Insightful practice: a reliable measure for medical revalidation

Douglas J Murphy,[1] Bruce Guthrie,[1] Frank M Sullivan,[1] Stewart W Mercer,[2] Andrew Russell,[3] David A Bruce[4]

► Additional appendices are published online only. To view these files please visit the journal online (http://qualitysafety.bmj.com/content/21/8.toc).

[1]Quality, Safety and Informatics Research Group, University of Dundee, Dundee, UK
[2]Institute of Health and Wellbeing, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK
[3]Medical Directorate, NHS Tayside, Dundee, UK
[4]Postgraduate General Practice Education, NHS Education for Scotland, UK

Correspondence to
Dr Douglas Murphy, Senior Clinical Research Fellow, University of Dundee, Mackenzie Building, Kirsty Semple Way, Dundee DD2 4BF, UK;
d.y.murphy@dundee.ac.uk

Accepted 20 April 2012
Published Online First 31 May 2012

ABSTRACT

Background: Medical revalidation decisions need to be reliable if they are to reassure on the quality and safety of professional practice. This study tested an innovative method in which general practitioners (GPs) were assessed on their reflection and response to a set of externally specified feedback.

Setting and participants: 60 GPs and 12 GP appraisers in the Tayside region of Scotland, UK.

Methods: A feedback dataset was specified as (1) GP-specific data collected by GPs themselves (patient and colleague opinion; open book self-evaluated knowledge test; complaints) and (2) Externally collected practice-level data provided to GPs (clinical quality and prescribing safety). GPs' perceptions of whether the feedback covered UK General Medical Council specified attributes of a 'good doctor' were examined using a mapping exercise. GPs' professionalism was examined in terms of appraiser assessment of GPs' level of *insightful practice*, defined as: engagement with, insight into and appropriate action on feedback data. The reliability of assessment of *insightful practice* and subsequent recommendations on GPs' revalidation by face-to-face and anonymous assessors were investigated using Generalisability G-theory.

Main outcome measures: Coverage of General Medical Council attributes by specified feedback and reliability of assessor recommendations on doctors' suitability for revalidation.

Results: Face-to-face assessment proved unreliable. Anonymous global assessment by three appraisers of *insightful practice* was highly reliable (G=0.85), as were revalidation decisions using four anonymous assessors (G=0.83).

Conclusions: Unlike face-to-face appraisal, anonymous assessment of *insightful practice* offers a valid and reliable method to decide GP revalidation. Further validity studies are needed.

INTRODUCTION

Revalidation of practising doctors has prompted a wave of worldwide interest and remains a high-stakes challenge.[1] Doctors' capacity to self-regulate has been questioned,[2] but the measurement of quality of patient care is complex and agreement on a UK revalidation system has been problematic and implementation repeatedly delayed (currently scheduled for introduction from late 2012). Unfortunately, there is a sparse evidence base to inform its implementation.[3] Understandably, the public and government want clinically effective, safe and person-centred care delivered by competent and, ideally, excellent doctors.[4] In the UK, the domains and attributes required of Good Medical Practice have been defined (box 1).[5]

Revalidation aims to promote quality improvement as well as demonstrate a doctor being up to date and fit to practise.[5] Current proposals in the UK include an annual appraisal to check the quantity and quality of workplace and continuous professional development data collected over a 5-year cycle.[6] Satisfactory completion will lead to recommendation by an appointed Responsible Officer to the General Medical Council (GMC) for successful revalidation.[6] This moves appraisal from its current focus on supporting professional development to judging evidence.[7] Two issues need to be considered. First, continuous professional development has at its heart practitioners' ability to self-assess his or her educational needs. However, difficulties in recognising one's own (in) competence can lead to inflated or pessimistic self-assessments.[8] Second, there is no evidence that assessment at appraisal of this type is reliable enough for use in such high-stakes as revalidation.[9] As a possible alternative, formal examinations, such as those used by the American Board of Medical Specialties, could be used for revalidation in the UK, but knowledge on its own is unlikely to measure all the professional attributes of a doctor.[10]

To protect patients and ensure trust in doctors, we argue that we need a system of revalidation that is valid, reliable and

Box 1 General Medical Council domains and attributes of a doctor for appraisal and revalidation

Domain 1: knowledge, skills and performance
1. Maintain your professional performance.
2. Apply knowledge and experience to practice.
3. Ensure that all documentation (including criminal records) formally recording your work is clear, accurate and legible.

Domain 2: safety and quality
4. Contribute to and comply with systems to protect patients.
5. Respond to risks to safety.
6. Protect patients and colleagues from any risk posed by your health.

Domain 3: communication, partnership and teamwork
7. Communicate effectively.
8. Work constructively with colleagues and delegate effectively.
9. Establish and maintain partnerships with patients.

Domain 4: maintaining trust
10. Show respect to patients.
11. Treat patients and colleagues fairly and without discrimination.
12. Act with honesty and integrity.

supports reflective practice. Medical professionalism has been defined as a partnership between patient and doctor based on mutual respect, individual responsibility and appropriate accountability.[11] This definition formed the rationale for a new concept tested in this study: *insightful practice*. *Insightful practice* was defined as doctors' willingness to engage with and show insight into independent credible feedback on their performance and, where applicable, take appropriate action for improvement.

The aim in promoting *insightful practice* was to help individuals build beyond the conscientious collection and reflection of evidence to include independently verified outcomes for professional improvement. A doctor's professionalism and suitability for revalidation would be evidenced by testing his or her levels of *insightful practice* by measuring his or her willingness to engage with revalidation (responsibility and accountability); to show insight[12 13] into external feedback on his or her performance (mutual respect); and take action as needed to improve his or her patient care (partnership, responsibility and accountability). The study design took account of GMC attributes[4] and was further underpinned by GMC guidance to Post-Graduate Deans and GP Directors on professional remediation.[14] The GMC guidance advises that remedial training is only a practicable solution if a doctor demonstrates insight into his or her deficiencies and accepts that a serious problem exists, and that a remedial training programme can only be successful with the doctor's willingness and commitment.[14] In addition, the same guidance advises that, when deciding whether the doctor is suitable for remedial training, the panel should consider whether the doctor has insight into and is willing to address the problem.[14]

The purpose of this study was to test if:
1) Specified independent feedback (box 2) could validly cover necessary GMC attributes (box 1)[15]
2) Participants' level of *insightful practice* offered a reliable basis for making recommendations on revalidation.

METHODS

Included here is a summary of the methods. More information is available as a data supplement in the web appendices 1 and 2.[16 17]

This was a study which involved recruited general practitioners (GPs) collecting a suite of specified feedback on their performance. Participants completed a mapping exercise to test their agreement of the perceived validity of specified sources of feedback content at the start and end of the study. Participants received an appraisal from a GP colleague approved by the Health Board to help demonstrate their *insightful practice* by showing appropriate reaction to collected feedback. Doctors' success in showing *insightful practice* was subsequently assessed by the face-to-face appraiser and then again by three other anonymous appraiser assessors. The reliability of assessment of *insightful practice* (AIP) and subsequent recommendations on GPs' revalidation by face-to-face and anonymous assessors was investigated using Generalisability G-theory.[9] Decision (D) studies were conducted to determine the number of assessors required to achieve a reliability of 0.8, as required for high-stakes assessment.[9]

Participants and sample size calculation

Sixty-one participants were recruited from all GPs (n=337) within the National Health Service in Tayside in Scotland. Three information meetings were held, in different geographical locations, at the end of which GPs signed a register to confirm their interest in taking part. A consent form was then sent to each participant along with a covering letter and study information sheet.

Box 2 Study's suite of independent feedback

Personal feedback
1. Colleague (clinical and non-clinical) feedback: multisource feedback.
2. Patient feedback: patient satisfaction questionnaires.
3. Open book self-evaluated knowledge test.

Team feedback
4. Clinical governance data: prescribing safety and quality of care data.
5. Patient complaints.

Table 1 Summary of tools used and processes followed*

	Tool	Source	Prepared by
Multi-source feedback (MSF)*	General Medical Council (GMC) colleague survey[18 19] 2Q MSF[18 20]	GMC Developed by study author	Practice manager and colleagues
Patient satisfaction questionnaires*	GMC patient survey[18 19] Consultation and relational empathy[18 21]	GMC Developed by study authors	Patients and practice staff
Open book self-assessed knowledge test	Consisted of 60 items focusing on chronic disease management, referral issues and prescribing	Royal College of General Practitioners (RCGP Scotland)	GP undertook test
Prescribing safety data feedback†	12 measures of undesirable co-prescriptions[18 22]	Developed for study	Web-based report
Quality of care data feedback	Single area of interest selected for each participant's practice by an external assessor[18]	Quality outcome framework	Web-based report
Patient complaints	–	As received	Practice staff including GP

*For the purpose of the research study programme, participants collected and reflected on output from two patient satisfaction questionnaires and two MSF questionnaires, both on two occasions, in order to test the reliabilities of individual tools. In any real system, only one tool would be used and the collection of data would likely be spread over a longer period of time. The reliabilities of individual tools are not reported here.
†These data on 12 undesirable co-prescriptions were developed for the purpose of this study.[18 22] Other tools used are available to GPs to include when considering data for current appraisal submission.
GP, general practitioner.

Participating GPs received financial reimbursement: equivalent to 17 h extra payment per GP participant in addition to existing reimbursement for participation in the Health Board's existing statutory annual appraisal system. This additional payment was to allow for the estimated additional time commitment to collect the study's multiple sources of evidence on more than one occasion. The power calculation was based on Fisher's Z_R transformation of the intraclass correlation coefficient.[9] Given a required reliability intraclass correlation coefficient R of 0.8 for a high-stakes assessment of portfolios,[9] specified SE of the reliability of 0.05 and three assessors of each subject, Fisher's Z_R transformation specified a minimum of 46 subjects.

Performance measures and data collection

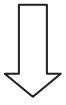
The study appraisal process was facilitated by a website called Tayside In-Practice Portfolio developed to administer, collect and assess all participant data,[18] making the allocation of tasks and feedback feasible. GPs were asked to collect specified data (patient and colleague feedback including complaints) and were also provided with feedback on their practice team's quality of care and prescribing safety (table 1). GPs were then asked to reflect on this specified suite of feedback in a portfolio to be submitted for appraisal.

Content validity of feedback

To ensure the content validity of the feedback in terms of the proposed suite of feedback covering the required GMC attributes,[5] each participant completed a mapping exercise of his or her perception (prestudy) and experience (poststudy) on each feedback tool's capacity to test the GMC attributes (see online appendix 1).

Study steps: reflection, appraisal and assessment

Step 1: Mapping exercise 1: June–July 2009 (online appendix 1).
This measured participant prestudy perceptions of the specified suite of feedback table 2.



Step 2: Collection of specified feedback: July–September 2009.
Study participants were provided with data via the study website including:
a. Colleague and patient feedback (existing available tools)
b. Report on undesirable co-prescriptions (developed for study)
c. Quality outcome framework data (currently used in UK General Practice System of Remuneration).

Some additional data were personally collected by participants:
d. Patient complaints
e. Self-evaluated knowledge test: developed by the Royal College of General Practitioners.

Step 3: Reflection on feedback and setting personal objectives for improvement (September–October 2009).
Having reflected on their performance feedback, participants used a reflective template with four 7-point Likert scales to rate each source of feedback data as having:

1) Highlighted important issues
2) Demonstrated concern in performance
3) Led to planned change
4) Given valuable feedback.

GPs then wrote a free-text commentary and framed any planned actions as Specific, Measurable, Achievable, Relevant and Timed (SMART) objectives (table 2).[23]

Step 4: Participants then received a face-to-face appraisal under the existing appraisal system, after which they had the opportunity to amend or add any personal objectives (October–December 2009).

Step 5: Assessment of participants' level of insightful practice by face-to-face appraiser postappraisal (October–December 2009).
Following the appraisal, the GP's appraiser rated the GP using an AIP template with four 7-point Likert scales. These related to GPs' *engagement* with the appraisal process, *insight* into the data collected, *planning of appropriate action* in response, and a global rating of their engagement, insight and action as a marker of GPs' *insightful practice*. Additionally, the appraiser was asked to assess whether the GP was 'on track for revalidation' (table 2).

Step 6: The anonymous postappraisal assessment of participants' level of insightful practice by three additional anonymous appraisers postappraisal was completed by the same process as in step 5 (October–December 2009).

Step 7: Mapping exercise 2: November 2009–January 2010 (online appendix 1).
This measured participant experience post study of the specified suite of feedback.

Reliability

The reliability of *insightful practice* as a measure was calculated using Generalisability G-theory following a web-based anonymous marking exercise after appraisal.[9] Anonymous assessors were recruited from study appraisers (n=5) and included one Deanery assessor. Two groups of assessors (n=3) each marked 30 GP portfolios (raters nested within group). Reliabilities (internal consistency and inter-rater) of anonymous assessor decisions for AIP (Questions 1–3) and inter-rater reliabilities, intraclass correlation coefficients, and the associated CIs were calculated for AIP Questions 4 and 5 using Generalisability G-theory.[9] Decision (D) studies were conducted to determine the number of assessors required to achieve a reliability of 0.8, as required in high-stakes assessment[9] (see online appendix 2).

Participant experience

Participants' evaluation of the provided suite of feedback was investigated by comparing four groups:

1. GPs with a satisfactory score (4 or above) in *insightful practice*.
2. GPs with an unsatisfactory score (<4) in *insightful practice*.
3. Face-to-face appraisers
4. Anonymous assessors.

Mean scores for each participant's rating of the value of each source of feedback were calculated and any significant differences between participant groups (1–4) examined using ANOVA with post hoc testing of differences.

RESULTS

Included here is a summary of the results. More information is available as a data supplement in the web appendices 1 and 2.

In all, 61 GP participants were recruited to the study. Of these, 60 were established independent GPs and one was a GP practice locum practitioner. Participants worked in a range of urban (n=48), accessible (n=9), and remote (n=3) practices.[24] Overall, 60 GPs (98.4%)

Table 2 Rating questions completed by general practitioner (GP) participants (preappraisal), by appraisers (after face-to-face appraisal) and by anonymous web-based portfolio assessors

Question	Rating scale	Completed by
Reflection template		
Source of feedback highlighted		
1. Important issues	Likert 1–7*	GP participant
2. Concern in performance		Face-to-face appraiser (preappraisal)
3. Led to planned change		
4. Gave valuable feedback		
Assessment of insightful practice template		
Doctor demonstrated		
1. Satisfactory engagement with the TIPP process	Likert 1–7*	Face-to-face appraiser (postappraisal)
2. Insight into the feedback provided on performance		Anonymous assessor (postappraisal)
3. Plans for appropriate action where applicable		
4. Engagement, insight and action (global rating of *insightful practice*)		
5. Suitability for recommendation as on track for revalidation without further opinion	Binary yes/no	► Face-to-face appraiser (postappraisal) ► Anonymous assessor (postappraisal)

*Likert scale descriptors (1–7): (1) strongly disagree; (3) disagree; (5) agree; (7) strongly agree.
TIPP, Tayside In-Practice Portfolio.

completed the study, with one dropping out after completing an initial content validity (mapping) exercise.

Mapping exercise

GP participants completed a mapping exercise of their perception (prestudy) and experience (poststudy) on each feedback tool's capacity to test the GMC attributes[5] (see online appendix 1).

Results for the poststudy mapping exercise are given in table 3.

Mean GP scores in the mapping exercise (1–7) for each GMC attribute (row) and tool (column) are given in table 3 with a score of 4 as the neutral point. All GMC attributes were covered (score>4) by at least one tool.

Reliability of participants' AIP as measured by face-to-face and anonymous assessors

There was a highly significant difference in the mean scores of global AIP (Q4) with face-to-face assessment scoring more highly than anonymous assessment (mean difference 1.07, 95% CI 0.73 to 1.41, t=6.29, 59 df,

Table 3 Mean general practitioner (GP) ratings of perceived ability of each feedback tool (columns) to assess the 12 General Medical Council (GMC) attributes (rows) after feedback received. Scale (1–7) for each GMC with a score of 4 as a neutral point*

The GP…	Colleague feedback	Patient feedback	Practice performance data	Knowledge test	Patient complaints
Maintains professional competence	**5.3**	4.4	*3.8*	**4.6**	*3.3*
Applies knowledge and experience to practice	**5.1**	3.9	4.0	**4.7**	*2.6*
Keeps clear, accurate and legible records	**4.8**	*2.0*	2.9	*1.6*	3.2
Puts into effect systems to protect patients and improve care	**4.8**	*3.1*	**4.1**	*2.7*	*3.4*
Responds to risks to safety	**4.5**	2.9	3.3	2.8	3.1
Protects patients and colleagues from any risk posed by his/her health	**4.7**	2.4	1.8	1.7	2.3
Communicates effectively	**5.7**	**5.9**	*2.3*	*2.0*	4.1
Works constructively with colleagues and delegates effectively	**6.1**	2.7	2.9	*1.9*	2.9
Establishes and maintains partnerships with patients	5.0	**5.9**	*2.2*	*1.7*	3.9
Shows respect for patients	**5.4**	**6.0**	*1.9*	*1.8*	4.3
Treats patients and colleagues fairly and without discrimination	**5.9**	5.1	*1.8*	*1.8*	3.8
Acts with honesty and integrity	**5.7**	4.8	*2.2*	*1.9*	3.7

*Tools or groups of tools significantly different from the rest as being the **most highly valued** for each attribute are represented in **bold** font. Tools or groups of tools significantly different from the rest as being the *least highly valued* for each attribute are represented in *italic* font (p=0.05).

Table 4 Reliability of assessment of insightful practice (AIP) questions 1–5

	AIP questions 1–3 (engagement, insight and action) 1–7 scale reliability (G)		AIP question 4 (global assessment) 1–7 scale reliability (G) (ICC)*		AIP question 5 (binary yes/no recommendation on revalidation) reliability (G) (ICC)*	
Raters	**Internal consistency**	**Inter-rater**	**Inter-rater†**	**Inter-rater (95% CI)‡**	**Inter-rater**	**Inter-rater (95% CI)***
1	0.94	0.71	0.66	–	0.54	–
2	0.96	**0.83**	0.79	(0.68 to 0.88)	0.7	(0.54 to 0.83)
3	0.96	**0.88**	**0.85**	**(0.78 to 0.91)**	0.78	(0.69 to 0.86)
4	0.97	**0.91**	**0.89**	**(0.84 to 0.93)**	**0.83**	**(0.75 to 0.89)**
5	0.97	**0.92**	**0.91**	**(0.87 to 0.94)**	**0.86**	**(0.80 to 0.91)**
6	0.97	**0.94**	**0.92**	**(0.89 to 0.95)**	**0.88**	**(0.83 to 0.92)**

Reliabilities greater than 0.8, as required for high-stakes assessment, are given in bold.[9]
*Intraclass correlation coefficients (ICCs) are G coefficients when you have a one facet design (rater).
†Inter-rater reliability is the extent to which one rater's assessments (or when based on multiple raters, the average of raters' assessments) are predictive of another rater's assessments.
‡95% CIs for reliabilities (ICCs) were calculated using Fisher's Z_R transformation which is dependent on raters (k) with a denominator value of (k-1), and so cannot be calculated when there is only one rater.[9]

p<0.001). Dichotomous judgment on GPs' suitability for revalidation (AIP Q5) also revealed significant differences between face-to-face and anonymous assessment. No portfolio was considered unsatisfactory at face-to-face assessment, while 42/180 (23.3%) of the three anonymous markings of each of the 60 portfolios were considered unsatisfactory (χ^2, value 16.97: p<0.001). Face-to-face appraisal did not discriminate between GPs and therefore could not be classed as reliable. In contrast, high reliability was demonstrated by anonymous global assessment by three assessors (G= 0.85) of GPs' *insightful practice*. A recommendation on GPs' suitability for revalidation was also highly reliable by four assessors (G=0.83) (table 4, online appendix 2).

Participant experience

The four groups of participants rated the suite of five feedback sources positively (mean value rating over all feedback tools for each participant group above a neutral score of 4), with anonymous assessors giving significantly higher ratings than other groups (mean 5.4 vs 4.7–4.9, p=0.05) (table 5).

DISCUSSION

Summary

This study demonstrates that a valid suite of independent feedback covering necessary GMC attributes can be created for use in GP appraisal and revalidation. Doctors' *insightful practice*, measured by GPs demonstrating accountability for making quality improvement where needed, offers a reliable basis for a recommendation on revalidation.

Context

A system of revalidation is needed that is valid and reliable.[23] Revalidation goals appear to include restoring public trust, promoting quality improvement and identifying doctors in difficulty, but there is a sparse evidence base to inform the introduction of an agreed system.[3] This is the first study of which we are aware to formally use medical professionals' *insightful practice* as a proxy of workplace-based performance and to include a form of knowledge testing, an element of competency testing demanded by the Shipman Inquiry.[2] Study methods were robust and the tested system included recently developed and innovative reliable indicators on high risk prescribing for participants to reflect on practice improvement.[23]

Interpretation

This work contributes to the limited evidence in this important area for both public and profession.[3 25] The proposed role of *insightful practice* is to act as the hub within a continuous cycle to generate, monitor and maintain objective evidence of personal responsibility

Table 5 Mean scores for reflective template questions (1–4) for feedback sources for each group (n=4)

Reflective template question	Groups	Mean RT score over all feedback tools (95% CI)
Value of feedback	GPs with unsatisfactory *insightful practice* global assessment	4.9 (4.6 to 5.2)
	GPs with satisfactory *insightful practice* global assessment	4.7 (4.6 to 4.9)
	Face-to-face appraisers	4.7 (4.4 to 5.0)
	Anonymous assessors	5.4 (4.9 to 5.9)

GP, general practitioner; RT, reflective template.

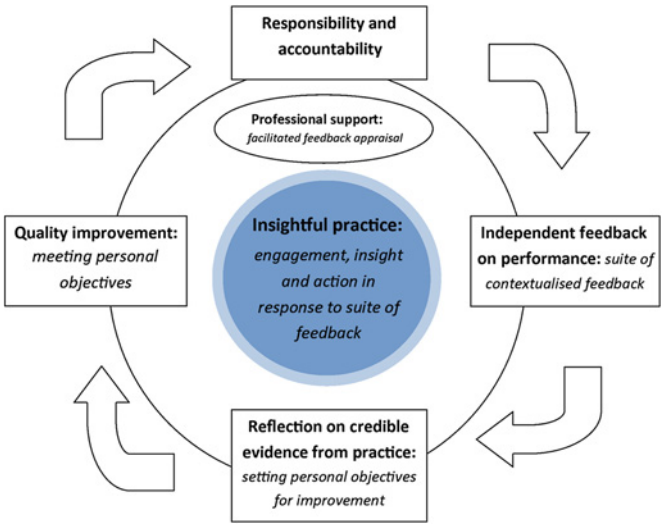


Figure 1 Cycle of insightful practice.

and accountability for quality improvement as needed (figure 1).

The nature of reflective practice makes its quantification a challenge.[26] Doctors' capacity to show insight, overcome challenges and incorporate new behaviours and attitudes have previously been described as a fundamentally personal and subjective concept called mindful practice.[27] Reflection and facilitation are known to prove useful for the assimilation of feedback and acceptance of change.[28] *Insightful practice* is arguably a useful conceptual development, which both lends itself to the reliable measurement of objective outcomes and combines the subjective consideration of self-perceptions with the reflections and facilitation by others on needed insight for improvement. In addition, the study's combination of feedback from multiple methods, reflection and mentoring is consistent with the call for innovation in assessing professional competence and shows how assessment instruments might be used together to promote performance improvement.[29 30] By placing a focus on productive reflection (engagement and insight) and needed action (life-long learning and appropriate response to performance feedback), measurement of *insightful practice* may also offer an answer to the call for innovation in measuring professionalism to cover previously poorly tested areas of seeking and responding to feedback and results of audit.[31]

A challenge for revalidation will be whether the system benefits all doctors, while still identifying those at risk of poor performance. If adopted, the tested system could meet this challenge by early and reliable identification of doctors' level of, and progress with, improvements in care, as well as allowing the monitoring of progress towards satisfactory revalidation. The collection of specified data is feasible if spread over the proposed 5-year cycle. The role, frequency and targeting of appraisal would need further consideration should such a system be implemented, with a possible reduction in ongoing scrutiny and support for those doctors shown to be 'on track'. In cases of unsatisfactory progress, early identification would give maximum opportunity to target professional support (figure 1). No system guarantees identification and protection from criminal behaviour, but valid and reliable external monitoring should help to reassure the public in the quality and safety of their doctors. The participant mapping exercise gave evidence of content validity of the specified feedback. The subsequent agreement between participants that the suite of feedback was of value added further face validity to the system. Anonymous assessors' significantly higher rating of the value of the suite of feedback possibly reflected its help in quantification and discrimination of those assessed. It is interesting that opinion of patient and self-evaluated knowledge testing feedback both improved significantly with experience.

Limitations

This study had limitations and there is a need for significant further research. The assessors' role and process of making judgements in a 'live' system of revalidation will need to be explicit to inform further research. While many health professionals believe that more objective is equivalent to better, this is not always the case. Much research in medical education has suggested that expertise is not always characterised by comprehensiveness. As a result, assessment processes that are scored by simple frequency counts of whether or not particular actions were taken tend to be less valid indicators of performance than more subjective global ratings provided by informed raters.[32] This concept underpinned this study's investigation of *insightful practice* as a possible foundation for revalidation recommendations.

While reliabilities reported in this study were generalised across assessors, using G-theory and associated D studies,[9] the participants were limited to GPs in a single region of Scotland. Future research needs to focus on the capacity of *insightful practice* to offer reliable and valid measure in the performance across other settings and specialties as the measurement properties of every instrument are specific to the population on which the instrument is tested.[9]

In addition, although the literature supports *insightful practice* as a proxy measure for successful performance improvement,[11–14] the construct validity of this was not possible to test within this study. Engagement in appraisal is needed to promote improved clinical management,[33] and GMC recommendations on remediation highlight the importance of insight and capacity to address problems.[14] Although there is evidence that

well-founded and well-planned change is still a reasonable surrogate for successful implementation,[34] it was not possible in this study to track whether GPs' SMART personal objectives were carried through.[23] This requires further research to demonstrate.

CONCLUSIONS

The real test of revalidation will be whether its introduction leads to improvement in the quality and safety of healthcare. Further research will be needed, but public trust in doctors requires them to be held to account for their own performance and urgent progress is long overdue. The appraisers' role in revalidation could lie among coaching, educational advocate and supporter at one end, and assessor accountable for revalidation and the quality of its outcome at the other. This study's findings suggest that a single face-to-face appraiser is unlikely to be able to make a valid or reliable judgement about fitness for revalidation, but that anonymous measurement of *insightful practice* offers an alternative platform from which a robust system of revalidation could be developed and implemented.

Acknowledgements We thank all the general practitioners and general practice appraisers who took part in the study; programmers Jill JeanBlanc and Keith Milburn who helped develop materials and the study website; and Selene Ross who acted as the study administrator.

Contributors All authors contributed to the design of the study. DJM, AR and DAB recruited the study participants. DJM analysed all data. DJM managed the literature review and wrote the manuscript. All authors contributed to the interpretation of the findings and the critical revision of the manuscript for intellectual content and were involved in the decision to submit the manuscript for publication. DJM acts as guarantor for the study.

Funding The study was funded by the Chief Scientist Office (CSO) Scottish Government, Royal College of General Practitioners (RCGP), NHS Education for Scotland (NES) and Scottish Patient Safety Research Network (SPSRN). DM, BG and FS are employed by University of Dundee. SM is employed by the University of Glasgow; AR is employed by NHS Tayside, and DB by NHS Education for Scotland. All authors had full access to all the data and agreed responsibility for the decision to submit for publication independently from any funding source. DM is supported by a Primary Care Research Career Award from the Chief Scientist Office, Scottish Government.

Competing interests None.

Ethics approval Formal application and submission of the research proposal was made and ethical approval granted for all of the work contained in this paper by the Tayside Committee on Medical Research Ethics A. Participants gave informed consent before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. Villanueva T. Revalidation wave hits European doctors. *CMAJ* 2010;182:E463–4.
2. Smith J. *The Shipman enquiry—Fifth Report: Safeguarding Patients: Lessons From the Past—Proposals for the Future*. 2004. http://www.shipman-inquiry.org.uk/fifthreport.asp (accessed 4 Aug 2011).
3. Greenhalgh T, Wong G. Revalidation: a critical perspective. *Br J Gen Pract* 2011;584:166–8.
4. The Scottish Government. *The Healthcare Quality Strategy for NHS Scotland*. 2010. http://www.scotland.gov.uk/Resource/Doc/311667/0098354.pdf (accessed 4 Aug 2011).
5. General Medical Council. *GMP Framework for Appraisal and Revalidation*. http://www.gmc-uk.org/doctors/revalidation/revalidation_gmp_framework.asp (accessed 5 Aug 2011).
6. General Medical Council. http://www.gmc-uk.org/Revalidation_The_Way_Ahead.pdf_32040275.pdf (accessed 4 Aug 2011).
7. Royal College of General Practitioners RCGP. http://www.rcgp.org.uk/PDF/PDS_Guide_to_Revalidation_for_GPs.pdf (accessed 4 Aug 2011).
8. Kruger J, Dunning D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J Pers Soc Psychol* 1999;77:1121–34.
9. Streiner DL, Norman GR. *Health Measurement Scales*. 3rd edn. Oxford: Oxford Medical Publications, 2003.
10. Wu J. A piece of my mind. Recertification. *JAMA* 2010;303:309–10.
11. Working Party of the Royal College of Physicians. Doctors in society: medical professionalism in a changing world. *Clin Med* 2005;5(Suppl 1):S5–40.
12. Hays RB, Jolly BC, Caldon LJ, *et al*. Is insight important? Measuring capacity to change performance. *Med Educ* 2002;36:965–71.
13. Hixon JG, Swann WB. When does introspection bear fruit? Self-reflection, self-insight, and interpersonal choices. *J Pers Soc Psychol* 1993;64:35–43.
14. General Medical Council. http://www.gmc-uk.org/Guidance_for_making_referrals_to_the_Postgraduate_Dean.pdf_25416687.pdf (accessed 4 Aug 2011).
15. RCGP Report on Tayside Revalidation Study. http://www.rcgp.org.uk/PDF/RCGP_Report_on_Tayside_Final_with_Abstract.pdf (accessed 4 Aug 2011).
16. Murphy DJ, Bruce DA, Eva KW. Workplace-based assessment for general practitioners: using stakeholder perception to aid blueprinting of an assessment battery. *Med Educ* 2008;42:96–103.
17. Brennan RL. http://www.education.uiowa.edu/casma/GenovaPrograms.htm (accessed 4 Aug 2011).
18. Tayside In-Practice Portfolio. https://www.tipportfolio.co.uk/tipp/TIPPPaper.aspx (accessed 4 Aug 2011).
19. Campbell JL, Richards SH, Dickens A, *et al*. Assessing the professional performance of UK doctors: an evaluation of the utility of the General Medical Council patient and colleague questionnaires. *Qual Saf Health Care* 2008;17:187–93.
20. Murphy DJ, Bruce DA, Mercer SW, *et al*. The reliability of workplace-based assessment in postgraduate medical education and training: a national evaluation in general practice in the United Kingdom. *Adv Health Sci Educ* 2009;14:219–32. http://dx.doi.org/10.1007/s10459-008-9104-8
21. Mercer SW, Maxwell M, Heaney D, *et al*. The development and preliminary validation of the Consultation and Relational Empathy (CARE) Measure: an empathy-based consultation process measure. *Fam Pract* 2004;21:699–705.
22. Guthrie B, McCowan C, Davey P, *et al*. High risk prescribing in primary care patients particularly vulnerable to adverse drug events: cross sectional population database analysis in Scottish general practice. *BMJ* 2011;342:d3514.
23. Blanchard K, Zigarmi P, Zirgarmi D. *Leadership and the One Minute Manager: S.M.A.R.T. goals*. http://www.primarygoals.org/books/OneMinuteManager.htm (accessed 14 Feb 2012).
24. The Scottish Government. *Urban Rural Classification 2009-2010*. http://www.scotland.gov.uk/Topics/Statistics/About/Methodology/UR2010 (accessed 4 Aug 2011).
25. Bruce DA, Philips K, Reid R, *et al*. Revalidation for general practitioners: randomised comparison of two revalidation models. *BMJ* 2004;328:687–91.
26. Mann K, Gordon J, MacLeod A. Reflection and reflective practice in health professions education: a systematic review. *Adv Health Sci Educ Theory Pract* 2009;14:595–621.
27. Epstein R. Mindful Practice. *JAMA* 1999;282:833–9.
28. Sargent JM, Mann KV, van der Vleuten CP, *et al*. Reflection: a link between receiving and using assessment feedback. *Adv Health Sci Educ Theory Pract* 2009;3:399–410.
29. Epstein RM, Hundert EM. Defining and assessing professional competence. *JAMA* 2002;287:226–35.
30. Miller A, Archer J. Impact of workplace based assessment on doctors' education and performance: a systematic review. *BMJ* 2010;341:c5064.
31. Wilkinson TJ, Wade WB. A Blueprint to assess professionalism: results of a Systemic review. *Acad Med* 2009;84:551–8.
32. Hodges B, Regehr G, McNaughton N, *et al*. OSCE checklists do not capture increasing levels of expertise. *Acad Med* 1999;74:1129–34.
33. Spurgeon P, Barwell F, Mazelan P. Developing a medical engagement scale. *Int J Clin Leadersh* 2008;16:213–23.
34. Wakefield JG. Commitment to change: exploring its role in changing physician behaviour through continuing education. *J Contin Educ Health Prof* 2004;24:197–204.

Insightful Practice: a reliable measure for medical revalidation

APPENDICES (web supplement files)

APPENDIX 1

METHODS

Mapping exercise

The content validity of each type of feedback was examined by participants completing a mapping exercise of each feedback tool's ability to test a doctor's alignment with GMC required attributes. Participating GPs were asked to rate perceived ability of feedback tools (n=5) to test the 12 attributes of Good Medical Practice using a 7-point Likert scale.[5,16] The mapping exercise was completed at the outset and on completion of the study to see if perceptions changed with experience. Descriptive statistics, ANOVA with associated Post–Hoc tests and Generalisability G- theory[9] were used to assess GP agreement and coverage of desired attributes by tools, with changes in perceptions at the beginning and end of the study examined using paired t-tests.

RESULTS

Mapping exercise

Mean GP scores in the mapping exercise (1-7) for each GMC attribute (row) and tool (column) are given in table 3 with a score of 4 as the neutral point on the Likert scale. Inter-rater reliability (participant agreement) was extremely high at 0.99, both before and after TIPP participation. For each GMC attribute (row) tested, there were significant differences identified in participants' assessment on the ability of different tools to test each attribute (P=0.001). Post-hoc tests (Tukey, Tukey's-b) were used to investigate where these significant differences between feedback tools were for each GMC attribute.

MSF was the tool most expected by participants to be the best test for 11/12 attributes. Patient satisfaction questionnaires were perceived to test communication, patient

partnership and respect best. Practice data were expected to test systems to protect patients and improve care. Conversely, knowledge testing was not expected to test any attribute well pre-study but, after experience of it in the study, it was thought to test application of knowledge, experience and the maintaining of professional performance. The suite of feedback was perceived as testing the required spectrum of GMC attributes with at least one tool rated above the neutral point of four for every attribute.

Paired t tests comparing pre- and post-study scores only showed significant differences for patient questionnaires (mean difference 0.17, 95%CI 0.03 to 0.30, t=2.78, 11df, p=0.02) and knowledge tests (mean difference 0.28, 95%CI 0.15 to 0.41, t=4.7, 11df, p=0.001), with participants valuing both more highly in the light of experience of using them.

APPENDIX 2

METHODS

Reliability of Assessment

Descriptive statistics were calculated using SPSS. Reliabilities (internal consistency and inter-rater) of anonymous assessor decisions for AIP (Questions 1-3, 4 and 5) were assessed using Generalisability G- theory and GENOVA.[9,17] 95% confidence intervals for reliabilities were calculated using Fisher's Z_R transformation.[9] Reliability is denoted by a number between zero and one and indicates the proportion of the variance in scores that can be attributed to true differences, as opposed to measurement error. Internal consistency refers to the extent to which the items within the instrument provide consistent information. Inter-rater reliability indicates the extent to which one rater's assessments are predictive of another rater's assessments. Decision (D) studies were conducted to determine the number of assessors required to achieve a reliability of 0.8, as required in high-stakes assessment.[9]

RESULTS (Table 4)

Reliability of Assessment

Internal consistency and inter-rater reliabilities (AIP questions: 1-3) and inter-rater reliability and associated confidence intervals (AIP questions: 4 and 5) were calculated for assessors' judgement on participants' portfolios for a specified number of assessors. Anonymous assessment of satisfactory portfolio performance was highly reliable (G>0.8, suitable for high-stakes assessment)[9] using a 1-7 Likert scale given only two assessors (AIP Q1-3: elements of *insightful practice*) and three assessors (AIP Q4: global rating of *insightful practice*). Dichotomous judgement of the suitability of GPs for revalidation was also highly reliable with four assessors (AIP Q5). Portfolio marking using AIP required a range of 15-45 minutes for each portfolio for five of the six assessors, with the sixth requiring longer than 90 minutes per portfolio.