Surg Endosc (2011) 25:207–216
DOI 10.1007/s00464-010-1161-4

Measuring learning gain during a one-day introductory bronchoscopy course

Henri G. Colt · Mohsen Davoudi · Septimiu Murgu · Nazanin Zamanian Rohani

Received: 2 February 2010 / Accepted: 23 May 2010 / Published online: 29 June 2010

Abstract
Background Rigorous assessment of medical knowledge and technical skill inspires learning, reinforces confidence, and reassures the public. Identifying curricular effectiveness using objective measures of learning is therefore crucial for competency-oriented program development in a learner-centric educational environment. The aim of this study was to determine whether various measures of learning, including *class-average normalized gain*, can be used to assess the effectiveness of a one-day introductory bronchoscopy course curriculum.
Methods We conducted a quasi-experimental one-group pre-test/post-test study at the University of California, Irvine. The group comprised 24 first-year pulmonary and critical care trainees from eight training institutions in southern California. Class-average normalized gain, single-student normalized gain, absolute gain, and relative gain were used as objective measures of cognitive knowledge and bronchoscopy technical skill learning. A class-average normalized gain of 30% was used to determine curricular effectiveness. Perceived educational value using Likert-scale surveys and post-course questionnaires was determined during and 3 months after course participation.
Results Mean test scores of cognitive knowledge improved significantly from 48 to 66% ($p = 0.043$). Absolute gain for the class was 18%, relative gain was 37%, *class average normalized gain* ⟨g⟩ was 34%, and the *average of the single-student normalized gains* g(ave) was 29%. Mean test scores of technical skill improved significantly from 43 to 77% ($p = 0.017$). Absolute gain was 34%, relative gain was 78%, *class average normalized gain* ⟨g⟩ was 60%, and the *average of the single-student normalized gains* g(ave) was 59%. Statistically significant improvements in absolute gain were noted in all five elements of technical skill ($p < 0.05$). Likert-scale surveys, questionnaires, and surveys demonstrated strong perceived educational value.
Conclusion The effectiveness of a one-day introductory bronchoscopy curriculum was demonstrated using a pre-test/post-test model with calculation of normalized gain and related metrics.

Keywords Pre–post testing · Normalized gain · Competency · Training · Bronchoscopy · Education

Learning bronchoscopy in the clinical setting promotes learner anxiety, exposes patients to the burden of procedure-related education [1], and results in a highly variable learning experience [2]. Clinical responsibilities often interfere with reading bronchoscopy-related material, and, in the absence of periodic assessments of bronchoscopic knowledge, trainees are unlikely to be compliant with educational endeavors they perceive as optional, especially if there are no pass/fail grading consequences [3]. Furthermore, the current bronchoscopy learning environment is less than ideal for beginners because of concerns for patient safety, fiscal constraints, and an increasing impetus to document procedural competency [4, 5].

Short postgraduate courses comprising lectures and simulation-based hands-on instruction have thus become a popular means for enhancing procedure-related learning [6,

Electronic supplementary material The online version of this article (doi:10.1007/s00464-010-1161-4) contains supplementary material, which is available to authorized users.

H. G. Colt (✉) · M. Davoudi · S. Murgu · N. Zamanian Rohani
Pulmonary and Critical Care Medicine, University of California, 333 City Drive West, Suite 400, Orange, CA 92868, USA
e-mail: hcolt@uci.edu

7]. In accordance with continuing medical education (CME) guidelines, these programs identify learner objectives and provide opportunities for feedback from students regarding the perceived quality of the course. Yet, to our knowledge, no study has objectively measured how much knowledge and skill students actually gain as a result of participating in such a program.

It is generally difficult to prove that a learning gain has occurred *as a result of* an educational intervention. This difficulty is due partly to controversies regarding the comparative use of pre- and post-test assessments and partly because of problems constituting control groups with which studies of an educational intervention can be compared [8–11]. Causality is also difficult to prove due to the confounding effects of retention and decay, normal maturation, and possible ongoing training [12].

Many educators in biological science, engineering, astronomy, mathematics, and physics have turned to using *class-average normalized gain* and related metrics to gauge a course's effectiveness [13–16]. *Class-average normalized gain* (known as $\langle g \rangle$) measures the ratio of a whole group's performance to the maximum achievable improvement. It is expressed mathematically as a fraction of maximum achievable pre-test/post-test gain [17]. Educators use this measure of performance to diminish the confounding effects of pre-course knowledge and other baseline group characteristics, thereby decreasing the need for a control group [18].

In this study we used a pre-test/post-test model to assess the effectiveness of a one-day introductory bronchoscopy course curriculum. We demonstrated how various metrics of learning, including *class-average normalized gain*, can measure the acquisition of a procedural skill such as bronchoscopy. We hypothesized that the educational intervention would result in significant class and individual student gains in bronchoscopic cognitive knowledge and technical skill.

Methods

Participants

Twenty-four first-year pulmonary and critical care fellows (from now on referred to as students) and four program directors from eight training institutions in southern California participated in this study.

Course curriculum

A one-day curriculum comprising educational content, learner assessments, and teaching guides [19] was designed. It incorporated multiple components (visual, auditory, verbal, role-playing, and analytic) in order to create intentional redundancy (see supplementary Appendix 1). Educational content was composed in modular fashion using different educational media and techniques [20]; these included ten classroom-based didactic lectures, three interactive audience-participation question-and-answer (Q/A) sessions, four small-group hands-on technical skill workshops using low- and high-fidelity simulation technology, and one clinical problem-solving case scenario. Students received a syllabus containing lectures, learning objectives, and a CD-ROM of the web-based *Essential Bronchoscopist*©, and *Bronchoscopy Step-by-Step*© exercises [21].

Lectures were structured to reinforce incremental learning using a combination of didactic, interactive, Q/A, and anonymous audience participation techniques. For the small-group learning sessions, the 24 students were randomly divided into five groups that rotated through four technical skill stations (airway anatomy and bronchoscopic inspection, endobronchial brushing and biopsy, transbronchial needle aspiration, and emergent bronchoscopic intubation) and one case scenario, spending 30 min at each station (2.5 h total). Specific learning objectives were designed to keep the students mentally and physically engaged so as to achieve "hands-on and heads-on" learning [13]. Teaching guides were distributed to instructors so that they could function as professional learning coaches rather than as experts merely stating facts and opinions [22].

Cognitive knowledge assessments

A multiple-choice question (MCQ) test of cognitive knowledge was developed by the authors. It comprised 40 items evenly divided into a pre-test and post-test, each containing 20 questions (the maximum score for each test was 20). Three of the authors (HC, MD, SM) had each written 25 Q/A sets based on information included in the web-based *Essential Bronchoscopist*©; the sets were found in a prior study [23] to be necessary or absolutely necessary for a test of bronchoscopic knowledge. The resulting 75 questions were pilot-tested using a group of seven second- and third-year University of California Irvine (UCI) pulmonary and critical care fellows; questions with extreme high or low difficulty indices ($\rho > 80\%$ or $<20\%$ correct response rate) were eliminated. The difficulty index or ρ of a test item is defined as the proportion of a group of test-takers who gets that item wrong, hence $\rho = 90\%$ is a very difficult question while $\rho = 10\%$ is a very easy one. Questions that were very similar in material or had controversial responses were also eliminated.

The 40 best questions were then assigned semirandomly to 24 sets of two tests (items 1–20 for pre-test, and 21–40 for post-test). Using a random numbers table, items were

randomly ordered from 1 to 40, 24 consecutive times for the 24 students. Each set of 1–20 and 21–40 was reshuffled to assure that the pre-test and post-test for each student had an equal number of difficult ($0.20 < \rho \leq 0.40$), intermediate ($0.40 < \rho \leq 0.60$), and easy ($0.60 < \rho \leq 0.80$) questions. Tests were normalized for topic-specific questions. For example, among four questions on BAL (bronchoalveolar lavage), two were in the first 20 and two in the second 20 for each of the 24 students. By using this semirandom process, each student received a different test, yet the tests were similar in level of difficulty and material covered. To assure that each student received the two halves of the same test, the tests were marked with random two-letter codes (AA, BB, CC, etc.), printed on each page. The same code was used on the students' identity badges and skills tests sheets, assuring anonymous processing of all data.

Technical skills assessments

An abbreviated version of the Bronchoscopic Skills and Tasks Assessment Tool—BSTAT© (Fig. 1) [24] was used to test individuals in a low-fidelity airway model (see supplementary Appendix 2). Students were asked to navigate the bronchoscope from the larynx to two designated segments (superior segment, left lower lobe; mediobasal segment, right lower lobe). After delivery of uniform instructions, and with one-on-one supervision from an independent instructor, students were allowed up to 45 s for each segment and were scored on five criteria: time, anatomic recognition, precision, economy of movement, and posture/hand position. To minimize inter-rater variability, the four instructors/testers performing the assessments had previously practiced scoring during a 2-h session, assuring similar definitions and parameters for the criteria being tested in test subjects. At the end of the 2-h session, the last three test samples were scored almost identically by all four testers.

Study protocol

Immediately following on-site registration, each student was asked to complete a short questionnaire about prior experience and perceptions regarding bronchoscopy and simulation. Individual students then underwent technical skill pre-testing, followed by administration of the multiple-choice pre-test (20 min was allowed, although no one needed the allotted time). During the day, students and program directors were asked to rate the educational content and quality of each learning session using a Likert-scale survey instrument. Post-testing was done at the end of the day after course completion. While technical skill post-testing was performed identically, the written post-test consisted of the second half of each student's specific MCQ test (e.g., student GG received GG-1–20 as pre-test and GG-21–40 as post-test).

Three months after the educational intervention, a four-item questionnaire was emailed to each student, and an independent psychologist research interviewer conducted telephone interviews with the four program directors to inquire about the perceived impact of the course on student performance, perceived weaknesses of the course, and whether the program directors would support similar courses in the future.

Outcome assessments

We hypothesized that participation in this educational intervention 3 months into subspecialty training would result in improved class and individual student learning gains in cognitive knowledge and technical skill. For the purpose of this study, curricular effectiveness was defined as the extent to which the program produced learning gains for the group and for individual students as demonstrated by several metrics of student performance.

A multiple-choice pre-test/post-test model of cognitive knowledge acquisition and a pre-test/post-test time-limited, low-fidelity simulation technical skill exercise were used to identify knowledge and skill gains. Pre-test and post-test scores were compared and the effectiveness of the educational intervention for the group as a whole was determined using a predefined target objective of at least 0.3 for *class average normalized gain* $\langle g \rangle$. This criterion is taken from Hake [13] whose 62-course $\langle g \rangle$ data suggested that $\langle g \rangle = 0.3$ (30%) is a lower bound of what he designated as "medium" normalized gain. Individual single-student g_i values to identify variability between individuals, and the average of *single-student normalized gain*, g(ave), were also calculated.

To complement these objective measurements, the program's perceived educational value was assessed by analyzing qualitative evaluations and Likert-scale survey measurements from each participant on-site and by reviewing results from post-course student questionnaires and recorded interviews of program directors 3 months after the educational intervention. Quantitative and qualitative analyses were performed on interview transcripts.

To determine the potential reproducibility of study results in a different cohort tested at a similar time period during the course of their subspecialty training (3 months into training), the same curriculum was delivered 1 year later to a fresh group of first-year pulmonary subspecialty trainees invited from the same institutions. Statistical analyses were performed on pre-test/post-test results using the same metrics. The study was considered exempt from UCI Institutional Review Board review.

Testing Session I			Testing Session II		Change
Time	_____ sec.		Time	_____ sec.	Delta T ___
Anatomic Recognition	0 ½ 1		Anatomic Recognition	0 ½ 1	_____
Precision	0 ½ 1		Precision	0 ½ 1	_____
Economy of Movement	0 ½ 1		Economy of Movement	0 ½ 1	_____
Posture Hand Pos.	0 ½ 1		Posture Hand Pos.	0 ½ 1	_____
Total I			Total II		Total Change

Anat. Rec.: 1 - Recognized anatomy, and went to target segment on first try
½ - Recognized anatomy, and went to target segment on second try
0 - Could not find target segment during allotted time

Precision: 1 - Scope always centered; no episodes of red-out or scraping airway wall
½ - Scope mostly centered, <= 3 episodes of red-out or scraping airway wall
0 - Scope not centered, > 3 episodes of red-out or scraping airway wall

EoM: 1 - Navigates directly to segment, with confidence and intent
½ - Does not navigate with full confidence, has doubts, stops at other segments
0 - Enters 2 or more other segments

Post/Hands.: 1 - Comfortable with bronchoscope, able to adjust hand positions
½ - Corrects hand positions and posture upon prompting
0 - Cannot use correct hand positions and posture even upon prompting

Fig. 1 Bronchoscopy technical skills pre-test and post-test score sheet

Statistical methods

This was a quasi-experimental, one-group pre-test/post-test study designed to assess learning gain and effectiveness of this single-day introductory bronchoscopy course curriculum [25]. Paired-samples t test with an $\alpha = 0.05$ was used to compare pre- and post-test scores, which were also plotted for descriptive purposes. A dissimilar pre-test/post-test study design was used to diminish the biasing effect. Individual actual gains G_i (where G_i = post-test score − pre-test score) were tabulated in order to calculate percent absolute gain (where Δ = average G_i/maximum score achievable), and percent relative gain, expressed as a percentage (where C = average G_i/pre-test score) for the class. As a measure of course effectiveness, the *class average normalized gain* $\langle g \rangle$ was calculated. The $\langle g \rangle$ is defined as the average actual gain divided by the maximum possible gain, where G is the actual gain and $\langle \%post \rangle$ and $\langle \%pre \rangle$ are the final (post) and initial (pre) class averages, and the angle brackets "$\langle \ldots \rangle$" indicate an average of the students taking the tests:

$$\langle g \rangle = \langle \%G \rangle / \langle \%G \rangle_{max}$$

$$\langle g \rangle = [\langle \%\text{post-test} \rangle - \langle \%\text{pre-test} \rangle] / [100\% - \langle \%\text{pre-test} \rangle]$$

A predefined target $\langle g \rangle$ of 30% was taken as defining the minimum value at which the educational intervention could be regarded as effective [13, 26].

In addition, *individual single-student normalized gains* (g_i) were calculated for all students and averaged as:

$$g(ave) = \left[\sum_{\text{from 1 to } N} (g_i) \right] / N$$

where N is the number of trainees taking both the pre- and post-tests. Because of the possibility of post-test scores inferior to pre-test scores (negative gain), g(ave) was also calculated by replacing negative g_i with zero and by deleting all negative-gain students. In physics education research, it has been found that the first two averages $\langle g \rangle$ and g(ave) are usually the same or within 5% for $N > 20$ and that this near equality is associated with a low correlation of the single-student gain (or g_i) with the single-student pre-test score [13].

Results

Twenty-four of 25 eligible first-year pulmonary and critical care trainees and four program directors from eight training institutions in southern California participated in this study. The director from UCI was not included to avoid obvious bias. Twenty-one students (88%) had assisted in 30 or fewer flexible bronchoscopies; the same percentage envisioned that bronchoscopy was a strong component of their future career. Two thirds (16/24) of the students were exposed to some form of medical simulation during residency training, of which nine had been previously exposed to a bronchoscopy simulator. None of those who had been exposed to medical simulation during residency training (0/16) reported a negative experience.

All 24 students took both the pre-test and the post-test (Fig. 2A, B). Mean test scores of cognitive knowledge improved significantly from 48% (9.6/20 ± 2.58) to 66% (13.2/20 ± 2.53) ($p = 0.043$). Absolute gain was 18% (3.5/20 ± 3.7) and relative gain was 37% (Fig. 3A). The *class average normalized gain* ⟨g⟩ was 34%, and the *average of the single-student normalized gains* g(ave) was 29% (SD ± 33) (Table 1).

Mean test scores of technical skill also improved significantly from 43% (6.9/16 ± 2.91) to 77% (12.4/16 ± 3.33) ($p = 0.017$). Absolute gain for the class was 34% (5.5/16 ± 3.7) and relative gain was 78% (Fig. 3A). The *class average normalized gain* ⟨g⟩ for technical skills was 60%, and the *average of the single-student normalized gains* g(ave) was 59% (SD ± 39) (Table 1). Statistically significant absolute gains were noted in all five elements of technical skill ($p < 0.05$); time (−29%), precision (27%), anatomic recognition (42%), posture/hand position (24%), and economy of movement (42%) (Fig. 3B).

Likert-scale surveys for cognitive learning sessions received mean scores ranging from 4.65/5 to 4.94/5 (5 was the best score attainable). Likert-scale scores for each of the technical skill stations ranged from 4.81/5 to 5/5. Perceptions of the educational program assessed 3 months later were based on a four-item questionnaire. The response rate was 75% (18/24 students). All but one person (17/18) stated they would recommend participation in this course to the following year's incoming trainees. Almost all (14/18) said the course had a very positive impact on their skills and performance. Fifteen said their senior colleagues in training had shown strong enthusiasm when asked if they would like to participate in a similar course. When asked for suggestions regarding future courses, most (16/18) requested more time at the skill stations or a program lasting more than one day (Table 2).

In their Likert surveys, all four program directors scored each of the cognitive and technical skill stations 5/5 at the time of the course. Follow-up interviews 3 months later

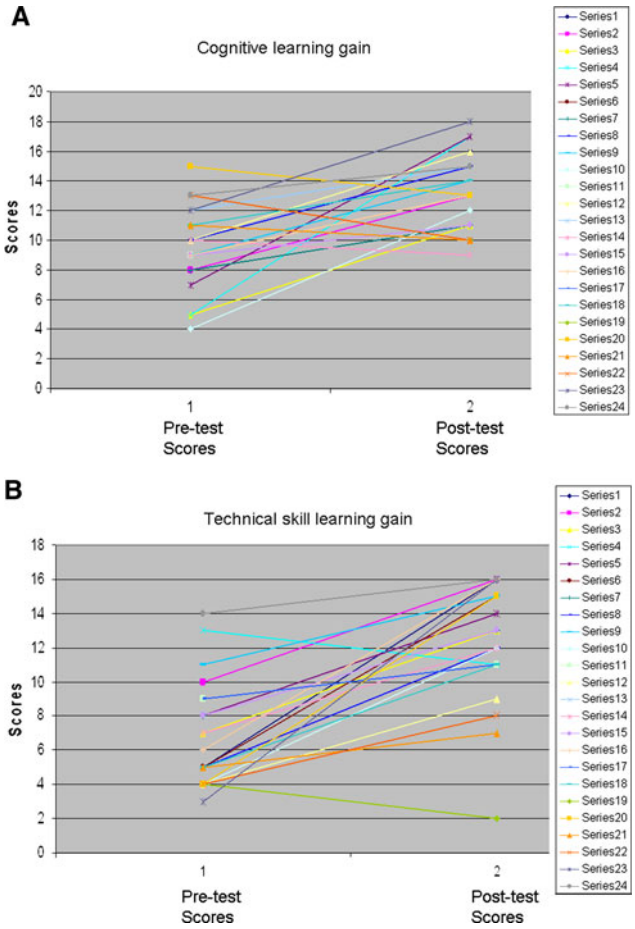

Fig. 2 **A** Cognitive learning pre- and post-test score plots showing improvement (positive slopes), no change (horizontal lines), or deterioration (negative slopes); 1 = pre-test, 2 = post-test. Maximum score was 20 for cognitive knowledge assessments. **B** Technical skill learning pre- and post-test score plots showing improvement (positive slopes), no change (horizontal lines) or deterioration (negative slopes); 1 = pre-test, 2 = post-test. Maximum score was 16 for technical skill assessments

revealed unanimity regarding the positive effect of this course on their trainees' bronchoscopy skills and performance. The directors' opinions were based on feedback obtained from trainees participating in the course, as well as on direct observation of their trainees' ability to perform bronchoscopy during the intervening 3 months. When asked if the program had any weaknesses, two said that it was "too-information-dense" for one day and one recommended greater emphasis on bronchial anatomy. All four program directors said they wanted their trainees to attend a similar course the following year.

For the cohort one year later (comprising 18 first-year pulmonary and critical care trainees), baseline knowledge and technical skill were similar to the earlier cohort. All measures of learning gain were again significantly increased, thereby corroborating findings from the initial

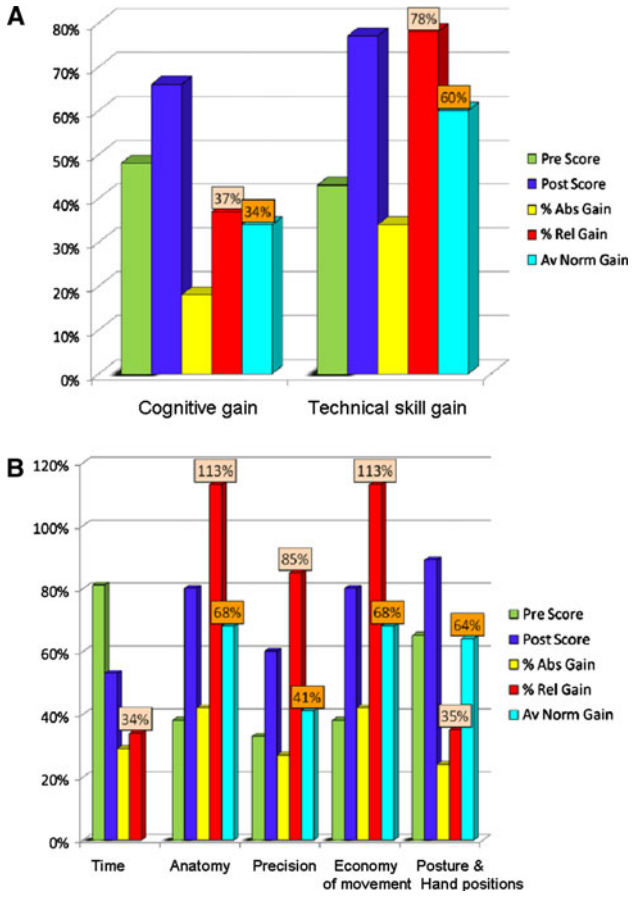

Fig. 3 **A** Cognitive and technical skill average learning gains with pre- and post-test scores (%) during one-day introductory bronchoscopy course ($N = 24$). **B** Average learning gains of five components of bronchoscopy technical skill with pre- and post-test skill scores (%). All changes statistically significant ($p < 0.05$)

cohort (Table 3). Mean scores of cognitive knowledge significantly increased from 39% (7.8/20 ± 2.4) to 66% (13.1/20 ± 2.5) ($p = 0.021$). Absolute gain for the class was 27% (5.3/20 ± 2.9) and relative gain was 69%. No negative gains were noted. The *class average normalized gain* ⟨g⟩ was 44%. The *average of the single-student normalized gains* g(ave) was 60% (SD ± 21).

Mean scores for technical skill also increased significantly from 41% (6.6/16 ± 3.2) to 76% (12.2/16 ± 2.4) ($p = 0.015$). Absolute gain for the class was 35% (5.6/16 ± 2.9) and relative gain was 85% (Table 3). Again, no negative gains were noted. The *class average normalized gain* ⟨g⟩ for technical skills was 60%, which, along with the large ⟨g⟩ for cognitive knowledge, confirmed the effectiveness of the educational intervention. The *average of the single-student normalized gains* g(ave) was 60% (SD = ±22).

Discussion

The competency-based paradigm is today's prevalent educational model. It warrants that procedure-related learning should lead to a level of verifiably measurable knowledge and technical skill [27, 28]. From an educator's perspective, identifying curricular effectiveness using objective measures of student learning that demonstrate gains in knowledge and skill is an important element of competency-oriented program development. Demonstrating learning gain can be viewed as analogous to the number-needed-to-treat metric that is required to prove the efficacy of new therapies [6]. It is often extremely difficult to prove the direct beneficial impact of educational interventions on clinical care. Yet measurements of curricular effectiveness can identify strengths and shortcomings of an educational intervention and help delineate the individual student's progress along the competency curve between novice, advanced beginner, proficient provider, and competent provider [29].

In this study we used a pre-test/post-test assessment model, several measures of learning gain, Likert-scale analyses, and post-intervention surveys to study the curricular effectiveness and perceived educational value of a

Table 1 Pre- and post-test scores and learning gain ($N = 24$)

	Pre-test scores	Post-test scores	p value	Absolute gain	Relative gain	⟨g⟩	g(ave)
Cognitive learning	9.6 ± 2.5 48%	13.2 ± 2.5 66%	0.043*	18%	37%	34%[a]	29% ± 33[b]
Technical skills learning	6.9 ± 2.9 43%	12.4 ± 3.3 77%	0.017	34%	78%	60%[a]	59% ± 39[c]

⟨g⟩ is the *class average normalized gain*: ⟨g⟩ = [⟨% post-test⟩ − ⟨% pre-test⟩]/[100% − ⟨% pre-test⟩]

g(ave) is the *average single-student normalized gain*: g(ave) = $[\sum_{\text{from 1 to } N} (g_i)]/N$

* $p < 0.05$

[a] Robustness of educational intervention defined if ⟨g⟩ is greater than 30%

[b] g(ave) calculated with all negative g_i's replaced by 0 was 35% ± 25 and g(ave) calculated with all negative g_i's ignored was 46% ± 17

[c] g(ave) calculated with all negative g_i's replaced by 0 was 63% ± 31and g(ave) calculated with all negative g_i's ignored was 68% ± 25

Table 2 Four-item questionnaire to assess educational value for pulmonary trainees 3 months after course participation ($n = 18$)

Question	Selected comments
Q-1: How would you rate the impact of the course on your skills and performance since you attended the course 3 months ago?	Good literature review[a]
	My skills/confidence improved greatly[a]
	Useful/essential review of information
	Good combination of theory and practical[a]
	Good technical skills[a]
Q-2: When your senior fellows (2nd and 3rd year) learn about your experience at the course, have they shown enthusiasm for a similar course?	They wished to have this opportunity as first fellows[a]
	They were enthusiastic/excited about having a potential course for them[a]
	They were annoyed at having to cover for us
	They would like to have more education regarding bronchoscopy[a]
Q-3: Would you encourage next year's 1st year fellows to partake in the program?	It is an excellent teaching and learning opportunity
	It was definitely worth the time[a]
	Great way to start an introduction to bronchoscopy before actual procedures[a]
Q-4: Do you have any other suggestions?	More than one station for the basic anatomy[a]
	Course should be more than one day[a]
	Need to spend more time teaching about the procedures
	More hands-on time learning and practicing the basic airway anatomy[a]

[a] Comment repeated by more than 30% of respondents

Table 3 Pre-/post-test scores and learning gain for two cohorts

	Pre-test scores	Post-test scores	Absolute gain	Relative gain	$\langle g \rangle$	g(ave)
Cognitive learning						
2008 cohort	9.6 ± 2.5	13.2 ± 2.5	18%	37%	34%[a]	29%
($N = 24$)	48%	66%				
2009 cohort	7.8 ± 2.4	13.1 ± 2.5	27%	69%	44%[a]	43%
($N = 18$)	39%	66%				
Technical skills learning						
2008 cohort	6.9 ± 2.9	12.4 ± 3.3	34%	78%	60%[a]	63%
($N = 24$)	43%	77%				
2009 cohort	6.6 ± 3.2	12.2 ± 2.4	35%	85%	60%[a]	60%
($N = 18$)	41%	76%				

[a] Robustness of educational intervention defined if $\langle g \rangle$ greater than 30%

one-day course for novice bronchoscopists. The true value of the pre-test/post-test assessment model has been controversial because of the effects of many extraneous variables, including the Hawthorne effect (knowing that one is being tested may affect the results), the halo effect (the human tendency to respond positively or negatively to an instructor), and the practice effect (of a pre-test on a subsequent post-test). These limitations are inherent to most measures of knowledge acquisition in social research. For this reason, quasi-experimental designs are frequently used in education research to evaluate interventions where randomization cannot be performed because of ethical considerations, difficulty with randomization, or small available sample sizes [30]. We thus favored a synthetic design that also involved the integration of numerous variables in order to add to the internal validity of a simple analysis of pre-test/post-test gains [31].

The effectiveness of this curriculum, independent of the study group's pre-test level of knowledge, was established using measures of *class-average normalized gain* ($\langle g \rangle$) and related metrics. This follows a practice in physics education research where it has been shown that $\langle g \rangle$ for courses with widely varying average pre-test scores ($\langle \%pre \rangle$) is nearly independent of the pre-test score, being dependent primarily on the effectiveness of the instruction [16]. Weiman and Perkins [32] described the $\langle g \rangle$ metric as the fraction of concepts that students master, on average, which they did not already know at the start of the class.

To mitigate the need for a control group, the curriculum and the pre-tests and post-tests were all administered on the

same day. 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-test/post-test design, including history, maturation, and testing effect [25]. Furthermore, the baseline knowledge and skill level of a second cohort of novice bronchoscopists one year later was similar to the first, and participation in the course again resulted in significant learning gains in both knowledge and skill.

To diminish the skewing effect of outlier students with very high or very low pre-test scores, we calculated individual *single-student normalized gain*, where $g_i = [\%\text{post-test} - \%\text{pre-test}]/100\% - \%\text{pre-test}]$. This is the actual gain divided by the maximum gain achievable by each student and has been described by McGowan and Davis [33] as "telling us what the student achieved in tests, given what was possible for her (him) to achieve." The use of the single-student g and its related calculations has received empirical justification as an easy-to-use gauge of course effectiveness in hundreds of classroom teaching and other varying types of courses with different instructors and student populations [15]. In addition, individualized learning needs can potentially be determined by *single-student normalized gain* assessments. In our study, this performance measure allowed us to document changes in individual scores in addition to explore the overall effectiveness of our one-day curriculum. As expected, we demonstrated that the largest improvements were seen in learners with the lowest pre-test scores (Fig. 2A, B). This is a fairly obvious observation, basically suggesting that those who have more to learn do actually learn more. Yet it provides additional justification for objective measurements of knowledge and skill acquisition in each novice trainee; when an individual's plotted learning curve does not meet established expectations, remedial intervention may be in order.

Our study has several limitations. First, our curriculum was designed for single-day delivery targeting a small number of participants. This was justified as our objective was to document only learning gain, not procedural competency. It reflects the logistical reality of the difficulties involved in bringing together trainees from eight institutions in southern California. Second, the study was intentionally limited to assess short-term acquisition of knowledge and skills and was not meant to follow the learners' long-term knowledge and skill retention or decay [34–36]. Third, despite the use of various skill stations, we focused only on bronchoscopic anatomy and inspection skill acquisition. We had two rationales for this: this is indeed the basis of all bronchoscopic procedures, and a considerable time commitment would have been required to test each participant at every other station. Last, as in many areas of medical education, the impact of this intervention on clinical practice and outcomes is unknown [37, 38]. Patient-related outcomes are generally not an ideal surrogate to demonstrate effectiveness of educational interventions. Despite this, most medical education research is founded on the basic logical assumption that knowledge and skill acquisition eventually leads to improved patient care [6].

Conclusion

It has long been recognized that assessment drives learning and that rigorous assessment inspires learning, reinforces confidence, and reassures the public [39, 40]. In the context of procedure-based training, we submit that the pre-test/post-test model with calculation of various measures of learning gain, including *class-average* and *single-student normalized gains,* provides an objective and informative means to document learner performance and demonstrate the effectiveness of the educational intervention. The presence of diverse opinions regarding educational methodologies [41, 42], curricular structure [43], and measures of effectiveness persist [44, 45], necessitating further studies to confirm and build on our findings.

Acknowledgments This study was funded in part by a grant from the Chest Foundation, California chapter, to Dr. Colt, primary investigator. We thank Drs. Momen Wahidi, Andrew Duke, Cristina Plencovich, Richard Hake, and Steven Downing for their generous input, and our many colleagues at UCI and other southern California training institutions.

Disclosures Drs. Colt, Davoudi, Murgu, and Zamanian Rohani have no conflicts of interest or financial ties to disclose.

References

1. Silvestri GA (2008) The evolution of bronchoscopy training. Respiration 76(1):92–101
2. Haponik EF, Russell GB, Beamis JF, Britt EJ, Kvale P, Mathur P, Metha A (2000) Bronchoscopy training: current fellows' experiences and some concerns for the future. Chest 118(3):572–573
3. Wahidi MM, Silvestri GA, Coakley RD, Ferguson JS, Shepherd RW, Moses L, Conforti J, Que L, Anstrom KJ, McGuire F, Colt H, Downie GH (2009) A prospective multi-center study of competency metrics and educational interventions in the learning of bronchoscopy among starting pulmonary fellows. Chest 137(5):1040–1049
4. Reznick RK, MacRae H (2006) Teaching surgical skills—changes in the wind. N Engl J Med 355:2664–2669

5. Carraccio C, Wolfsthal SD, Englander R, Ferentz K, Martin C (2002) Shifting paradigms: from Flexner to competencies. Acad Med 77(5):361–367
6. Norman G (2009) The American College of Chest Physicians evidence-based educational guidelines for continuing medical education interventions: a critical review of evidence-based educational guidelines. Chest 135(3):834–837
7. Schijven MP, Schout BM, Dolmans VE, Hendrikx AJ, Broeders IA, Borel Rinkes IH (2008) Perceptions of surgical specialists in general surgery, orthopaedic surgery, urology and gynaecology on teaching endoscopic surgery in The Netherlands. Surg Endosc 22(2):472–482
8. Cronbach LJ, Furby L (1970) How should we measure "change" or should we? Psychol Bull 74:68–80
9. Hake RR (2009) Should We Measure Change? Yes!. http://www.physics.indiana.edu/~hake/MeasChangeS.pdf. Accessed 16 May 2010
10. Melzer DE (2002) Normalized learning gain: a key measure of student learning [Addendum to: Melzer DE (2002) The relationship between mathematics preparation and conceptual learning gains in physics: a possible "hidden variable" in diagnostic pretest scores. Am J Phys 70:1259–1267]. http://scitation.aip.org/getpdf/servlet/GetPDFServlet?filetype=pdf&id=AJPIAS000070000012001259000001&idtype=cvips. Accessed 16 June 2009
11. U.S. Department of Education (USDE) (2003) Identifying and implementing educational practices supported by rigorous evidence: a user friendly guide. Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. http://www.ed.gov/rschstat/research/pubs/rigorousevid/index.html. Accessed 16 May 2010
12. Shadish WR, Cook TD, Campbell DT (2002) Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin, New York
13. Hake RR (1998) Interactive-engagement vs traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses. Am J Phys 66:64–74
14. Prather EE, Rudolph AL, Brissenden G, Schlingman WM (2009) A national study assessing the teaching and learning of introductory astronomy. Part I. The effect of interactive instruction. Am J Phys 77(4):320–330
15. Epstein J (2009) The calculus concept inventory—new data. Correlation with teaching methodology. Joint Mathematics Meeting, Washington, DC, 5–10 January 2009; abstract. http://www.ams.org/amsmtgs/2110_abstracts/1046-h1-1458.pdf
16. Hake RR (2008) Design-based research in physics education research: a review. In: Kelly AE, Lesh RA, Baek JY (eds) A handbook of design research methods in education: innovations in science, technology, engineering, and mathematics learning and teaching. Routledge, New York
17. Hovland CI, Lumsdaine AA, Sheffield FD (1965) A baseline for measurement of percentage change. In: Hovland CI, Lumsdaine AA, Sheffield FD (eds), Experiments on mass communication. New York: Wiley (first published in 1949). Reprinted as pages 77–82 in Lazarsfeld PF, Rosenberg M (eds) (1955) The language of social research: a reader in the methodology of social research. Free Press, New York
18. Hake RR (2002) Lessons from the physics education reform effort. Conserv Ecol 5(2):28. http://www.consecol.org/vol5/iss2/art28/. Accessed 16 June 2009
19. Confrey J, Stohl V (2004) On evaluating curricular effectiveness: judging the quality of k-12 mathematics evaluations. National Academies Press, Washington, pp 36–37
20. McGaghie WC, Siddall VJ, Mazmanian PE, Myers J (2009) Lessons for continuing medical education from simulation research in undergraduate and graduate medical education. Chest 135:62S–68S
21. Bronchoscopy International, The Essential Bronchoscopist© (2009) http://www.essential-bronchoscopy.org/intro_en.asp. Accessed 16 June 2009
22. Harden RM, Crosby JR (2000) The good teacher is more than a lecturer: the twelve roles of the teacher. Med Teach 22:334–347
23. Quadrelli S, Galíndez F, Davoudi M, Colt HG (2009) Reliability of a 25-item low-stakes multiple choice assessment of bronchoscopic knowledge. Chest 135(2):315–321
24. Davoudi M, Osann K, Colt HG (2008) Validation of two instruments to assess technical bronchoscopic skill using virtual reality simulation. Respiration 76(1):92–101
25. Campbell D, Stanley J (1963) Experimental and quasi-experimental designs for research. Rand-McNally, Chicago
26. Prather EE, Rudolph AL, Brissenden G (2009) Teaching and learning astronomy in the 21st century. Phys Today 62(10):41–47
27. Miller GE (1990) The assessment of clinical skills/competence/performance. Acad Med 65 Suppl:S63-S67
28. Anderson LW, Krathwohl D (eds) (2001) A taxonomy for learning, teaching and assessing: a revision of bloom's taxonomy of educational objectives. Addison Wesley Longman, Upper Saddle River
29. Dreyfus HL, Dreyfus SE (1992) Mind over machine: the power of human intuition and expertise in the age of the computer. Free Press, New York
30. Thomson O'Brien MA, Freemantle N, Oxman AD, Wolf F, Davis DA, Herrin J (2001) Continuing education meetings and workshops: effects on professional practice and health care outcomes. Cochrane Database Syst Rev (2):CD003030
31. Lynch DC, Whitley TW, Willis SE (2000) A rationale for using synthetic designs in medical education research. Adv Health Sci Educ 5:93–103
32. Wieman C, Perkins K (2005) Transforming physics education. Phys Today 58(11):36–41
33. McGowan MA, Davis G (2005) Individual gain and engagement with teaching goals. In: McDougall D (ed), Proceedings of the 26th Annual Conference of the International Group for the Psychology of Mathematics Education, North American Chapter. OISE, Toronto
34. Willingham DB, Dumas JA (1997) Long-term retention of a motor skill: implicit sequence knowledge is not retained after a one-year delay. Psychol Res 60:113–119
35. Marinopoulos SS, Baumann MH (2009) Methods and definitions of terms: effectiveness of continuing medical education: American College of Chest Physicians evidence-based educational guidelines. Chest 135:17S–28S
36. Bordage G, Carlin B, Mazmanian PE (2009) Continuing medical education effect on physician knowledge: effectiveness of continuing medical education: American College of Chest Physicians evidence-based educational guidelines. Chest 135:29S–36S
37. Sweet RM, McDougall EM (2008) Simulation and computer-animated devices: the new minimally invasive skills training paradigm. Urol Clin North Am 35(3):519–531
38. Wang EE, Quinones J, Fitch MT, Dooley-Hash S, Griswold-Theodorson S, Medzon R, Korley F, Laack T, Robinett A, Clay L (2008) Developing technical expertise in emergency medicine—the role of simulation in procedural skill acquisition. Acad Emerg Med 15(11):1046–1057
39. Cooke M, Irby DM, Sullivan W, Ludmerer KM (2006) American medical education 100 years after the Flexner report. N Engl J Med 355:1339–1344
40. Whitcomb ME, Nutter D (2006) Learning medicine in the 21st century: the education of medical students & the general professional education of the physician. http://www.carnegiefoundation.org/sites/default/files/Learning_Medicine.pdf. Accessed 16 May 2010
41. Chin MW, Forbes GM (2008) Should simulator use become mandatory in endoscopy training? J Gastroenterol Hepatol 23:996–997

42. Kruidering-Hall M, O'Sullivan PS, Chou CL (2009) Teaching feedback to first-year medical students: long-term skill retention and accuracy of student self-assessment. J Gen Intern Med 24(6):721–726
43. Christensen C, Horn MB, Johnson CW, Horn MB (2008) Disrupting class: how disruptive innovation will change the way the world learns. McGraw Hill, New York, pp 110–111
44. Michelson JD, Manning L (2008) Competency assessment in simulation-based procedural education. Am J Surg 196:609–615
45. General Accounting Office (2009) Program evaluation: a variety of rigorous methods can help identify effective interventions (GAO-10-30), 23 November. http://www.gao.gov/Products/GAO-10-30. Accessed 16 May 2010