Changing estimates of leadership ability before a programme: Retrospective self-assessments and response-shift bias

Authors:
Lyons, Oscar (Corresponding Author)
University of Oxford Medical Sciences Division, Nuffield Department of Surgical Sciences
Oxford, Oxfordshire, UK
University of Oxford Medical Sciences Division, Nuffield Department of Primary Care Health Sciences
Oxford, Oxfordshire, UK

Kingsley-Smith, Harry
Oxford University Hospitals NHS Foundation Trust, Department of Medical Education
Oxford, Oxfordshire, UK

Kotze, Koot
University of Oxford Medical Sciences Division, Nuffield Department of Primary Care Health Sciences
Oxford, Oxfordshire, UK

Nandra, Karandeep
Oxford University Hospitals NHS Foundation Trust, Department of Medical Education
Oxford, Oxfordshire, UK
Registrar in Public Health
East of England deanery
karandeep.nandra1@nhs.net

Galante, Joao R
Oxford University Hospitals NHS Foundation Trust, Department of Medical Education
Oxford, Oxfordshire, UK
Maidstone and Tunbridge Wells NHS Trust, Kent Oncology Centre, Maidstone, Kent, UK.
joao.galante@nhs.net

Fahy, Nicholas
University of Oxford Medical Sciences Division, Nuffield Department of Primary Care Health Sciences
Oxford, Oxfordshire, UK

Canter, Richard
University of Oxford Medical Sciences Division, Nuffield Department of Surgical Sciences
Oxford, Oxfordshire, UK

Corresponding Author Details

Oscar Lyons
oscar.lyons@nds.ox.ac.uk
Nuffield Department of Surgical Sciences
University of Oxford
John Radcliffe Hospital, Headington
Oxford
OX3 9DU
United Kingdom

Word Count: 1580
Figures: 1
Tables: 1
Supplementary Materials: 4

ABSTRACT

Background:

Most evaluations of clinical leadership development programmes rely on self-assessments. Self-assessments are vulnerable to response-shift bias. Using retrospective then-tests may help to avoid this bias.

In this study, we investigate whether, paired with post-test self-assessments, post-programme then-tests (retrospective self-assessments) are more sensitive to change in clinical leadership development programme participants than traditional pre-programme pre-tests.

Methods:

17 healthcare professionals participated in an eight-month single-centre multidisciplinary leadership development programme. Participants completed prospective pre-test, retrospective then-test and traditional post-test self-assessments using the Primary Colours Questionnaire (PCQ) and Medical Leadership Competency Framework Self-Assessment Tool (MLCFQ). Pre-post pairs and then-post pairs were analysed for changes using Wilcoxon Signed-Rank Tests and compared with a parallel multi-method evaluation organised by Kirkpatrick levels.

Results:

A greater number of significant changes were detected using then-test pairs than pre-test pairs for both the PCQ (11/12 vs 4/12 items) and MLCFQ (7/7 vs 3/7 domains). The multi-methods data showed positive outcomes at all Kirkpatrick levels.

Conclusions:

In ideal circumstances, both pre-test and then-test evaluations should be conducted. We cautiously suggest that if only one post-programme evaluation can be conducted, then-tests may be appropriate means of detecting change.

BACKGROUND

The NHS is estimated to spend more than a quarter of their organisational development budget on leadership development. This amounts to more than £1 billion sterling per year.[1] Such expenditure relies on an assumption that leadership development programmes lead to significant clinical impact. Lack of effective evaluation of these programmes leaves this a mostly untested assumption.[2–4]

Over two thirds of the clinical leadership development programmes that are evaluated rely solely on self-assessments for programme evaluation.[2] Self-assessments, however, often do not have a stable reference point. During educational interventions where a participant's experience increases, participants could re-calibrate their confidence to lower levels as they become aware of their relative lack of knowledge. Howard and colleagues refer to this shift as "Response-Shift Bias".[5]

To account for Response-Shift Bias, Howard introduced the retrospective "then-test", where participants are asked after an intervention to rate their previous selves before the intervention ("what were you like back *then*"). These then-tests are paired with a post-intervention rating of themselves after the intervention (at the same time as the then-test) to determine the absence or presence of an improvement.[6,7]

The application of then-tests to clinical leadership development programmes using self-assessment as a primary means of evaluation could provide the dual benefit of time-efficiency and reduced risk of response-shift bias. There have however been limited studies of then-tests in clinical leadership development programmes, and those studies which have been completed have been criticised for their lack of triangulation of results.[7]

In this pilot project, we evaluate whether retrospective then-tests are more able to detect meaningful change in medical leadership development programme participants than traditional prospective pre-tests.

METHODS

Setting

We focused on the second cohort of the Emerging Leaders Programme (ELP), an eight-month long multidisciplinary leadership development programme based in a single UK hospital. Programme learning was through a combination of monthly full-day workshops and team leadership projects. Workshops included small and large group teaching sessions, skills sessions and facilitated group work. Content was based on the Medical Leadership Competency Framework (MLCF)[8], the Faculty of Medical Leadership and Management (FMLM) leadership standards[9] and Pendleton and Furnham's Primary Colours model of leadership[10]. Participants worked in teams on a leadership project centred on complex quality improvement projects in their hospital, as a vehicle for application of their learning.[11]

Seventeen health professionals volunteered for the programme. Participants included 11 medical trainees, 2 nurses, 2 midwives and 2 physiotherapists.

Evaluation

To investigate the relative effectiveness of then-tests in detecting change, we added a then-test to each item on the two existing programme evaluation questionnaires, which were already used as pre- and post-programme measures.

The two questionnaires used were the Primary Colours Questionnaire (PCQ) (developed as part of OL's doctoral research) and an adapted version of the Medical Leadership Competency Framework Self-Assessment Tool (MLCFQ). In the PCQ, participants rated their leadership ability on twelve items using a 10-point Likert scale anchored to 1=very poor and 10=excellent. In the MLCFQ, participants rated themselves on 56 behaviours grouped into seven domains, using an adapted seven-point Likert scale anchored to 1=strongly disagree and 7=strongly agree (the original questionnaire uses a three-point scale).[12] The design of these questionnaires is described in Supplementary Material 2, and the full questionnaires are available as Supplementary Material 3 and Supplementary Material 4.

In addition to the PCQ and MLCFQ questionnaires, we collected data using multiple methods in the form of free text questions, systematic observations throughout the programme, post-workshop feedback (both quantitative and qualitative), and outcomes from programme projects (henceforth, "multi-methods data").

Data Analysis

Median changes in participant scores on the PCQ and MLCFQ for pre-post/post-test pairs and then-post/post-test pairs were compared using Wilcoxon Signed-Rank Tests for each of the twelve PCQ items and each of the seven MLCFQ domains. Alpha was set at the 0.05 level and the Bonferroni correction was applied in each set of tests. The combined score for each of the MLCFQ domains was scaled to 1-7 for ease of interpretation.

Multi-methods data were analysed using Kirkpatrick's levels[13] as a sensitising framework. The main analysis was completed by OL, a medical doctor with a DPhil (PhD) in evaluating medical leadership development programmes, with input from all other authors.

Ethical Approval

The University of Oxford Clinical Trials and Research Governance was consulted, and ethics exemption was granted as an evaluation of service delivery. Participant responses were anonymised and paired using pre-assigned participant codes.

RESULTS

All 17 participants completed the full evaluation. In the PCQ, participants improved significantly on 11/12 items in the then-post test pairs, and only 4/12 in the pre-post pairs. In the MLCFQ, participants improved significantly on all seven domain scores in then-post pairs, and only three domains in the pre-post pairs. Outcomes are shown in Figure 1, with more detail in Supplementary Material 1. Impact was shown across all four levels of Kirkpatrick's framework, as shown in Table 1.

[Figure 1 about here]

Figure 1 caption: Median score on retrospective then-test, prospective pre-test and traditional post-test self assessment. Symbols indicate statistical significance after correction for multiple comparisons. MLCF scores were from 1-7, PCQ scores from 1-10. MLCF, Medical Leadership Competency Framework; PCQ, Primary Colours Questionnaire.

Kirkpatrick Level	**Programme Outcomes**
Level One: Reaction *Participants' views on the learning experience, its organisation, presentation,*	Participants rated each session from 1=poor to 5=excellent. Mean ratings were 4.04 (Fun), 4.21 (Learning), 4.56 (Engagement) and 4.39 (Organisation).

content, teaching methods and quality of instruction.	
Level Two: Attitudes/Knowledge *Changes in the attitudes or perceptions among participant groups towards teaching and learning; acquisition of concepts; acquisition of thinking/problem-solving, psychomotor and social skills.*	Participants reported increased self-awareness alongside improved knowledge in leadership, management, engagement with stakeholders and health system structure: "It was invaluable in providing insight into the structural processes involved in effecting change at individual, Trust and even National level."
Level Three: Behaviour *Transfer of learning to the workplace or willingness of learners to apply new knowledge and skills.*	Participants reported spending an additional 60-120 minutes per week (median) outside working on their leadership projects. Most participants (15/17) reported starting other clinical improvement projects outside the programme, during the programme. Two graduates delivered workshops for the subsequent cohort. Three further graduates contributed to the subsequent as leadership coaches.
Level Four: Results *Changes in the organisation, attributable to the educational program; improvement in student or resident learning/performance as a direct result of the educational intervention.*	One participant successfully advocated for funding of the next ELP. One clinical improvement project introduced peer support groups for staff which were well-attended and have been funded for two more years.

Table 1: Multi-methods data, grouped by Kirkpatrick Level. ELP, Emerging Leaders Programme.

DISCUSSION

Then-test pairs showed a greater number of significant changes than pre-test pairs for both the PCQ and the MLCFQ. The multi-methods data presented in Table 1 showed positive outcomes from the programme at all four Kirkpatrick levels. The closer alignment of then-test results with the multi-method evaluation suggests that then-tests may have been more accurate than pre-tests as well as more sensitive. Participants may have been better informed after the programme and therefore better equipped to assess their pre-programme abilities retrospectively (then-test) than prospectively (pre-test).

Then-test responses tended to be the same or lower than pre-test responses. In real terms, leadership programme evaluations usually look at changes over a programme. Our study was therefore designed to analyse differences in changes in pre-test/post-test pairs compared with changes in then-test/post-test pairs. The study was not powered to compare pre-test responses and then-test responses, so it is not possible to evaluate the significance of differences in the two alternative baseline measurements. This may be an interesting area for further investigation.

The increase in number of significant changes is similar to that found in other studies.[7] While many of these studies used a single method to evaluate programme outcomes, we used two questionnaires paired with a multi-methods evaluation. A lack of objective measures specifically aligned to each questionnaire item limits our ability to infer whether the increased number of significant changes is an accurate representation of participant learning or a result of otherwise inflated changes. Despite this limitation, we note that our multi-methods data aligned with the findings of the retrospective approach. The descriptors of the four MLCF domains that only showed a statistically significant change using a retrospective approach align with the positive findings in Table 1 above. There were expressions of improved self-awareness (Domain 1, "Developing self awareness by being aware of their own values, principles, and assumptions, and by being able to learn from experiences"); increased involvement in quality improvement (Domain 4, "Facilitating transformation by actively contributing to change processes that lead to improving healthcare"); involvement in supporting and motivating others to develop their leadership (Domain 6, "Communicating the vision and motivating others to work towards achieving it"), and deliberate engagement with stakeholders at a range of levels (Domain 7, "Developing the strategy by engaging with colleagues and key stakeholders).

Studies have supported the use of then-tests and identified them as more closely aligned than pre-tests with changes in objective performance in a range of settings.[14–17] Moreover, recent psychological modelling shows objectively low-performing individuals are prone to overconfidence when self-scoring their performance.[18]

Retrospective tests are however considered vulnerable to biases associated with overrepresenting change. Participants exposed to an intervention aimed at increasing knowledge are prone to believing their knowledge has increased post-intervention under an implicit theory of change.[19] There is also evidence that inclusion of then-test questions in evaluations may alter the post-test responses.[20] When individuals invest time and effort into participating there is a risk of 'error-justification' bias, under the assumption that such time and effort must not have gone to waste.[21] Similarly, there is a predisposition towards reporting positive change because it is a desirable outcome for all parties involved. This is termed 'social desirability' bias or 'impression management'.[22]

Recent study has suggested then-tests generate more noninvariant data[23], which could indicate that participants are merely "satisficing" rather than actually considering each response[24]. If participants tended towards satisficing in then-tests, this could potentially account for the greater significance detected compared using Wilcoxon signed-rank tests, which rely on consistency of change rather than magnitude. There is evidence to suggest separation of then-test and post-tests onto two forms, each without referencing the other, is less prone to satisficing as well as error-justification, impression management and implicit theories of change than the parallel form that we used in this study.[19,25,26]

Previous studies have conflicted over the extent these biases manifest themselves in programme evaluations.[7] Many factors influence how vulnerable a question is to response-shift bias. Previous studies suggest that behaviour questions are less vulnerable than ability questions.[21] This could explain the smaller magnitude of change in MLCFQ scores compared to PCQ scores.

From this study we have concluded that, in ideal circumstances, both pre-test and then-test evaluations should be conducted. Although our interpretation of the above evidence would lead us to suggest that if only one evaluation can be conducted, then-tests can be used alongside post-tests at the end of a programme to determine change in participants. Given that there remains

uncertainty as to which methods produce the most accurate results in this field, we would strongly recommend careful combination of these assessments and engagement in ongoing research to determine procedures which are least vulnerable to confounding.

Author Contributions

OL, RC, KN and JG designed the work and acquired the data. All authors interpreted the data, drafted the manuscript and revised it critically for intellectual content. All authors approved the final version of the manuscript. All authors agreed to be accountable for all aspects of the work, ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Acknowledgements

Administrative support was provided for free by Wendy Washbourn and the Oxford Simulation, Teaching and Research (OxSTaR) centre.

Teaching sessions were delivered for free by Gionathan Amante, Rajinder Andev, Richard Canter, Tamsin Cargill, Clare Dollery, Mike English, Nick Fahy, João Galante, Sarah Gandhi, Richard Gleave, Helen Higham, Shona Johnston, Oscar Lyons, Karandeep Nandra, Manny Nijjar, Claire Pulford, Peter Sullivan, and Keith Willett.

Graduates of the Oxford University Hospitals Foundation Trust's Future Leaders Programme supported the Emerging Leaders Programme by facilitating workshops and mentoring projects.

Funding

The Oxford University Hospitals Foundation Trust's Department of Medical Education kindly provided rooms and catering for the workshops.

Oscar Lyons was supported during this work by a Rhodes Scholarship, a Goodger and Schorstein Research Scholarship (University of Oxford) and the Shirtcliffe Fellowship (Universities New Zealand)

Competing Interests

OL, KN, JG and RC contributed to the programme as course faculty. KN and JG are graduates of the first Emerging Leaders Programme.

Patient Consent for Publication

Not required

Data Availability

Available on request.

REFERENCES

1 West M, Armit K, Loewenthal L, *et al.* Leadership and Leadership Development in Health Care: The Evidence Base. *Kings Fund* 2015;:1–36. doi:19022015

2 Lyons O, George R, Galante J, *et al.* Evidence-based medical leadership development: a systematic review. *BMJ Lead* Published Online First: 2020. doi:10.1136/leader-2020-000360

3 Geerts JM, Goodall AH, Agius S. Evidence-based leadership development for physicians: A systematic literature review. *Soc Sci Med* 2020;246:112709. doi:10.1016/j.socscimed.2019.112709

4 Frich JC, Brewster AL, Cherlin EJ, *et al.* Leadership Development Programs for Physicians: A Systematic Review. *J Gen Intern Med* 2015;30:656–74. doi:10.1007/s11606-014-3141-1

5 Howard GS, Ralph KM, Gulanick NA, *et al.* Internal Invalidity in Pretest-Posttest Self-Report Evaluations and a Re-evaluation of Retrospective Pretests. *Appl Psychol Meas* 1979;3:1–23. doi:10.1177/014662167900300101

6 Campbell DT, Stanley JC, Gage NL. *Experimental and quasi-experimental designs for research*. Chicago: : Rand McNally 1969.

7 Hill LG. Back to the future: Considerations in use and reporting of the retrospective pretest. *Int J Behav Dev* 2020;44:184–91. doi:10.1177/0165025419870245

8 NHS Institute for Innovation and Improvement and Academy of Medical Royal Colleges. *Medical Leadership Competency Framework*. 2009.

9 Faculty of Medical Leadership and Management. Leadership and Management Standards for Medical Professionals. 2016.

10 Pendleton D, Furnham A. *Leadership : all you need to know*. Second edi. Basingstoke: : Palgrave Macmillan 2016.

11 Rabin R. Blended Learning for Leadership: The Centre for Creative Leadership Approach. 2014.

12 NHS Leadership Academy. Medical Leadership Competency Framework: Self assessment tool. 2012.

13 Kirkpatrick DL. Evaluation of training. In: Craig R., Bittel LR, eds. *Training and Development Handbook*. New York: : McGraw-Hill Book Company 1967.

14 Howard GS, Dailey PR. Response-shift bias: A source of contamination of self-report measures. *J Appl Psychol* 1979;64:144–50. doi:10.1037/0021-9010.64.2.144

15 Geldhof GJ, Warner DA, Finders JK, *et al.* Revisiting the utility of retrospective pre-post designs: The need for mixed-method pilot data. *Eval Program Plann* 2018;70:83–9. doi:10.1016/j.evalprogplan.2018.05.002

16 Pratt CC, Mcguigan WM, Katzev AR. Measuring program outcomes: Using retrospective pretest methodology. 2000. doi:10.1177/109821400002100305

17 Taylor PJ, Russ-Eft DF, Taylor H. Gilding the outcome by tarnishing the past: Inflationary biases in retrospective pretests. *Am J Eval* 2009;30:31–43. doi:10.1177/1098214008328517

18 Jansen RA, Rafferty AN, Griffiths TL. A rational model of the Dunning–Kruger effect supports insensitivity to evidence in low performers. *Nat Hum Behav* 2021;5:756–63. doi:10.1038/s41562-021-01057-0

19 Schwartz CE, Sprangers MAG. Guidelines for improving the stringency of response shift research using the thentest. *Qual Life Res* 2010;19:455–64. doi:10.1007/s11136-010-9585-9

20 Nolte S, Elsworth GR, Sinclair AJ, *et al.* The inclusion of 'then-test' questions in post-test questionnaires alters post-test responses: a randomized study of bias in health program evaluation. *Qual life Res* 2012;21:487–94. doi:10.1007/s11136-011-9952-1

21 Hill LG, Betz DL. Revisiting the retrospective pretest. *Am J Eval* 2005;26:501–17. doi:10.1177/1098214005281356

22 Pearson RW, Ross MA, Dawes RM. Personal recall and the limits of retrospective questions in

surveys. *Quest about Quest Inq into Cogn bases Surv* 1992;:65–94.

23 Nolte S, Elsworth GR, Sinclair AJ, *et al.* Tests of measurement invariance failed to support the application of the “then-test”. *J Clin Epidemiol* 2009;62:1173–80. doi:10.1016/J.JCLINEPI.2009.01.021

24 Simon HA. *Administrative Behavior: A study of decision-making processes in administrative organizations*. New York: : Macmillan 1947.

25 Lam TCM, Bengo P. A Comparison of Three Retrospective Self-reporting Methods of Measuring Change in Instructional Practice. *Am J Eval* 2003;24:65–80. doi:10.1177/109821400302400106

26 Nimon K, Zigarmi D, Allen J. Measures of program effectiveness based on retrospective pretest data: Are all created equal? *Am J Eval* 2011;32:8–28. doi:10.1177/1098214010378354