The Retrospective Pre-Post: A Practical Method to Evaluate Learning from an Educational Program

Article *in* Academic Emergency Medicine · February 2012

DOI: 10.1111/j.1553-2712.2011.01270.x · Source: PubMed

5 authors, including:

Farhan Bhanji
McGill University Health Centre
107 PUBLICATIONS **7,169** CITATIONS
SEE PROFILE

Ronald Gottesman
McGill University
60 PUBLICATIONS **3,004** CITATIONS
SEE PROFILE

Willem De Grave
Maastricht University
120 PUBLICATIONS **4,115** CITATIONS
SEE PROFILE

Laura Winer
McGill University
39 PUBLICATIONS **491** CITATIONS
SEE PROFILE

Academic Emergency Medicine
Official Journal of the Society for Academic Emergency Medicine

EDUCATIONAL ADVANCE

The Retrospective Pre–Post: A Practical Method to Evaluate Learning from an Educational Program

Farhan Bhanji, MD, MSc(Ed), Ronald Gottesman, MD, Willem de Grave, PhD, Yvonne Steinert, PhD, and Laura R. Winer, PhD

Abstract

Objectives: Program evaluation remains a critical but underutilized step in medical education. This study compared traditional and retrospective pre–post self-assessment methods to objective learning measures to assess which correlated better to actual learning.

Methods: Forty-seven medical students participated in a 4-hour pediatric resuscitation course. They completed pre and post self-assessments on pediatric resuscitation and two distracter topics. Postcourse, students also retrospectively rated their understanding as it was precourse (the ''retrospective pre'' instrument). Changes in traditional and retrospective pre- to postcourse self-assessment measures were compared to an objectives-based multiple-choice exam.

Results: The traditional pre to post self-assessment means showed an increase from 1.9 of 5 to 3.7 of 5 ($p < 0.001$); the retrospective pre to post scores also increased from 1.9 of 5 to 3.7 of 5 ($p < 0.001$). Although the group means were the same, individual participants demonstrated a response shift by either increasing or decreasing their traditional pre to retrospective pre scores. Scores on the 22-item objective multiple choice test also increased, from a median score of 13.0 to 18.0 ($p < 0.001$). There was no correlation between the change in self-assessments and objective measures as demonstrated by a Spearman correlation of −0.02 and −0.13 for the traditional and retrospective pre–post methods, respectively. Students reported fewer changes on the two distracters using the retrospective pre–post versus the traditional method (11 vs. 29).

Conclusions: Students were able to accurately identify, but not quantify, learning using either traditional or retrospective pre–post ''self-assessment'' measures. Retrospective pre–post self-assessment was more accurate in excluding perceived change in understanding of subject matter that was not taught.

ACADEMIC EMERGENCY MEDICINE 2012; 19:189–194 © 2012 by the Society for Academic Emergency Medicine

From the Montreal Children's Hospital (FB, RG), Centre for Medical Education (FB, RG, YS, LRW), Teaching and Learning Services (LRW), McGill University, Montreal, Canada; and the University of Maastricht (WdG), Maastricht, The Netherlands.

Received April 9, 2011; revision received August 1, 2011; accepted August 8, 2011.

Presented at The Association of Medical Education in Europe, Prague, Czech Republic, September 2008; The Canadian Association of Emergency Physicians annual conference, Montreal, Quebec, June 2010. The research findings have also been presented to a small audience at the master's thesis defense of the principal author at the University of Maastricht.

The authors have no relevant financial information or potential conflicts of interest to disclose.

Supervising Editor: Jacob Ufberg, MD.

Address for correspondence and reprints: Farhan Bhanji, MD, MSc; e-mails: farhan.bhanji@muhc.mcgill.ca, fbhanji@hotmail.com.

Program evaluation remains a critical, but underutilized, step in the educational process. It supports educators in improving instruction,[1] thus contributing to the ultimate goal of improved student learning and, through it, improved patient care. Kirkpatrick's classic work outlined four levels of training evaluation: learner reaction, learning, transfer, and results.[2] In emergency medicine training, as in clinical medical education in general, evaluation is frequently overlooked or limited to self-report of learner satisfaction due to time and resource constraints. Evaluating the effect of a specific instructional intervention on subsequent patient outcomes is often not feasible. A more realistic program evaluation therefore would focus on levels 2 (learning) and 3 (transfer) of the Kirkpatrick hierarchy.

Morrison[3] describes an ideal program evaluation as reliable, valid, acceptable, and inexpensive. While an easy-to-use self-assessment of learning could, in theory, satisfy all of Morrison's criteria, the literature in

doi: 10.1111/j.1553-2712.2011.01270.x
ISSN 1069-6563
PII ISSN 1069-6563583

medical education has indicated that students' and physicians' self-assessments do not correlate well with objective measures of performance.[4–6] Whether learners are any more capable of accurate self-assessment of learning than they are of performance has not been answered definitively.[7,8]

Self-assessment changes rely on a ''common metric,'' i.e., the participant's standard of measurement for the dimension being assessed is stable from one data point to the next.[9] When learners' understanding of the dimension(s) being measured changes, they recalibrate their criteria for self-rating (the response shift bias).[10] This poses a risk to the validity of the traditional pre–post design (TPP), where data are collected *before* and *after* the intervention, with the change in learner self-ratings attributed to the educational intervention.[11] The retrospective pre–post method (RPP) offers an alternate method. Data are collected at the same point in time (i.e., at the conclusion of training); thus the ratings of understanding before (''retrospective pre'') and after (post) the intervention use the same metric. The RPP method of evaluation has been studied empirically and shown to be effective in the context of faculty development;[10,12] however, few studies validate the RPP with medical students or residents.[7] Occasional publications have claimed learning via the RPP method, but these studies have not validated the learning in an objective manner.[13–15] Our primary objective was to compare TPP and RPP with objective measures of learning to determine their potential for use as program evaluation tools.

METHODS

Study Design

This study evaluates two separate strategies of self-assessment of learning (TPP and RPP) surrounding an educational intervention and compares each to accepted objective measures of learning. The Research Ethics Board of the Faculty of Medicine, McGill University, granted ethical approval for the study. Written informed consent was obtained from all participants, including the validation group.

Study Setting and Population

The study was conducted at the McGill University. The population was a convenience sample of third-year medical students (clinical clerks) undergoing a core clinical rotation in pediatrics. Students were previously certified in Advanced Cardiac Life Support (ACLS; adult version). The pediatric resuscitation course built on this basic knowledge. The only exclusion criterion was prior successful completion of a formal Pediatric Advanced Life Support (PALS) course. Subject recruitment was from February 2007 to May 2007.

Study Protocol

The authors modified the standardized PALS course, which was developed using widely accepted international consensus-based standards of care (International Liaison Committee on Resuscitation, 2005).[16] All subjects participated in the 4-hour modified pediatric resuscitation course. Each subject completed both self- and objective assessments immediately before (pre) and after (post) the course. To minimize the effect of objective testing on the self-assessments, the self-assessments were filled out before the objective tests at each measurement time (see Figure 1).

Educational Intervention. The PALS course has a well-identified content domain with clearly defined learning objectives. These were modified to an appropriate student level by the two content-expert authors (RG and FB) and two physicians DM and SR; see acknowledgments with extensive experience in resuscitation training and medical education. The two content-expert authors then developed the introductory session and the nine core-content interactive ''teaching cases'' of the 4-hour pediatric resuscitation course.

The course was designed for groups of six to eight students per session. It was piloted with a group of eight students, with particular attention paid to the feasibility of achieving the stated objectives in the allotted time. Consistency of instruction, between subjects and across all content areas, was achieved by having a single physician deliver the entire course to all subjects.

Self-assessment Using Pre–Post and Retrospective Pre Methods. Subjects rated their current level of understanding of pediatric resuscitation, before and after the intervention, on a five-point Likert scale from low (1) to high (5). At the end of the course, they were also asked to think back to the beginning of the course and rate their understanding at that time (the retrospective pre). The difference between the retrospective pre and traditional pre scores was the ''response shift.''

Subjects also rated their understanding of two related distracter topics on the same five-point traditional pre, retrospective pre, and post formats. The distracters (pediatric fractures and toxicology) were not covered in the pediatric resuscitation course; therefore, any perceived change would reflect something other than a real change in understanding.

Objective Pre and Post Tests. As with much of medical education, there is no internationally accepted criterion standard examination for assessing resuscitation knowledge. To permit objective assessment of learning without the potential bias of a locally created test, we used the test for certification in PALS courses given by the Heart and Stroke Foundation of Canada and the American Heart Association. This multiple-choice test is the most widely used pediatric resuscitation examination, at least for pediatric residents, in North America.[17] Two equivalent, nonidentical tests were created as pre and post tests. All questions from the three versions of

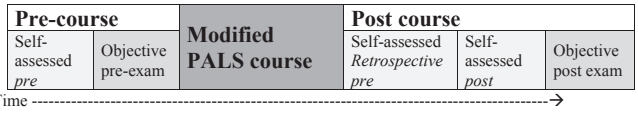

Figure 1. Study outline. PALS = Pediatric Advanced Life Support.

the PALS course written test were pooled and assessed for relevance to the objectives of the modified course. Redundant, outdated, or irrelevant questions were eliminated, resulting in a bank of 44 questions.

Questions were then classified according to the specific content area and the cognitive level (according to Bloom's taxonomy).[18] For example, a question requiring the student to assess the presenting problem and determine the best course of action for a child who presented with an abnormal heart rhythm would be classified as an ''arrhythmia application'' question. The majority of the questions were at the application level.

The objective tests were scored with one point for correct answers, zero for unanswered questions, and −0.5 points for each incorrect answer; with 22 questions, potential scores ranged from −11 to +22. Students were informed of the marking scheme prior to taking the test. The objective knowledge gain was calculated as the change in score from the pre to post multiple-choice tests.

Validating the Equivalence of Pre and Post Tests. Once all questions were coded, the questions were paired on their classification and perceived level of difficulty. The paired questions were randomly assigned to exams A and B, creating two theoretically equivalent exams of 22 questions. The exams were administered as one test (counterbalanced with half AB and half BA) to a group of nonstudy subjects for evaluation. There were no significant differences in median test results (exam A = 17.5, exam B = 17.5; p = 0.62) or in test order (first exam = 17.25, second exam = 17.75; p = 0.42) for a validation group of 18 learners.

Data Analysis

Objective test scores were analyzed using the Wilcoxon rank sum test. Self-assessment measures of resuscitation understanding had a limited range of responses (three outcome responses for any variable) and were also analyzed using the Wilcoxon rank sum test. Spearman's rank correlation was used to calculate correlations between self-assessed and objective measures.

For the distracter variables (toxicology and fractures), pre and post measures (both traditional and retrospective) were expected to be identical for each subject. The absolute difference between the two respective pre scores and the post score was calculated at the individual subject level and then analyzed using a one-sample t-test against the expected value of zero (i.e., pre and post scores should be identical). Data were analyzed using SPSS software (version 11, IBM SPSS, Armonk, NY).

RESULTS

All 49 eligible subjects were enrolled over the 4-month study period. For logistic reasons, two enrolled subjects did not complete the study protocol and their data could not be analyzed. Complete data were available from the remaining 47 subjects.

Demographic Data

The mean age of participants was 24.6 years (range = 22 to 29 years). There were 30 females and 17 male participants. None had taken the PALS course; all but one had ACLS certification.

Students' scores on the self-assessments of resuscitation increased from pre to post course (see Table 1) and from retrospective pre to post measures. Although there was no difference between the mean traditional pre and retrospective pre scores, 10 subjects did change their precourse self-assessments, with five increases and five decreases. The retrospective pre–post measure resulted in an increase for all 47 subjects; the traditional pre–post resulted in 46 of the 47 subjects indicating an increase. The single subject who did not increase on the self-assessment scores actually did increase on the objective measures. Students' scores increased significantly on the objective resuscitation multiple choice tests from a pretest median of 13.0 (interquartile range [IQR] = 10.5 to 14.5) to a posttest median of 18.0 (IQR = 16.5 to 20.5; $p < 0.001$).

Correlation between the changes in self-assessment scores, via the traditional pre–post method, to the changes in objective scores revealed a Spearman correlation of −0.02, indicating no association between the variables. As expected (given that the pre and retrospective pre scores were not significantly different), the correlation between change on self-assessment scores, via the retrospective pre to post, and objective measures was very similar at −0.13. This small difference is not practically relevant.

Scores on the distracters were expected to be identical for the pre, retrospective pre, and post measurements given that these topics were not taught and students had no other opportunity to learn the subject matter. With respect to fractures, 13 subjects demonstrated a change (positive or negative) between the TPP measures, while only three subjects did so using the RPP design. Similarly for toxicology, the traditional pre to post method had 16 subjects who changed their scores, while only eight changed via the retrospective pre to post method. One-sample t-tests revealed that the absolute difference between traditional pre and post scores differed from the expected value of 0, for both the toxicology and the fracture domains (see Table 2).

Table 1
Comparing Self-assessments for the Subject of Interest on a Five-point Likert Scale

Comparison	Self-assessed Pre	Retrospective Pre	Post	p-value
Pre to post	1.9		3.7	<0.001
Retrospective pre to post		1.9	3.7	<0.001
Pre to retrospective pre	1.9	1.9		>0.999

Table 2
Ability to Discriminate Distracters: Mean Difference From Post Score

	Mean Absolute Difference	p-value
Traditional pre		
Fracture	0.5	<0.001
Toxicology	0.3	<0.001
Retrospective pre		
Fracture	0.1	0.10
Toxicology	0.2	0.01

A similar analysis for the retrospective pre to post score revealed no difference on the domain of fractures, but a small difference for toxicology.

DISCUSSION

Using the Retrospective Pre–Post to Evaluate Learning From an Educational Program

Program evaluation remains a challenge for medical educators. The selection of the (Kirkpatrick) level of evaluation[2] should therefore be the result of an analysis of the benefits of knowing the effect of a given educational intervention versus the costs (time, money, people, etc.) of conducting such an evaluation. A review by Belfield et al.[19] showed that even among published articles on education for clinical practitioners, a minority of studies focused on patient outcome (1.6%) or learner performance (18.9%). Learners' successful completion of the program or their satisfaction was the focus of 21%, while the remaining studies focused on learning. Experience suggests that evaluation of unpublished courses focuses predominantly at the level of learner satisfaction.

In our study, the self-assessment data do not correlate well with objective measures (the Spearman correlation was close to zero) and therefore would not be useful for assessing individual student achievement; i.e., the ideal evaluation of learning from an educational program should involve objective testing. However, self-assessment still has utility for program evaluation in that it can identify when learning occurred. Our study suggests that students can accurately define that they learned but not how much they learned on an individual topic. This accurate identification of learning is helpful in the evaluation of educational programs, and the feedback it provides is useful to future program decision-making. Student learning is a more relevant criterion than learner appreciation for modifying programs.

Although objective measures of learning have important advantages, they are not appropriate in all cases, and choosing a suitable method of self-assessment is important. The RPP has several important advantages over the TPP. The RPP can be valuable in situations where a formal ''preevaluation,'' either self-assessment or objective, would sensitize the learner and have a negative influence on experiential learning.[20] For example, in EM, a program developer for a simulation-based workshop may not want to ask learners about their understanding of ventricular fibrillation prior to the scenario and debrief on the same topic.

Second, the RPP avoids potential response shift bias inherent in TPP. To illustrate, consider the case of a hypothetical course titled ''Child abuse for primary care physicians.'' Family physicians may rate themselves as fairly knowledgeable and subjectively score themselves as 7 of 10 prior to the course. However, as a result of being sensitized during the course to the true scope of the issue, the physicians may change their internal standard for evaluation. They may then rate themselves as 7 of 10 on completion of the course. A traditional pre–post self-assessment evaluation would conclude that the physicians did not learn from the program. Using a retrospective pre–post questionnaire (i.e., asking participants after completing the course to assess their understanding both before and after) would allow physicians to score their understanding at both times with the same internal standard. In our example, a given physician may have retrospectively rated her pretest understanding as 4 of 10, even though before the course she might have said 7. This would more accurately reflect her self-assessment of learning from the course. Finally, the RPP is easier to administer, as it only requires data collection at one point in time.

In the current study, the retrospective pre–post was nearly identical to traditional pre–post on the areas of instruction, but was better able to exclude distracters. A response shift between TPP and RPP was not evident and may be related to students' prior ACLS training and potential understanding of the parameters being assessed (pediatric resuscitation) at the outset of the course. Alternatively, the similarity of the TPP and RPP responses may be a reflection of the study design, where students functioned as their own controls and completed a score for both the TPP and the RPP. A research design with subjects randomized to either the traditional pre or the retrospective pre groups might yield different results. The ability of the traditional pre–post to better discriminate distracters may reflect the need for a ''common metric'' when the self-assessments are made.

LIMITATIONS

There are some methodologic and theoretical limitations to this study. Most importantly, the study only assessed learning through a paper-based test. The majority of the questions focused on the application of learning, but performance on a test does not necessarily transfer to performance in the clinical setting. While there is limited research that suggests that written tests may not predict resuscitation performance in a simulated environment for adult resuscitation courses,[21,22] the general medical education literature suggests that written tests can predict performance in an objective structured clinical exam[23,24] and that written tests may predict actual performance in the clinical environment as well as a multiple station examination.[25] Nonetheless, studies evaluating the TPP and RPP methods against performance based measures should be pursued in the future.

The study focused on the immediate learning of resuscitation. No attempt was made to evaluate long-term retention, a more important measure of learning.

However, given limited evidence that self-assessment of learning would correlate to objective measures, the current study was a necessary initial step.

There is a final caution related to the context in which the present study was conducted. The subjects' participation was voluntary and did not contribute to the evaluation of their clerkship rotation. Social desirability may play a much larger role, with students more willing to claim learning, if their performance is being formally evaluated as part of their training program.

CONCLUSIONS

Morrison suggested the need for program evaluation tools that are reliable, valid, acceptable, and inexpensive.[3] While the ideal evaluation of learning from a program should be an objective measure, this is not always possible or feasible. The subjective, easy-to-administer RPP fulfills many of these criteria and, from our study, appears to be at least as effective as the TPP design. It demonstrates validity in its ability to identify learning and rule out distracters; however, reliability was not explicitly addressed. The RPP requires few resources, either financial or instructor development time, and is easy to use as it is completed at one point in time. This ease of development and usability may help with acceptability among clinical educators in emergency medicine. Finally, it allows subjective assessment of learning without sensitizing the learner to the subject matter with a prequestionnaire (subjective or objective), often a consideration when instructional strategies such as simulations are used. As this is the first study looking specifically at the emergency medicine context, further research would help contribute to greater acceptance.

Dr. Farhan Bhanji is the Richard and Sylvia Cruess Faculty Scholar in Medical Education at McGill University. The authors are grateful to the Heart and Stroke Foundation of Canada for their permission to conduct the study and their willingness to share course information. The authors thank Drs. Lambert Schuwirth, Kelly Skeff, and John Norcini for their expert advice on the design of the research project; Sebastien Dubé for his help with the statistical analysis; Drs. David McGillivray and Saleem Razack for their expertise in designing the resuscitation course; and Dr. Peter Cantillon for his editorial suggestions. None of these individuals received any financial compensation for their valuable assistance. Finally, we thank the students who participated in the study, as without their support the study would not have been possible.

References

1. Wilkes M, Bligh J. Evaluating educational interventions. Br Med J. 1999; 318:1269–72.
2. Kirkpatrick DL. Evaluating Training Programs: The Four Levels. San Francisco, CA: Berrett-Koehler, 1994.
3. Morrison J. ABC of learning and teaching in medicine: evaluation. Br Med J. 2003; 326:385–7.
4. Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence - a systematic review. JAMA. 2006; 296:1094–102.
5. Vnuk A, Owen H, Plummer J. Assessing proficiency in adult basic life support: student and expert assessment and the impact of video recording. Med Teach. 2006; 28:429–34.
6. Ward M, Gruppen L, Regehr G. Measuring self-assessment: current state of the art. Adv Health Sci Educ. 2002; 7:63–80.
7. D'Eon M, Sadownik L, Harrison A, Nation J. Using self-assessments to detect workshop success. Do they work? Am J Eval. 2008; 29:92–8.
8. Lam TCM. Do self-assessments work to detect workshop success? An analysis of argument and recommendation by D'Eon et al. Am J Eval. 2009; 30:93–105.
9. Cronbach LJ. How we should measure change - or should we. Psychol Bull. 1970; 74:68–80.
10. Levinson W, Gordon G, Skeff, K. Retrospective versus actual pre-course self-assessments. Eval Health Profess. 1990; 13:445–52.
11. Howard GS. Response-shift bias - a problem in evaluating interventions with pre-post self-reports. Eval Rev. 1980; 4:93–106.
12. Skeff KM, Stratos GA, Bergen MR. Evaluation of a medical-faculty development program - a comparison of traditional pre/post and retrospective pre/post self-assessment ratings. Eval Health Prof. 1992; 15:350–66.
13. Pirrallo RG, Wolff M, Simpson DE, et al. Analysis of an international emergency medical-service train-the-trainer program. Ann Emerg Med. 1995; 25:656–9.
14. Yeazel MW, Center BA. Demonstration of the effectiveness and acceptability of self-study module use in residency education. Med Teach. 2004; 26:57–62.
15. Razack S, Meterissian S, Morin L, et al. Coming of age as communicators: differences in the implementation of common communications skills training in four residency programmes. Med Educ. 2007; 41:441–9.
16. International Liaison Committee on Resuscitation. 2005 International consensus on cardiopulmonary resuscitation and emergency cardiovascular care science with treatment recommendations. Part 6: pediatric basic and advanced life support. Circulation. 2005; 112:III73–90
17. Halamek LP, Kaegi DM. Who's teaching neonatal resuscitation to housestaff? Results of a national survey. Pediatrics. 2001; 107:249–55.
18. Bloom BS, Engelhart MD, Furst EJ, Hill WH, Krathwohl DR. Taxonomy of Educational Objectives: The Classification of Educational Goals: Handbook I. Cognitive Domain. New York, NY: McKay, 1956.
19. Belfield C, Thomas H, Bullock A, Eynon R, Wall D. Measuring effectiveness for best evidence medical education: a discussion. Med Teach. 2001; 23:164–70.
20. Kolb D, Fry R. Towards an applied theory of experiential learning. In: Cooper C (ed.). Theories of Group Processes. London, UK: John Wiley, 1975.
21. Napier F, Davies RP, Baldock C, et al. Validation for a scoring system of the ALS cardiac arrest simulation test (CASTest). Resuscitation. 2009; 80:1034–8.
22. Rodgers DL, Bhanji F, McKee BR. Written evaluation is not a predictor for skills performance in an

Advanced Cardiovascular Life Support course. Resuscitation. 2010; 81:453–6.
23. Kramer AW, Jansen JJ, Zuithoff P, et al. Predictive validity of a written knowledge test of skills for an OSCE in postgraduate training for general practice. Med Educ. 2002; 36:812–9.
24. Remmen R, Scherpbier A, Denekens J, et al. Correlation of a written test of skills and a performance based test: a study in two traditional medical schools. Med Teach. 2001; 23:29–32.
25. Ram P, van der Vleuten C, Rethans JJ, Schouten B, Hobma S, Grol R. Assessment in general practice: the predictive value of written-knowledge tests and a multiple-station examination for actual medical performance in daily practice. Med Educ. 1999; 33:197–203.