included eight traditional didactic lectures and four interactive sessions. Learning was based on incremental structure of lectures where knowledge gained from one lecture could be added to a foundation established during preceding lectures. Lectures were delivered by internationally recognized experts, who also served as instructors at the hands-on technical skill stations. Day 2 included hands-on training in patient positioning (using actors), trocar insertion and flex-rigid pleuroscopy (using an Olympus inanimate thorax model), chest tube placement/suturing exercises (using the Laerdal SimMan, discarded animal parts, and silicone-suturing pads), and rigid thoracoscopy (using a series of Storz video laparoscopy-thoracoscopy simulation box stations; fig. 1).

Participants were divided into small groups of 4 persons spending 30 min at each skill station. Specific learning objectives identified for each station were provided to the participants and

Table 1. Introductory thoracoscopy course curriculum

Day 1

Didactic lectures and interactive sessions
Thoracoscopy and pleuroscopy in pleural disease management
Radiology of the pleura: interactive session
Pleural ultrasonography: interactive session
Rigid thoracoscopy: indications, patient selection, techniques, results, and complications
Flex-rigid pleuroscopy: indications, patient selection, techniques, results, and complications
Interactive session: procedure techniques and results
Anesthesia and moderate sedation
Pleurodesis for malignant pleural effusions: pleuroscopy and small-bore chest tubes
Malignant and paramalignant effusions and mesothelioma
Pleuroscopy and pleurodesis in uncomplicated recurrent benign effusions
Rigid thoracoscopy in complicated parapneumonic effusion and empyema
Interactive session: procedure techniques and results

Day 2

Hands-on training
Rigid thoracoscopy stations
Flex-rigid pleuroscopy stations
Suture stations
Live patient positioning stations
Chest tube insertion stations

each station was monitored by 2 instructors. The instructors' emphasis at each station was to keep the trainees mentally and physically engaged in the skill being practiced at that station. Deliberate practice was encouraged. After the groups had completed all skill stations, individuals were given 30 min to return to whichever stations they felt might be most likely to help them fulfill their educational needs.

Cognitive Knowledge Assessments

A 37-item, MCQ pre-test (maximum score of 37) was administered to the group prior to the first didactic lecture. A post-test (same questions, differently ordered) was administered at the end of the 2nd day. The order was intentionally changed in order to minimize risk for recall or practice bias. Answers to test questions were not provided until after the post-test was administered. MCQs were developed by one of the investigators (M.D.) who did not give any of the lectures. Because it would have been unfair to include test items that were not addressed in the lectures (scoring of such items would be a measure of the group's pre-course knowledge level, rather than a measure of what was learned), questions were designed after reviewing lectures submitted by instructors before the course. This was done to prevent 'teaching to the test'.

Technical Skill Assessments

On day 2, before skill station training, participants were asked to perform a complex task using a rigid thorascopic grasping forceps to pick up eight 1-cm plastic rings of two different colors

Fig. 1. Video laparoscopy-thoracoscopy simulation box trainer (Storz) used for dexterity, accuracy, and time assessment.

Fig. 2. Using the Storz station, speed, accuracy, and dexterity were practiced using rigid thoracoscopy grasping forceps to retrieve eight 1-cm plastic rings and place them on matching colored pins. Foam padding in the center of the model was used to practice biopsy using thorascopic cutting biopsy forceps.

and place them one at a time onto matching colored pins (fig. 2) inside the Storz simulation station. A maximum of 3 min was allowed to complete the task, which had a maximum possible score of 8. This specific task was not practiced at the skill station sessions. For the second assessment, the same task was performed at the end of day 2, this time using a 2-min time limit to achieve the maximum possible score of 8. The reduced time was intended to minimize the ceiling effect for post-test scores, whereby increased speed gained from training might lead most learners to achieve

Fig. 3. Pre- and post-test cognitive knowledge scores (n = 17).

Fig. 4. Pre- and post-test technical skill scores (n = 17).

the maximum possible score. All tests were scored anonymously by independent scorers who were not study investigators.

Statistical Methods

This was a quasi-experimental, one-group pre-/post-test study designed to assess learning gain and robustness of this course curriculum. Individual actual gains (post-test score – pre-test score) were tabulated in order to calculate mean change scores for the class expressed as absolute gain (average actual gain/maximum score achievable), and relative gain (average actual gain/pre-test score) expressed as a percentage. Paired-samples t test with an α of 0.05 was used to compare pre- and post-test scores for both cognitive and technical skill components. Results were confirmed using Wilcoxon rank sum non-parametric testing.

As an additional measure of curricular effectiveness, *class-average normalized gain* ($\langle g \rangle$) was calculated; $\langle g \rangle$ is defined as the average actual gain divided by the average maximum possible gain. The robustness of the educational intervention was determined using the predefined threshold of $\langle g \rangle > 30\%$ [15, 16]. The study was exempted from review of the Institutional Review Board at the University of California, Irvine.

Results

All 17 participants completed cognitive knowledge and technical skill pre- and post-tests. The group's mean cognitive knowledge score improved from 20.9 to 28.7 (7.8 ± 1.3 points, $p < 0.001$), representing a relative gain of 37% and an absolute gain of 21% (fig. 3). Mean techni-

cal skill scores improved from 5.20 to 7.82 (2.62 ± 0.33 , $p < 0.001$), representing a relative gain of 50% and an absolute gain of 33%, even with the 33% reduction in time allotted for the post-test task completion (fig. 4). Non-parametric testing confirmed t test results ($p < 0.001$ for both cognitive and technical skill tests). *Class-average normalized gain* was 48% for cognitive knowledge and 92% for technical skills.

Discussion

In accordance with the Accreditation Council for Graduate Medical Education Outcome Project [16], the focus of physician education in the United States is shifting away from process-oriented measures, such as the number of lecture hours attended or number of invasive procedures performed, to a competency-based model whereby learner performance is measurable using competency-based metrics (cognitive knowledge and technical skill assessments) and patient outcomes. We agree with Norman [17], however, that while continued medical education can improve knowledge and application of knowledge, 'pursuing patient outcomes amounts to looking in bigger and bigger haystacks for smaller and smaller needles'.

Procedure-related postgraduate programs with specific predesignated objectives, including determination of curricular robustness, provide excellent opportunities for competency-based education. Because it makes little sense to pursue technical performance gains in the absence of knowledge changes, we measured the effectiveness of our curriculum in meeting two distinct and predetermined objectives: (1) to increase cognitive (factual) knowledge of thoracoscopy, with reference to indications, contraindications, patient safety, techniques, and patient selection; and (2) to augment technical skill as measured by improved dexterity, accuracy, and speed in the manipulation of thoracoscopic instruments while performing a complex task using a video laparoscopy-thoracoscopy simulation box station.

The effectiveness of our educational intervention, independent of the study group's pre-test level of knowledge, was supported for these two objectives using measures of *class-average normalized gain* ($\langle g \rangle$ of 48% on cognitive knowledge testing and 92% on technical skill tests). Educational research has shown that for courses with widely varying average pre-test scores, $\langle g \rangle$ is nearly independent of baseline average pre-test score, and primarily dependent on the effectiveness of the instruction [18]. A $\langle g \rangle$ value $>30\%$ is an accepted marker of a robust educational intervention for students participating in 'heads-on and hands-on activities which yield immediate feedback through discussion with peers and/or instructors' [19, 20].

Using pre-/post-test models to ascertain curricular effectiveness raises several concerns that can be extrapolated to other fields of medical and surgical postgraduate education. Methodologically, for example, one might suppose that without a control group we cannot say whether the participants' improvement would have occurred even without the educational intervention. The design and execution of an educational study using a parallel group that is deprived of all educational material, however, is illogical. Furthermore, the use of normalized gain diminishes the noise and confounders of pre-test knowledge and group characteristics, thereby decreasing the need for a comparison group [18, 21]. In order to further obviate the need for a control group, the curriculum as well as the pre- and post-tests were all administered on 2 consecutive days. It is not plausible that the significant gains seen over such a short time span could have occurred without the intervention. Hence, because of its short duration, our educational intervention was immune to many of the external factors that could otherwise threaten the validity of a single-group pre-/post-test design.

Improvements demonstrated through increases in post-test scores may not prove the effectiveness of a course because pre-/post-test models are only as good as the tests employed. In the absence of national databanks, it is neither practical nor cost-effective for most postgraduate course organizers to create valid and consistently reliable test questions. In the realm of pulmonary procedures, validated assessment tools are only now being proposed for bronchoscopy [8, 22] and none exists for thoracoscopic procedures.

As stated above, the short duration of our educational intervention makes it unlikely that results were impacted by extraneous variables such as history or maturation, or be subject to issues of retention and decay that are inherent in most measures of knowledge acquisition in social research [23]. Practice effect, another factor that might threaten the validity of a pre-/post-test study design was unlikely in our study because the order of post-test questions was different, lecturers blinded to the test questions had no opportunity for 'teaching to the test', and the specific complex technical skill used for the testing was not practiced during hands-on sessions. The effect of deliberate practice of a specific skill set, compared with random or independent practice, however, could be the subject of future investigations.

Our study has several limitations. First, we focused only on dexterity, accuracy, and speed as a reflection of thoracoscopic technical skill ascertained using an inanimate box trainer. Surgical training using such models has been shown to translate to improved performance in patients [24]. Second, we did not perform technical skill assessments at all the hands-on stations, even though each of them was potentially amenable to an objective and reproducible test of that specific procedure-related skill. This would have detracted participants from the objectives of the exercise, and been logistically unrealistic. Time constraints represent a considerable hurdle to most explorations of educational modalities of short postgraduate courses, where greatest emphasis is placed on ensuring maximum learning rather than on assessing curricular effectiveness. Third, our sample size was small and self-selected, reflecting the reality of most highly specialized courses where the number of attendees seldom exceeds thirty. Lower attendance, however, means more time at each skill station and thus may be beneficial for learning. Even with this small sample size, our results demonstrated statistical significance ($p < 0.001$) and educational significance ($\langle g \rangle > 30\%$ for both knowledge and skill improvement). Finally, as in many areas of medical education, the impact

of this educational intervention on clinical practice and outcomes is unknown because short-term improvements in cognitive knowledge and technical skill might not translate to long-term retention without continued reinforcement and practice [25]. Long-term retention versus decay depends on numerous confounding variables, including but not limited to incentive for initial training, opportunities for repetition in the practice setting, incorporating the procedure into one's clinical practice, and motivation [26].

Conclusion

Participation in short postgraduate programs, while not the ideal strategy for learning [27], remains a popular means by which practitioners become familiar with new instruments, procedures, and techniques. Because physicians incorporate these newly learned procedures into their clinical practice, concerns for patient safety warrant a change in training methodologies from being solely

process oriented to one that is competency based [28, 29]. This study of a 2-day introductory thoracoscopy course for a small group of chest physicians provides just one example of how competency-based metrics, including *class-average normalized gain*, can be used to measure course effectiveness and determine if a program meets predesignated objectives of knowledge acquisition and psychomotor technical skill. Points raised by this study can also be extrapolated to other areas of minimally invasive surgery and thus be of interest to medical educators working within the competency-based paradigm of procedural training.

Acknowledgments

We thank our instructors, assistants, and physicians who graciously participated in this course. We are grateful to the Laerdal Corporation, the Olympus Corporation, Karl Storz Corporation, and Ethicon Corporation for their participation.

References

- 1 Miller GE: The assessment of clinical skills/competence/performance. *Acad Med* 1990; 65(suppl):S63–S67.
- 2 Brasel KJ, Bragg D, Simpson DE, Weigelt JA: Meeting the Accreditation Council for Graduate Medical Education competencies using established residency training program assessment tools. *Am J Surg* 2004;188:9–12.
- 3 Long DM: Competency-based residency training: the next advance in graduate medical education. *Acad Med* 2000;75:1178–1183.
- 4 Apeltgren K: ACGME e-Bulletin. August 2006, ACGME Competencies. <http://www.acgme.org> (accessed June 1, 2009).
- 5 Davoudi M, Colt HG: Bronchoscopy simulation: a brief review. *Adv Health Sci Educ Theory Pract* 2009;14:287–296.
- 6 Colt HG, Davoudi M, Quadrelli S: Pilot study of web-based bronchoscopy education using the Essential Bronchoscopist in developing countries (Mauritania and Mozambique). *Respiration* 2007;74:358–359.
- 7 Crawford SW, Colt HG: Virtual reality and written assessments are of potential value to determine knowledge and skill in flexible bronchoscopy. *Respiration* 2004;71:269–275.
- 8 Quadrelli S, Davoudi M, Galíndez F, Colt HG: Reliability of a 25-item low-stakes multiple choice assessment of bronchoscopic knowledge. *Chest* 2009;135:315–321.
- 9 Hake RR: Should we measure change? Yes! <http://www.physics.indiana.edu/~hake/MeasChangeS.pdf> (accessed June 1, 2009).
- 10 Cronbach LJ, Furby L: How should we measure 'change' or should we? *Psychol Bull* 1970;74:68–80.
- 11 Meltzer DE: The relationship between mathematics preparation and conceptual learning gains in physics: a possible 'hidden variable' in diagnostic pretest scores. *Am J Phys* 2002;70:1259–1267.
- 12 Hovland CI, Lumsdaine AA, Sheffield FD: A baseline for measurement of percentage change; in *Experiments on Mass Communication*. New York, Wiley, 1949; reprinted in Lazarsfeld PF, Rosenberg M (eds): *The Language of Social Research: A Reader in the Methodology of Social Research*. New York, Free Press, 1955, pp 77–82.
- 13 Wise SL, DeMars CE: Examinee motivation in low-stakes assessment: problems and potential solutions. Seattle, Annual Meeting of the American Association of Higher Education Assessment Conference, 2003.
- 14 Hake RR: Lessons from the physics education reform effort. *Conserv Ecol* 2002;5:28. <http://www.ecologyandsociety.org/vol5/iss2/art28/> (accessed June 1, 2009).
- 15 Melzer DE: Normalized learning gain: A key measure of student learning. (Addendum to: Melzer DE: The relationship between mathematics preparation and conceptual learning gains in physics: a possible 'hidden variable' in diagnostic pretest scores. *Am J Phys* 2002;70:1259–1267). <http://scitation.aip.org/getpdf/servlet/GetPDFServlet?filetype=pdf&id=AJPIAS000070000012001259000001&idtype=cvips> (accessed June 1, 2009).
- 16 ACGME Outcome Project: Enhancing residency education through outcomes assessment. Accreditation Council for Graduate Medical Education. Chicago, 2000, <http://www.acgme.org/outcome/comp/General-CompetenciesStandards21307.pdf>; <http://www.acgme.org/outcome/Comp/compFull.asp> (accessed June 1, 2009).
- 17 Norman G: The American College of Chest Physicians evidence-based educational guidelines for continuing medical education interventions. *Chest* 2009;135:834–837.
- 18 Hake RR: Interactive engagement vs traditional methods: a six-thousand student survey of mechanics test data for introductory physics courses. *Am J Phys* 1998;66:64–74.
- 19 Redish EF, Steinberg RN: Teaching physics: figuring out what works. *Phys Today* 1999; 52:24–30.

- 20 Hobson R, Rolland S, Rotgans J, Schoonheim-Klein M, Best H, Chomyszyn-Gajewska M, Dymock D, Essop R, Hupp J, Kundzina R, Love R, Memon RA, Moola M, Neumann L, Ozden N, Roth K, Samwel P, Villavicencio J, Wright P, Harzer W: Quality assurance, benchmarking, assessment and mutual international recognition of qualifications. *Eur J Dent Educ* 2008;12(suppl 1):92–100.
- 21 Hake RR: Suggestions for administering and reporting pre/post diagnostic tests. <http://www.physics.indiana.edu/~hake/Testing-Suggestions051801.pdf> (accessed June 1, 2009).
- 22 Goldberg R, Colt HG, Davoudi M, Cherrison L: Realistic and affordable lo-fidelity model for learning bronchoscopic transbronchial needle aspiration. *Surg Endosc* 2009;23:2047–2052.
- 23 Campbell D, Stanley J: *Experimental and Quasi-Experimental Designs for Research*. Chicago, Rand-McNally, 1963.
- 24 Abastakis DJ, Regehr G, Rexnick RK, Cusimano M, Murnaghan J, Brown M, Hutchison C: Assessment of technical skills transfer from the bench model to the human model. *Am J Surg* 1999;177:167–170.
- 25 Willingham DB, Dumas JA: Long-term retention of a motor skill: implicit sequence knowledge is not retained after a one-year delay. *Psychol Res* 1997;60:113–119.
- 26 Wise VL, Wise SL, Bhola DS: The generalizability of motivation filtering in improving test score validity. *Educ Assess* 2006;11:65–83.
- 27 Davis DA, Thomson MA, Oxman AD, et al: Evidence for the effectiveness of CME. A review of 50 randomized controlled trials. *JAMA* 1991;268:1111–1117.
- 28 Carraccio C, Wolfsthal SD, Englander R, et al: Shifting paradigms: from flexner to competencies. *Acad Med* 2002;77:361–367.
- 29 Michelson JD, Manning L: Competency assessment in simulation-based procedural education. *Am J Surg* 2008;196:609–615.