

.....  
General Practice Specialist Training

# Workplace Based Assessment for nMRCGP

*External Tools Pilot Report  
2005-2006*



COGPED



•  
•  
•  
•  
•  
•  
•  
•

---

## Acknowledgements

The completion of the pilot is thanks to the help and enthusiasm of all staff in and 171 GP registrars from the Wales, Northern Ireland, Mersey, KSS, East Scotland, North and North East Scotland, South East Scotland and West Midlands Deaneries.

The authors would like to thank Mrs Angela Inglis (Team Leader and Personal Assistant to Dr David Bruce, GP Director in the East of Scotland Deanery) and her team, (Lee-Ann Troup, Linda Kirkcaldy, Susan Smith, Carol Ironside and Gill Ward) for their help, support, and contribution to the work contained in this report.

## Authors

Douglas J. Murphy, David A. Bruce, Kevin W. Eva

© MSF Tool – NHS Education for Scotland



# Contents

---

1.	Executive Summary	Page 4
2.	Background to the Pilot	Pages 8 to 10
3.	Statistical Methods	Page 11
4.	Piloted Assessment Tools	Pages 13 to 55
	<ul style="list-style-type: none"><li>• <i>Patient Satisfaction Questionnaire (PSQ)</i></li><li>• <i>Multi Source Feedback (MSF)</i></li><li>• <i>Video</i></li><li>• <i>Referrals</i></li><li>• <i>Criterion Audit</i></li><li>• <i>Significant Event Analysis</i></li></ul>	
5.	Summary of Results	Page 56
6.	Conclusions	Page 57 to 58
7.	Appendices	Page 59



# Executive Summary

Following a literature review and blueprinting exercise involving educational experts, GP trainers, GP registrars and patient representation, six potential workplace based assessment tools were piloted to determine their value in performance testing of GP registrars. The reliability, validity, feasibility, acceptability and educational impact of each tool were measured.

During the pilot year from August 05 to July 06, 171 GP registrars and their GP trainers, 64 assessors, and medical and administrative staff from 9 Deaneries in all four UK countries, participated and completed an allocated selection of the assessment tools.

The assessment tools piloted were; Video of Consultation, Significant Event Analysis (SEA), Criterion Audit, Multi-source Feedback (MSF), Patient Satisfaction Questionnaire (PSQ) and Analysis of Referrals.

Previously validated tools were used for SEA, Criterion Audit and the PSQ. The video tool was constructed from items from both the Membership of the Royal College of General Practitioners (MRCGP) and Summative Assessment video tools. The MSF and Analysis of Referrals tools were designed and piloted by the project steering group.

Results were collated by East of Scotland Deanery. Psychometric analysis was completed within NHS Education for Scotland and at McMaster University, Ontario Canada. Qualitative data was gathered from participant questionnaires completed in each Deanery and from individual interviews conducted within a number of Deaneries

The MSF and PSQ tools had high reliability (MSF - G 0.75 clinical raters / G 0.81 non-clinical raters; PSQ - G 0.75). Both tools were judged to be acceptable and helpful to GPR learning. The MSF and PSQ tools meet accepted psychometric standards for high stakes assessment. In addition in combination with the Acquired Knowledge Test (AKT) and Clinical Skills Assessment (CSA), both tools were perceived by educationalists, GP trainers and GP registrars to cover that same range of qualities as when all six tools were tested.

Video, SEA, Criterion Audit and Analysis of Referrals did not demonstrate the capability of informing a high stakes decision. However these tools were valued by GP Registrars' and GP Trainers' and still have a potential role in teaching and training.

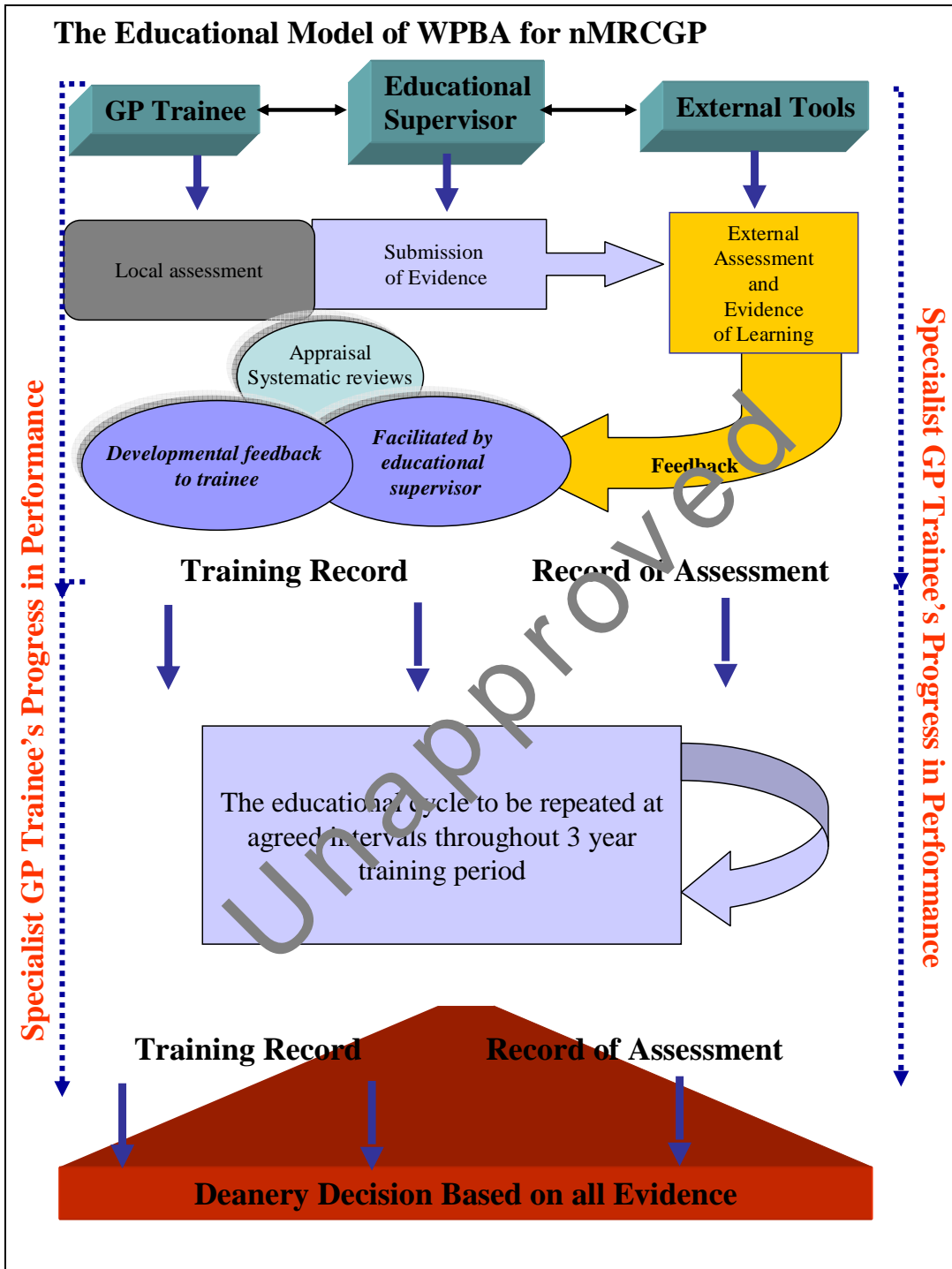
Feedback provided from the assessment tools was helpful to both the GPR and GP trainer. A suggested WPBA process has been developed by the national group which integrates assessment and learning (Table 1).

Results from the blueprinting exercise demonstrate that using the high reliability WPBA tools along with the AKT and CSA (both demonstrating high reliability in other contexts), will measure the qualities required for general practice and the spectrum of the RCGP curriculum. A suggested WPBA framework is therefore proposed (Table 2).

It should be noted that the findings of this pilot are specific to doctors in GP specialist training and can not be generalised to other groups without further testing in the appropriate context. It is likely however, that as GP registrars are a similar group in age and experience, testing with the PSQ and MSF in more heterogeneous groups is likely to be even more able to discriminate between individual's performance. If proven, these measures may have other applications in re-licensure and re-accreditation of practicing doctors

Unapproved

Table 1



**Table 2**  
**WPBA - PROPOSED ASSESSMENT FRAMEWORK**

	HOSPITAL Year 1	HOSPITAL Year 2	PRACTICE Year 3
<b>External Tools</b> <b>Core Framework</b> <b>WPBA</b> <b>UK CORE WPBA TOOLS</b>	<b>PSQ X 1</b>  <b>MSF X 2</b> <b>clinical raters</b>	<b>PSQ</b> <b>Re-sit option</b>  <b>MSF</b> <b>Re-sit option</b>	<b>PSQ X1</b>  <b>MSF X 2</b> <b>clinical &amp; non-clinical raters</b>
<b>Training Record</b> (additional assessment tool requirements)	CBD	CBD	CBD COT (video)
<b>Additional Teaching &amp; Training Resources</b>  (local feedback where applicable)	(Possible locally decided application) DOPS (Possible locally decided application) mini CEX		DOPS    N/A GP context  mini CEX  SEA Referrals Audit

KEY	CBD	Case based discussion
	COT	Consultation observation tool
	CEX	Clinical evaluation exercise
	DOPS	Direct observation procedural skills
	MSF	Multi-source feedback (360 degree appraisal)
	PSQ	Patient satisfaction questionnaire
	SEA	Significant event analysis

⋮

---

# Background

---

General Practice Specialist Training (GPST) starting August 2007 will, for the first time, comprise a three year managed educational programme based in general practice and underpinned by the RCGP curriculum. GPST programmes, GP trainers, the curriculum and assessments and will require approval by the Postgraduate Medical Education and Training Board (PMETB). Assessment of GPST will involve three parts; an applied knowledge test (AKT), a clinical skills assessment (CSA) and Workplace based assessment (WPBA). These assessments will form the new Membership of the Royal college of General Practitioners (nMRCGP) and successful completion of all three parts will lead to the award of a Certificate of Completion of Training (CCTs).

## ***Workplace assessment***

While performance (what the learner does) is related to both knowledge (what the learner knows) and competency (what the learner can do), it is also influence by other factors (attitudes, fatigue, circumstances etc.). Workplace assessment therefore aims to measure the behaviours or actions of the GP registrar in day to day working.

## ***The purpose of workplace assessment is:***

- To provide an accurate measure of the learner's performance at work
- To assess whether the learner's clinical practice achieves agreed standards for independent practice
- To provide the learner with feedback and promote reflective practice

It is recognised that the implementation of a system of assessment with impact on the care of patients and careers of individuals would be unethical if not piloted to reassure stakeholders that the outcomes are valid and reliable.

An ideal system of assessment will (a) provide reproducible assessments, allowing inferences to be drawn that are valid, and (b) take into account feasibility and acceptability, thereby creating a system that is less likely to be challenged. In addition it will be important to take account of (c) the consequences and impact of the assessment process on patients and the profession. To be meaningful and encourage acceptance, educational feedback should be provided to all doctors on the spectrum of their progress and performance over time. An understanding of factors and principles of assessment is required if the implementation of a flawed system with adverse consequences is to be avoided.

It is generally agreed that an overall reliability co-efficient of  $>$  or  $=$  to 0.75 is required for a high stakes assessment such as licensure for practice<sup>1</sup>.

The assessment tools developed and tested for this pilot took account of the above and the following important principles of assessment:

**a) Multiple samples:** Given that an individual's performance on any one problem is non-predictive of that individual's performance on other problems, requires adoption of multiple sampling approaches to measurement.<sup>2</sup> Just as one would never trust a single multiple-choice question to provide an accurate indicator of knowledge, performance indicators should be sampled with sufficient breadth of content and context.

**b) Multiple tools:** No single assessment tool can determine an individual's level of competence in every role demanded in the practices of medicine.<sup>3</sup> As a result, a single assessment by, for example, multiple-choice testing alone would be simplistic. Appropriate testing will require a range of tools to inform an overall reproducible judgement of the doctors' performance for it to be fair and trustworthy.

**c) Population specificity:** Decisions should not be made based on the result of the assessment protocol until reliability and validity testing is performed on the population of interest, because the measurement properties of every instrument are specific to the population on which the instrument was tested.<sup>1</sup> For example, a tool known to be a reliable and valid indicator of performance (with identified levels of minimum competence) for GP registrars (UK physicians training in general practice) may or may not be a reliable and valid indicator of performance for independent general practitioners. The latter group is more heterogeneous in age, profile, and experience and, as a result, may reveal a different pattern of scores. The authors are aware that bluntly transferring assessment tools across groups may yield inappropriate inferences, to the detriment of the older doctor and/or the public. This pilot hopes to provide lessons which can be researched in other contexts with the relevant population(s).

**d) Blueprinting of content of assessment:** The implementation of an assessment system is likely to be enhanced by stakeholder agreement that appropriate qualities are being tested. Pre-pilot, educationalists, general practice trainers and registrars completed a blueprinting exercise (appendix 1) to rate the extent to which each of the evaluation tools was perceived to assess each of eight competencies largely derived from the General Medical Council (GMC) document Good Medical Practice.

.....

There were high levels of agreement amongst stakeholders of the perceived qualities tested by the proposed tools ( $G = 0.82$  to  $0.93$ ).

**e) Ethical Approval:** Formal application and submission of the research proposal was made and ethical approval granted for all of the work contained in this report by NHS Ethics Committee (Glasgow West).

Assessment tools should not be developed using old data unrepresentative of current population performance if meaningful inferences are to be drawn. For example, a tool may be reliable enough to detect differences in individuals' performance prior to training or knowledge of an assessment's content but not able to differentiate differences in a reliable manner when the task is explained.

It is against this background that the opportunity to develop and pilot multiple assessment tools for specialist GP training was started.

It is hoped that the results of this pilot will be important in underpinning recommendations for Specialist GP Training and also point to future directions for development for national and international licensing authorities for other groups of doctors.

## References

1. Streiner DL, Norman GR. Health Measurement Scales (3<sup>rd</sup> ed.). Oxford Medical Publications, 1995.
2. Eva KW. On the generality of specificity. Medical Education 2003; 37: 587-588.
3. Schuwirth LWT, van der Vleuten C. Changing education, changing assessment, changing research. Medical Education 2004; 38: 805-812.

# Statistical Methods

---

All the data from the participating deaneries were analysed using analysis of variance (SPSS and GENOVA software). Intra-class correlation co-efficient for forms of reliability were calculated using generalisability theory. Decision studies were conducted to determine the optimum number of questionnaires required for a high stakes assessment

The design, methodology and analyses were conceived and carried out by Douglas J Murphy, NHS Education Scotland and Kevin W Eva, McMaster University, Ontario, Canada.

Unapproved



# Piloted Assessment Tools

Each individual tool is discussed in turn. The assessment tools are described with reference to their utility i.e validity, reliability, acceptability, feasibility, and educational impact.<sup>1</sup>

1. Patient Satisfaction Questionnaire (CARE)	pages 13 to 19
2. Multi-Source Feedback (MSF)	pages 20 to 28
3. Video of Consultations	pages 29 to 35
4. Assessment of Referral Letters	pages 36 to 42
5. Criterion Audit	pages 43 to 48
6. Significant Event Analysis	pages 49 to 55

Unapproved

1. van der Vleuten, CPM. The assessment of professional competence: developments, research and practical implications. *Advances in Health Sciences Education* 1996 1: 41-67.

# Patient Satisfaction Questionnaire (CARE)

## Research Question

Does the CARE patient satisfaction questionnaire provide a valid, reliable, acceptable, and feasible measure of workplace-based performance to inform a high stakes decision on individuals' outcomes for general practice specialist training?

## Background and Validity

The Consultation and Relational Empathy (CARE) Measure was developed by Dr Stewart Mercer a Senior Clinical Research Fellow in the Department of General Practice and Primary Care at University of Glasgow. CARE is a questionnaire of 10 questions testing patient opinion on a scale of 1 - 7 on the outcome of their experience of contact with a physician.

The CARE measure has been developed and validated in primary care<sup>1</sup>. The CARE measure has also been piloted in secondary care by the Scottish Executive Health Department (Centre for Change and Innovation) across a range of specialties<sup>2</sup>. Recent studies have also endorsed its validity in other healthcare settings<sup>3,4</sup>.

In the blueprinting exercise designed to examine the perceived validity for the purpose of the external tools WPBA pilot, patient satisfaction was perceived by stakeholders to test the *Good Medical Practice* quality of *Relationship with Patients*.

## Participants (deaneries, registrars)

Four UK deaneries (Northern Ireland, Mersey, KSS (Kent, Surrey and Sussex) and East Scotland) took part and a total of 66 registrars participated.

•  
•  
•  
•  
•  
•  
•

---

## Results (CARE)

### Deanery mean scores

The mean scores of registrars provided by each deanery were remarkably similar:

Deanery	Registrar Mean score
N Ireland	5.602
Mersey	5.671
KSS	5.624
E Scotland	5.673

The above results give confidence to the hypothesis that results generalise to the UK as a whole.

### Reliability (CARE)

**Overall Reliability:** This is represented by G as a number between 0 and 1 which represents the degree to which an assessment tool discriminates in the performance between individuals. A score of (0) would represent making a random judgement and (1) a decision free of all potential sources of error.

**Inter-rater Reliability:** This is represented by G as a number between 0 and 1 which represents the degree to which the patients rating doctors with CARE agree with each other on the degree of their satisfaction with an individual doctor's performance.

**Internal Consistency:** This is represented by G as a number between 0 and 1 which represents the degree to which each question in the CARE measure is related and how well each CARE question adds to the purpose of the overall questionnaire.

**Reliability Results (CARE)**

<b>G – patient satisfaction</b>		<b>Overall reliability</b>	<b>Inter-rater reliability</b>	<b>Internal consistency</b>
Questions	Number of Patients			
1	1	0.069	0.071	0.741
<b>10</b>	<b>20</b>	<b>0.665</b>	<b>0.604</b>	<b>0.966</b>
10	1	0.090	0.071	0.966
<b>10</b>	<b>30</b>	<b>0.748</b>	<b>0.696</b>	<b>0.966</b>

Unapproved

⋮

## Discussion (CARE: Validity and Reliability)

This study represents the first trial of CARE to discriminate performance in the context of general practice. CARE has been studied in secondary care (work in progress) and has demonstrated a high level of reliability to discriminate performance in the secondary care setting.

Testing “*Relationships with Patients*”, adds width to the sampling of qualities to be perceived to be tested (Appendix 1). Intuitively attractive as an area to test, the authors were concerned that the challenge to discriminate between such a similar group (age, experience) as GP registrars, would be beyond the capabilities of a “patient satisfaction questionnaire” or would demand analysis of an unfeasible number of patient responses. Ability to discriminate in performance between registrars was, however, proven by study and indicates that only 30 separate patient questionnaires are required for each registrar.

It is interesting to note that there was significant, expected, and reassuring correlation (Pearson 0.922, significant at, 0.01 level) between patient overall global rating of their experience and answers to mean individual question scores (Appendix 2).

It is also important to note that although patient satisfaction and video were expected by the *pre-pilot blueprinting* (appendix 1) to be both testing *Relationships with Patients*, this was not demonstrated, there being a very poor correlation between the two tools (Pearson 0.015).

Appendix 2 gives support to the expected need for multiple tools there being no area of significant correlation between tools due to the context specificity of individuals’ performance. It is interesting that a correlation (Pearson, 0.41 – 0.50) was found between *patient satisfaction* and *MSF* using non-clinical colleagues. Intuitively, this may have been predicted.

This pilot has shown that the involvement of patients by analysis of CARE competes psychometrically with other potentially robust assessments such as Objective Structured Clinical Examination (OSCE). Work to test the *predictive validity*<sup>5</sup> of CARE to predict and potentially give early warning of future concern in performance is an area for future study.

The application of CARE with other more heterogeneous groups in age, experience, and other factors such as GP principals holds the expected promise of even higher reliability<sup>5</sup> and offers potential for longitudinal study.

**Questionnaire and Results (CARE: Acceptability, Feasibility and Educational Impact)**

**Questionnaire**

Concern				No Concern			
GPR'S	1	2	3	4	5	6	7
	Strongly Agree		Disagree		Agree		Strongly Agree
Acceptability	n	n	n	n	n	n	n
Feasibility	n	n	n	n	n	n	n
Educational Impact	n	n	n	n	n	n	n

**Results**

CARE	GPR concern n	GPR total n	GPR concern %
<b>GPR'S</b>			
Acceptability	1.0	20.0	5.0
Feasibility	1.0	20.0	5.0
Educational Impact	3.0	20.0	15.0
<b>CARE</b>			
<b>TRAINERS</b>			
Acceptability	0.0	25.0	0.0
Feasibility	0.0	24.0	0.0
Educational Impact	3.0	24.0	12.5

.....

## Discussion (CARE: Acceptability, Feasibility and Educational Impact)

The above results suggest that CARE is perceived to be an acceptable and feasible measure. It will be difficult to truly estimate the actual educational impact of any of the proposed assessment tools until used in the applied context of assessment for nMRCGP.

The collection of 30 patient satisfaction questionnaires should not be onerous. The current GP contract recommends the collection of opinion from 50 patients. As the adoption of CARE requires no training or employment of assessors, this will limit the costs involved.

## Conclusions

### CARE: Validity and Reliability

CARE provided a valid and highly reliable measure of workplace-based performance suitable for a high stake assessment with an overall reliability co-efficient of 0.75 requiring only 30 patient satisfaction questionnaires. This result is equivalent to Multiple Choice examination (MCQ) and Objective Structured clinical Examination (OSCE) employed in other contexts.

### CARE: Acceptability, Feasibility and Educational Impact

Data collected suggest that the CARE patient satisfaction questionnaire offers an acceptable, feasible assessment for specialist GP training with generally positive feedback on perceived educational impact.

## References (CARE)

1. Mercer SW, Watt, GCM, Maxwell M, and Heaney DH. The development and preliminary validation of the Consultation and Relational Empathy (CARE) Measure: an empathy-based consultation process measure. *Family Practice* 2004, 21 (6), 699-705
2. Mercer SW. Using the CARE measure in secondary care. Report to the Centre for Change and Innovation, Scottish Executive Health Department (2005). [www.cci.nhs.scot.uk](http://www.cci.nhs.scot.uk)
3. Bikker AP, Mercer SW, Reilly D. A pilot prospective study on the consultation and relational empathy, patient enablement, and health changes over 12 months, in patients going to the Glasgow Homoeopathic Hospital. *J Alt. Comp. Med.* 2005, 11 (4), 591-600
4. McPherson H, Mercer SW, Scullion T, Thomas KJ. Empathy, enablement, and outcome: an exploratory study of acupuncture patients' perceptions. *J Alt. Comp. Med.* 2003, 9(6), 869-876.
5. Streiner DL, Norman GR. *Health Measurement Scales* (3<sup>rd</sup> ed.). Oxford Medical Publications, 1995.

Unapproved

# Multi-Source Feedback (MSF)

## Research Question

Does the MSF assessment tool developed for the WPBA pilot provide a valid, reliable, acceptable, and feasible measure of workplace-based performance to inform a high stakes decision on individuals' outcomes for general practice specialist training?

## Background and Validity

The MSF tool was developed by Douglas J Murphy and David Bruce, (NHS Education for Scotland), and Kevin W Eva (University of McMaster, Ontario, Canada). The instrument built on previously work in the hospital contexts in both the UK and the United States<sup>1,2</sup> and aimed to investigate the application of MSF in the GP context. The tool was developed as a web-based application<sup>3</sup>.

Case variance (number of applications of MSF) has been established as more important psychometrically than the number of questions asked<sup>2</sup>. It was decided to limit the number of questions to two global enquiries of colleagues' opinions. Clinical colleagues rated professional and clinical qualities and non-clinical colleagues rated professional qualities only.

There is no evidence that the limitation of the number of questions asked by MSF is detrimental to the provision of feedback. One might hypothesise that the limitation of length of the questionnaire would encourage participation in free text comments and aid the provision of specific examples of alleged behaviour.

It is known that for the purposes of educational feedback, the provision of scores, free text comments and appraisal by a colleague to aid insight are of benefit<sup>4</sup>.

In the blueprinting exercise designed to examine the perceived validity for the purpose of the external tools WPBA pilot, MSF was perceived by

stakeholders to test a wide range of qualities described by *Good Medical Practice*<sup>5</sup> (Appendix 1).

### **Participants (deaneries, registrars)**

Four UK deaneries (North and North East of Scotland, Northern Ireland, Wales, and West Midlands) took part and a total of 46 registrars participated.

Unapproved

.....

---

## Results (MSF)

### Deanery mean scores

The mean scores of registrars provided by each deanery were remarkably similar:

Clinical		Means			
N Scotland	5.421	Time 1	5.585	Question 1	5.763
N East Scotland	5.367	Time 2	5.589	Question 2	5.411
N Ireland	5.85				
Wales	5.683				
W Midland	5.488				

Non-Clinical		Means			
N Scotland	5.867	Time 1	5.886	Question 1	5.947
N East Scotland	5.76	Time 2	6.01		
N Ireland	6.108				
Wales	5.9				
W Midland	5.933				

The above results give support to the hypothesis that results generalise to the UK as a whole.

### Reliability Results (MSF)

**Overall Reliability:** This is represented by G as a number between 0 and 1 which represents the degree to which an assessment tool discriminates in the performance between individuals. A score of (0) would represent making a random judgement and (1) a decision free of all potential sources of error.

**Inter-rater Reliability:** This is represented by G as a number between 0 and 1 which represents the degree to which the colleagues rating doctors with MSF agree with each other on the degree of their satisfaction with an individual doctor's performance.

**Internal Consistency:** This is represented by G as a number between 0 and 1 which represents the degree to which each question in the MSF measure

(clinical only) is related and how well each MSF question adds to the purpose of the overall questionnaire.

**Test-retest Reliability:** This is represented by G as a number between 0 and 1 which represents the degree to which the score given to individuals by colleagues is stable when taken at different times.

**Reliability Results (MSF)**

<b>G - Clinical</b>			<b>Overall</b>	<b>Inter-rater</b>	<b>Internal consistency</b>	<b>Test-Retest</b>
Questions	Raters	Times				
1	1	1	0.257	0.300	0.647	0.290
<b>2</b>	<b>5</b>	<b>2</b>	<b>0.778</b>	<b>0.682</b>	<b>0.785</b>	<b>0.573</b>
2	5	1	0.644	0.682	0.785	0.402
2	5	3	0.836	0.682	0.785	0.668
1	3	2	0.641	0.563	0.647	0.548

<b>G – non clinical</b>		<b>Overall</b>	<b>Inter-rater</b>	<b>Test-Retest</b>
Times	Raters			
1	1	0.311	0.311	0.335
<b>2</b>	<b>5</b>	<b>0.814</b>	<b>0.693</b>	<b>0.501</b>
3	5	0.854	0.693	0.601
2	2	0.724	0.575	0.501

⋮

---

## Discussion (MSF: Validity and Reliability)

Multi-Source Feedback (360 degree appraisal) was developed in secondary care in the USA by Ramsay <sup>6</sup>. It has subsequently been developed by a number of researchers including surgical trainees in the United States and in the United Kingdom by researchers including West Midlands <sup>2,3</sup>.

In the pre-pilot blueprinting exercise (appendix 1), MSF was valued highly by educationalists, GP trainers and GP registrars in expected ability to assess a wide range of qualities.

A recurring difficulty, however, in the application of MSF has been to recruit enough different opinions to form a reliable discriminatory judgement on individuals' performance. Recent work points to a benefit in increasing the number of applications as it requires fewer opinions. This, if correct when tested in the UK general practice context, may aid implementation in smaller practices.

For the first time, in primary care in the UK, we were involving non clinical assessor colleagues in the assessment of doctors. This meant that the whole Primary Care Team could be involved in making an evaluation of a doctor's "professionalism".

Five opinions on two occasions by either clinical or non-clinical colleagues nominated by the candidate provide a highly reliable judgement appropriate for a summative assessment (0.78, 0.81).<sup>7</sup>

There was a significant correlation in scores between clinical and non-clinical colleague assessors and between *professionalism* and *clinical* question by the clinical colleague assessors. (appendix 2)

Appendix 2 gives support to the expected need for multiple tools, there being no area of significant correlation between tools due to the context specificity of individuals' performance.

This pilot has shown that administration of this web based MSF competes psychometrically with other potentially robust assessments such as Objective Structured Clinical Examination (OSCE). Work to test the *predictive validity*<sup>7</sup>

of MSF to predict and potentially give early warning of future concern in performance is an area for future study.

The application of this MSF with other more heterogeneous groups in age, experience, and other factors such as GP principals holds the expected promise of even higher reliability and offers potential for longitudinal study<sup>7</sup>.

Unapproved

Questionnaire and Results (MSF: Acceptability, Feasibility and Educational Impact)

Questionnaire

GPR'S	CONCERN			NO CONCERN			
	1	2	3	4	5	6	7
	Strongly Disagree		Disagree		Agree		Strongly Agree
Acceptability	n	n	n	n	n	n	n
Feasibility	n	n	n	n	n	n	n
Educational Impact	n	n	n	n	n	n	n

Results

MSF GPR'S	GPR concern n	GPR total n	GPR concern %
Acceptability	4.0	28.0	14.3
Feasibility	4.0	28.0	14.3
Educational Impact	5.0	28.0	21.4
MSF TRAINERS	GPR concern n	GPR total n	GPR concern%
Acceptability	1.0	24.0	4.2
Feasibility	1.0	24.0	4.2
Educational Impact	5.0	25.0	20.0

## Discussion (MSF: Acceptability, Feasibility and Educational Impact)

The above results suggest that MSF is perceived to be an acceptable and feasible measure. It will be difficult to truly estimate the actual educational impact of any of the proposed assessment tools until used in the applied context of assessment for nMRCGP. The above survey indicates a generally positive perception.

The administration of MSF on two occasions in both hospital and general practice contexts via a web based system should not be onerous. As the adoption of MSF requires no training or employment of assessors this will limit the costs involved.

## Conclusions

### MSF: Validity and Reliability

MSF provided a highly reliable and feasible measure of workplace-based performance suitable for a high stakes assessment with an overall reliability co-efficient of 0.78 (five clinical colleagues) and 0.81 (five non-clinical colleagues) with the MSF required to be conducted twice only. This result is psychometrically similar to Multi-Choice examination (MCQ) and Objective Structured Clinical Examination (OSCE) employed in other contexts.

### MSF: Acceptability, Feasibility and Educational Impact

Data collected suggests this web based Multi-Source Feedback offers an acceptable, feasible assessment for specialist GP training with a generally positive feedback on perceived educational impact.

•  
•  
•  
•  
•  
•  
•

---

## References (MSF)

1. Multi-Source Feedback: 360° Team Assessment of Behaviour (TAB) West Midlands Deanery, UK <http://www.wmdeanery.org/Downloads/360download.asp>
2. Reed G Williams, Steven Verhulst, Jerry A Colliver, and Gary L Dunnington. Assessing the reliability of resident performance appraisals: More items or more observations? *Surgery* 2005; 137: 141-147.
3. [wbapilot@chs.dundee.ac.uk](mailto:wbapilot@chs.dundee.ac.uk) available [www.dundee.ac.uk/gptraining](http://www.dundee.ac.uk/gptraining)
4. Sargeant, Mann & Ferrier, *Medical Education* 2005;39: 497-504.  
Making available a mentoring service to support physician feedback, reflection, learning and change, can increase acceptance and use of feedback.
5. General Medical Council. *Maintaining Good Medical Practice*, General Medical Council, London, 1998.
6. Ramsey P, Wenrich M. Peer ratings: an assessment tool whose time has come. *Journal of General Internal medicine* 1999; 14: 581-2.
7. Streiner DL, Norman GR. *Health Measurement Scales* (3<sup>rd</sup> ed.). Oxford Medical Publications, 1995.

# Video

## Research Question

Does the video assessment tool developed for the WPBA pilot based on merged content of current Committee of General Practice Education Directors (COGPED) Summative Assessment and MRCGP video marking schedules provide a valid, reliable, acceptable, and feasible measure of workplace-based performance to inform a high stakes decision on individuals' outcomes for general practice specialist training?<sup>1,2</sup>

## Background and Validity

The video tool in this pilot was developed to build on previous work undertaken by COGPED and the RCGP. An exercise was undertaken to agree merger of the two schedules by statistical agreement utilising the opinion of COGPED assessors, MRCGP assessors, and GP trainers. A pre-pilot was conducted to test the correlation of proposed items and to finalise the marking schedule. Agreed items of assessment were rated on a continuous rating scale (1-7), which is known to aid the expected reliability<sup>3</sup>

In the blueprinting exercise designed to examine the perceived validity for the purpose of the external tools WPBA pilot, video was perceived by stakeholders to test *Good Clinical Care* and *Relationships with Patients*, qualities described by *Good Medical Practice*. In addition, video was perceived to test *Professionalism and Critical Thinking*. (Appendix 1).

## Participants (deaneries, registrars)

Three UK deaneries (East of Scotland, Northern Ireland, and North and North East of Scotland) took part and a total of 30 registrars participated. Each registrar submitted four consultations assessed by four assessors in each deanery.



### Reliability Results (Video)

G - Video			Overall	Inter-rater	Internal consistency	Inter-case
Questions	Raters	Number of Consultations (case)				
1	1	1	0.024	0.209	0.092	0.202
<b>10</b>	<b>4</b>	<b>4</b>	<b>0.393</b>	<b>0.514</b>	<b>0.503</b>	<b>0.503</b>
25	6	6	0.566	0.614	0.717	0.603
10	4	6	0.426	0.514	0.503	0.603
10	4	7	0.437	0.514	0.503	0.640

Unapproved

•  
•  
•  
•  
•  
•  
•

---

## Discussion (Video: Validity and Reliability)

Video was expected by educationalists, GP trainers and GP registrars to assess a range of qualities by the pre-pilot blueprinting exercise (*Good Clinical Care, Relationship with Patients, Professionalism and Critical Thinking*)<sup>4</sup>. It was not expected to test any quality not already sampled by MCQ, OSCE, PSQ, and MSF.

Video is currently used by the RCGP as one part of the current MRCGP examination. In 1999 research indicated that 28 observations (four assessors examining the same seven consultations) provided robust agreement on performance (*kappa*, 0.79) using two point dichotomous, (present/absent), criteria<sup>5</sup>. The overall reliability (discrimination of candidates) was not published.

A pre-pilot study of the seven point video assessment employing two RCGP and two COGPED assessors (with no prior calibration) had been encouraging requiring only 12 observations for overall reliability (G, 0.80), with good internal consistency (0.92, 13 items). Ram P et al previously successfully demonstrated reliability with 16 observations using a seven point rating schedule (G, 0.81).<sup>6</sup>

Given this pre-pilot success, it was disappointing that the results of this pilot indicated that the video schedule was not fit for a high stakes assessment. Four assessors looking at four consultations (16 observations) had poor reliability (G overall, 0.39). Inter-rater agreement was only fair (G, 0.51)

It is also important to note that although patient satisfaction and video were expected by the *pre-pilot blueprinting* to both be testing *Relationships with patients*, this was not demonstrated, there being a very poor correlation between the two tools (Pearson 0.015)<sup>4</sup>.

Appendix 1 gives support to the expected need for multiple tools, there being no area of significant correlation between tools due to the context specificity of individuals' performance.

The above results support the observation of consultation by video as an appropriate means of providing teaching and training in GP education, but not as a valid or reliable method to inform individuals' licensure.

## Questionnaire and Results (Video: Acceptability, Feasibility and Educational Impact)

### Questionnaire

Concern				No Concern			
GPR'S	1	2	3	4	5	6	7
	Strongly Disagree		Disagree		Agree		Strongly Agree
Acceptability	n	n	n	n	n	n	n
Feasibility	n	n	n	n	n	n	n
Educational Impact	n	n	n	n	n	n	n

### Results

<b>Video</b>	<b>GPR concern n</b>	<b>GPR total n</b>	<b>GPR concern %</b>
<b>GPR'S</b>			
<b>Acceptability</b>	0	11	0.0
<b>Feasibility</b>	1	11	9.1
<b>Educational Impact</b>	1	11	9.1
<b>Video</b>	<b>Trainers concern n</b>	<b>Trainers total n</b>	<b>Trainers concern %</b>
<b>Trainers</b>			
<b>Acceptability</b>	1	8	12.5
<b>Feasibility</b>	0	8	0.0
<b>Educational Impact</b>	1	8	12.5
<b>Video</b>	<b>Assessors concern n</b>	<b>Assessors total n</b>	<b>Assessors concern %</b>
<b>Assessors</b>			
<b>Acceptability</b>	2	7	28.6
<b>Feasibility</b>	4	7	57.0
<b>Educational Impact</b>	2	7	28.6



## References (video)

1. National Office for Summative Assessment. First level assessor's instructions and marking schedule. <http://www.nosa.org.uk>
2. RCGP: Video assessment in consulting skills, 2005.
3. Streiner DL, Norman GR. Health Measurement Scales (3<sup>rd</sup> ed.). Oxford Medical Publications, 1995.
4. General Medical Council. Maintaining Good Medical Practice, General Medical Council, London, 1998.
5. Assessing physician's interpersonal skills via videotaped encounters: a new approach for the RCGP's membership examination. Journal of Health Communication 1999 vol 4(2): 143-152
6. Ram P, Grol R, Rethans J J, Schouten B, van der Vleuten C, Kester A. Assessment of general practitioners by video observation of communicative and medical performance in daily practice: issues of validity, reliability, and feasibility. Medical Education 1999; 33(6): 447-454.

Unapproved

# Referrals

## Research Question

Does the referrals assessment tool developed for the WPBA provide a valid, reliable, acceptable and feasible measure of workplace-based performance to inform a high stakes decision on individuals' outcomes for general practice specialist training?

## Background and Validity

The referrals tool in this pilot was developed specifically for the purpose of the WPBA pilot. The content of the assessment tool was agreed by implementation of a *content validity index* exercise. Agreed items of assessment were rated on a continuous rating scale (1-7) to inherently aid the expected reliability<sup>1</sup>.

In the blueprinting exercise designed to examine the perceived validity for the purpose of the external tool WPBA pilot, referrals was not tested. Originally case-based discussion was proposed but was rejected for final inclusion due to expected reliability difficulties and replaced by assessment of referral letters.

## Participants (deaneries, registrars, assessors)

Four UK deaneries (West Midlands, Wales, East of Scotland and South-East of Scotland) took part and a total of 72 registrars participated. Each registrar submitted 10 referral letters assessed by four assessors in each deanery using five questions. A sixth question (question 1) was excluded due to failure to endorse sufficiently by assessors for 20 registrars (n=72).

## Results (Referrals)

### Deanery mean scores (registrars) (1-7 scale)

Means Deanery		Means Question	
West Midlands	4.579	Q2	5.332
Wales	4.638	Q3	5.097
SE Scotland	4.768	Q4	4.784
E Scotland	4.741	Q5	4.528
		Q6	3.774

The above results give confidence to the hypothesis that results generalise to the UK as a whole.

## Reliability (Referrals)

**Overall Reliability:** This is represented by G as a number between 0 and 1 which represents the degree to which an assessment tool discriminates in the performance between individuals. A score of (0) would represent making a random judgement and (1) a decision free of all potential sources of error.

**Inter-rater Reliability:** This is represented by G as a number between 0 and 1 which represents the degree to which the assessors rating doctors letters of referral agree with each other on the degree of their satisfaction with an individual doctor's performance.

**Internal Consistency:** This is represented by G as a number between 0 and 1 which represents the degree to which each question in the referral measure is related and how well each Referral question adds to the purpose of the overall assessment schedule.

**Inter-letter Reliability:** This is represented by G as a number between 0 and 1 which represents the degree to which the score given to individuals is stable when assessed between different letters.

## Reliability Results (Referrals)

.....

---

G - Referrals			G (overall)	G (inter-rater)	G(internal consistency)	G(inter-letter)
Questions	Raters	Number of Letters				
1	1	1	0.016	0.219	0.216	0.078
<b>5</b>	<b>4</b>	<b>10</b>	<b>0.311</b>	<b>0.529</b>	<b>0.579</b>	<b>0.458</b>
10	6	20	0.475	0.628	0.734	0.628
5	4	20	0.392	0.529	0.579	0.628
5	4	15	0.361	0.529	0.579	0.559

Unapproved

## Discussion (Referrals: Validity and Reliability)

The assessment of referral letters (Referrals) was not assessed as part of the blueprinting of piloted assessment tools to desired qualities. Referrals replaced “case-based discussion”, a tool which did not lend itself to the demonstration of reliability.

The pilot results were disappointing and indicated that the piloted Referrals schedule was not fit for a high stakes assessment. Four assessors looking at 20 letters had poor but the same reliability as video reliability (G overall, 0.39). Inter-rater agreement was only fair and similar to video (G, 0.53). This was impressive as this was a new tool with no attempt to calibrate between participating deaneries.

Low levels of internal consistency (G, 0.58, 5 questions) and inter-letter reliability (G, 0.63, 20 letters) suggest that the assessment schedule should be lengthened and/ or number of letters increased. This may affect the feasibility of the assessment.

Appendix 1 gives support to the expected need for multiple tools, there being no area of significant correlation between tools due to the context specificity of individuals’ performance.

The above results support the evaluation of referral letters for the purpose of providing teaching and training in GP education, but not as a valid or reliable method to inform individuals’ licensure.

.....

**Questionnaire and Results (Referrals: Acceptability, Feasibility and Educational Impact)**

**Questionnaire**

Concern				No Concern			
GPR'S	1	2	3	4	5	6	7
	Strongly Disagree		Disagree		Agree		Strongly Agree
Acceptability	n	n	n	n	n	n	n
Feasibility	n	n	n	n	n	n	n
Educational Impact	n	n	n	n	n	n	n

**Results**

Referrals	GPR concern n	GPR total n	GPR concern %
<b>GPR'S</b>			
Acceptability	2.0	21.0	9.5
Feasibility	2.0	21.0	9.5
Educational Impact	6.0	21.0	28.6
<b>Referrals</b>			
<b>TRAINERS</b>			
Acceptability	1.0	34.0	2.9
Feasibility	1.0	34.0	2.9
Educational Impact	6.0	34.0	17.6
<b>Referrals</b>			
<b>ASSESSORS</b>			
Acceptability	0.0	13.0	0.0
Feasibility	2.0	13.0	15.4
Educational Impact	0.0	13.0	0.0

## Discussion (Referrals: Acceptability, Feasibility and Educational Impact)

Assessors had concern regarding feasibility (15.4%), and trainers and registrars were concerned regarding educational impact (17.6%, 28.6%).

Future work will be needed to research and develop the reliability (number of observations needed to discriminate performance) if referral is to demonstrate psychometric qualities equivalent to alternative assessments (MCQ, OSCE, CARE, piloted MSF).

## Conclusions

### Referrals: Validity and Reliability

The assessment of referral letters did not provide a reliable enough measure to discriminate between different registrars' performance for a high stakes assessment.

The above results do illustrate that assessment of 10 referral letters by four assessors with the piloted schedule ( $C = 0.31$ ) may be considered as giving data suitable for the purposes of feedback, but not for the purpose of drawing inferences on suitability for licensure. This format of formative feedback to their registrars is currently common within GP trainers' groups.

### Referrals: Acceptability, Feasibility and Educational Impact

There was no major level of concern with regard to the acceptability, feasibility, or educational impact of the piloted Referrals assessment tool. It appeared to be well supported and offers a useful addition for the purposes of teaching and training.

•  
•  
•  
•  
•  
•  
•

## References Referrals

1. Streiner DL, Norman GR. Health Measurement Scales (3<sup>rd</sup> ed.). Oxford Medical Publications, 1995.
2. Crossley JGM, Howe A, Newble D, Jolly B and Davies HA. (2001) Sheffield Assessment Instrument for Letters (SAIL): performance assessment using outpatient letters. Medical Education, 35:1115-1124
3. Scottish Intercollegiate Guideline Network (SIGN). Report on a Recommended Referral Document. SIGN, Edinburgh 1998.
4. New Manchester Rating Scale. Available from:  
<http://www.gp-training.net/training/nmrs/nmrs.htm>
5. Referral Advice. A guide to appropriate referral from general to specialist services. National Institute for Clinical Evidence (NICE) London 2001

Unapproved

# Criterion Audit

## Research Question

Does the criterion audit assessment tool developed for the WPBA provide a valid, reliable, acceptable, and feasible measure of workplace-based performance to inform a high stakes decision on individuals' outcomes for general practice specialist training?

## Background and Validity

The criterion audit assessment tool in the pilot was amended from the criterion audit assessment tool currently used for Summative Assessment for GP registrars in the UK.<sup>1</sup> In application in the WPBA pilot, the existing eight item assessment tool was extended to nine items to include all descriptors given in the original schedule. The nine items were then rated by four assessors in each deanery on a continuous rating scale (1-7), which is known to aid the expected reliability.<sup>2</sup> Analysis of data was subsequently restricted to two assessors due to local deanery limitations.

In the blueprinting exercise designed to examine the perceived validity for the purpose of the external tools WPBA pilot, criterion audit was expected to test qualities of *Critical Thinking* (Appendix 1).

## Participants (deaneries, registrars, assessors)

Five UK deaneries (Mersey, KSS, Northern Ireland, West Midlands and Wales) took part and a total of 70 registrars participated. Each registrar submitted one audit project for Summative Assessment which was marked by two assessors with the existing and piloted schedules concurrently.

•  
•  
•  
•  
•  
•  
•  
•  
•  
•  
•

## Results (Criterion Audit)

### Deanery mean scores (registrars) (1-7 scale)

#### Means

Mersey	4.753
KSS	4.429
N Ireland	4.036
W Midlands	4.261
Wales	4.219

The above results suggest that mean scores vary to an extent across deaneries. This points to differences in performance, but it is not clear whether that is due to the projects or different standards applied by assessors in different deaneries.

## Reliability (Criterion Audit)

**Overall Reliability:** This is represented by G as a number between 0 and 1 which represents the degree to which an assessment tool discriminates in the performance between individuals. A score of (0) would represent making a random judgement and (1) a decision free of all potential sources of error.

**Inter-rater Reliability:** This is represented by G as a number between 0 and 1 which represents the degree to which the assessors rating doctors' audit projects agree with each other on the degree of their satisfaction with an individual doctor's performance.

**Internal Consistency:** This is represented by G as a number between 0 and 1 which represents the degree to which each question in the criterion audit measure is related and how well each criterion audit assessment item adds to the purpose of the overall assessment schedule.

## Reliability Results (Criterion Audit)

G – Criterion audit		Number of Raters	Overall reliability	Inter-rater reliability	Internal consistency
Number of questions	1	1	0.143	0.216	0.837
	<b>9</b>	<b>2</b>	<b>0.480</b>	<b>0.355</b>	<b>0.844</b>
	<b>9</b>	<b>4</b>	<b>0.637</b>	<b>0.524</b>	<b>0.844</b>

### Discussion (Criterion audit: Validity and Reliability)

The above pilot results were disappointing and indicated that the piloted Criterion Audit schedule was not fit for a high stakes assessment. Better results had been anticipated. The use of a continuous rating scale applied to virtually unchanged questions (one added item) would improve the reliability compared to the existing schedule<sup>2</sup>.

The employment of two assessors had poor overall reliability (G overall, 0.48). Inter-rater agreement was very poor (G, 0.36). This was disappointing as national calibration exercises exist for the purpose of current summative assessment.

Appendix 1 gives support to the expected need for multiple tools there being no area of significant correlation between tools due to the context specificity of individuals' performance.

The above results support the evaluation of criterion audit for the purpose of providing teaching and training in GP education, but not as a valid or reliable method to inform individuals' licensure.

Questionnaire and Results (Criterion Audit: Acceptability, Feasibility and Educational Impact)

Questionnaire

Concern				No Concern			
GPR'S	1	2	3	4	5	6	7
	Strongly Disagree		Disagree		Agree		Strongly Agree
Acceptability	n	n	n	n	n	n	n
Feasibility	n	n	n	n	n	n	n
Educational Impact	n	n	n	n	n	n	n

Results

Criterion Audit	GPR concern n	GPR total n	GPR concern %
<b>GPR'S</b>			
Acceptability	1.0	14.0	7.1
Feasibility	1.0	14.0	7.1
Educational Impact	3.0	15.0	20.0
<b>Criterion Audit</b>			
<b>TRAINERS</b>	Trainers concern n	Trainers total n	Trainers concern%
Acceptability	1.0	12.0	8.3
Feasibility	1.0	12.0	8.3
Educational Impact	4.0	12.0	33.3
<b>Criterion Audit</b>			
<b>ASSESSORS</b>	Assessors concern n	Assessors total n	Assessors concern %
Acceptability	1.0	12.0	8.3
Feasibility	1.0	12.0	8.3
Educational Impact	1.0	12.0	8.3

## Discussion (Criterion Audit: Acceptability, Feasibility and Educational Impact)

GP registrars and their trainers had concerns (20%, 33.3%) regarding the educational impact of this established assessment tool. It is difficult to compare these figures with the other piloted tools' "perceptions of educational impact" with the exception of video, the other piloted current form of assessment.

## Conclusions

### Criterion Audit: Validity and Reliability

The assessment of criterion audit projects did not provide a reliable enough measure to discriminate between different registrars' performance for a high stakes assessment.

The above results do illustrate that assessment of criterion audit projects by two assessors with the piloted schedule ( $\kappa = 0.48$ ) may be considered as giving data suitable for the purposes of feedback but not for the purpose of drawing inferences on suitability for licensure. This format of formative feedback to their registrars is currently common within GP trainers' groups.

### Criterion Audit: Acceptability, Feasibility and Educational Impact

This assessment tool is part of current GP registrar Summative Assessment. Trainers, and to a lesser extent their registrars, expressed concern regarding the tool's educational impact.

.....

\_\_\_\_\_

## References Criterion Audit

1. Lough JR, Murray TS. Audit and summative assessment: a completed audit cycle. *Medical Education* 2001; 35(4): 357-63.
2. Streiner DL, Norman GR. *Health Measurement Scales* (3<sup>rd</sup> ed.). Oxford Medical Publications, 1995.

Unapproved

# Significant Event Analysis (SEA)

## Research Question

Does the SEA tool developed for the WPBA provide a valid, reliable, acceptable and feasible measure of workplace-based performance to inform a high stakes decision on individuals' outcomes for general practice specialist training?

## Background and Validity

The SEA tool piloted was amended from an existing SEA assessment tool used for provision of feedback on significant event analysis in the West of Scotland<sup>1</sup>. The existing assessment tool items were put on a continuous rating scale (1-7) and then assessed by four assessors in each deanery, as this is known to improve reliability<sup>2</sup>. Deanery assessors received a half-day's training. The marking schedule was made explicit to the registrars submitting projects.

In the blueprinting exercise designed to examine the perceived validity for the purpose of the external tools WPBA pilot, SEA was expected to test qualities of *Critical Thinking* (Appendix 1).

## Participants (deaneries, registrars, assessors)

**Four UK deaneries (Mersey, KSS, East Scotland and North/East Scotland) participated and included a total of 36 participating registrars.** Each registrar submitted two SEA projects, which were marked by 16 assessors.

.....

---

## Results (SEA)

### Deanery mean scores (registrars) (1-7 scale)

Deanery	Means(significant event analysis)		
N Scotland	4.051	Q1	4.269
Mersey	3.632	Q2	3.94
KSS	3.972	Q3	4.06
SE Scotland	3.905	Q4	4.00
		Q5	3.523
		Q6	4.042
		Q7	4.037
		Q8	3.907
		Q9	3.486

The above results give confidence to the hypothesis that results generalise to the UK as a whole.

### Reliability (SEA)

**Overall Reliability:** This is represented by G as a number between 0 and 1 which represents the degree to which an assessment tool discriminates in the performance between individuals. A score of (0) would represent making a random judgement and (1) a decision free of all potential sources of error.

**Inter-rater Reliability:** This is represented by G as a number between 0 and 1 which represents the degree to which the assessors rating doctors' significant event projects agree with each other on the degree of their satisfaction with an individual doctor's performance.

**Internal Consistency:** This is represented by G as a number between 0 and 1 which represents the degree to which each question in the significant event measure is related and how well each significant event assessment item adds to the purpose of the overall assessment schedule.

**Inter-case Reliability:** This is represented by G as a number between 0 and 1 which represents the degree to which SEA report(s) capture the variance in quality from different reports by the same person. As number of reports submitted increases so does the inter-case reliability co-efficient. Thus more potential variation in standard is explained and the inferences made on an individual's ability to write a report are made more trustworthy.

**Reliability Results (SEA)**

**G – Significant event analysis**

Number of questions	Number of raters	Number of projects (cases)	Overall reliability	Inter-rater reliability	Internal consistency	Inter-case reliability
1	1	1	0.045	0.248	0.396	0.797
<b>9</b>	<b>4</b>	<b>2</b>	<b>0.247</b>	<b>0.569</b>	<b>0.855</b>	<b>0.887</b>
9	4	4	0.303	0.569	0.855	0.940
9	2	2	0.178	0.398	0.855	0.887
9	2	4	0.205	0.398	0.855	0.940
<b>9</b>	<b>2</b>	<b>1</b>	<b>0.141</b>	<b>0.398</b>	<b>0.855</b>	<b>0.797</b>

Unapproved

⋮

---

## Discussion (SEA: Validity and Reliability)

Significant event analysis (SEA) is required as part of the new General Medical Services contract for general practice. SEA involves having the physician submit a written analysis of a near-miss, adverse or beneficial event (e.g. a patient complaint or commendation), including clarification of the reason for the event, insight that has been gained, and how the risk of further adverse events will be minimised.

However, the above pilot reliability results were disappointing and indicated that the piloted significant event assessment schedule was not fit for a high stakes assessment.

The employment of two assessors had very poor overall reliability (G overall, 0.14) rising to only 0.25 when two project reports were marked by four assessors.

Inter-rater agreement was encouraging and reached previously found levels of (G, 0.57). This was impressive as only one afternoon training calibration exercise had been held.

Inter-case reliability had not previously been analysed with this instrument. The high inter-case reliability (G, 0.8) with assessment of only one project and (G, 0.89) with two projects indicates that evaluation of two projects would suffice to capture variation in reports on SEA written by an individual.

Better results for overall reliability had been anticipated. Past experience with assessment of projects submitted previously by GP principals and GP registrars analysed in the West of Scotland had better and encouraging results. The poor reliability on this occasion may be explained by the homogenous nature of the assessed group (GP registrars), prior knowledge of the marking schedule, and perceived stakes of the assessment improving and limiting the range of captured performance. Further analyses of descriptive statistics and correlations of scores with other tools are to be conducted.

SEA as a practice activity is extremely useful to facilitate insight and change. As an individual assessment instrument, it tests the completion of a written report to reflect teamwork involved. As such, significant event analysis is an important team technique and should form part of teaching and training for a career in general practice but, in piloted form, not as a valid or reliable method to inform individuals' licensure.

## Questionnaire and Results (SEA: Acceptability, Feasibility and Educational Impact)

### Questionnaire

Concern				No Concern			
GPR'S	1	2	3	4	5	6	7
	Strongly Disagree		Disagree		Agree		Strongly Agree
Acceptability	n	n	n	n	n	n	n
Feasibility	n	n	n	n	n	n	n
Educational Impact	n	n	n	n	n	n	n

### Results

<b>SEA</b>	<b>GPR</b>	<b>GPR</b>	<b>GPR</b>
<b>GPR'S</b>	concern n	total n	concern %
Acceptability	1.0	24.0	4.2
Feasibility	0.0	25.0	0.0
Educational Impact	4.0	24.0	16.7
<b>SEA</b>	<b>Trainers</b>	<b>Trainers</b>	<b>Trainers</b>
<b>TRAINERS</b>	concern n	total n	concern%
Acceptability	1.0	32.0	3.1
Feasibility	1.0	34.0	2.9
Educational Impact	3.0	34.0	8.8
<b>SEA</b>	<b>Assessors</b>	<b>Assessors</b>	<b>Assessors</b>
<b>ASSESSORS</b>	concern n	total n	concern %
Acceptability	0.0	6.0	0.0
Feasibility	0.0	6.0	0.0
Educational Impact	0.0	6.0	0.0



## References Significant Event Analysis

1. Bowie P, McKay J, Lough M. Peer assessment of significant event analyses: being a trainer confers an advantage. *Education for Primary Care* 2003; 14(3): 338-344
2. Streiner DL, Norman GR. *Health Measurement Scales* (3<sup>rd</sup> ed.). Oxford Medical Publications, 1995.

Unapproved

.....

---

# Results Summary

The intent of any system of workplace-based performance testing is to provide physicians with an early warning of difficulty, enabling them to fulfill their professional responsibility while also providing reassurance and protection to the public. The findings in this report provide workplace-based tools CARE (Patient Satisfaction Questionnaire) and Multi-Source Feedback (MSF) as a basis on which to guide future planning and research for the RCGP in the UK and other agencies in the UK and abroad in the development of credible systems of workplace based assessment for licensure.

Unapproved

# Conclusions

This pilot study was designed to test the utility of six workplace based assessment tools. Following review of the literature it is our belief that this is the single largest study that has been undertaken to test multiple assessment tools in a scientifically rigorous manner.

General practice has a strong history of high quality assessment methods. The aim of this pilot was to build on previous work in both Summative Assessment and MRCGP assessment methodology. The introduction of new GPST programmes, based in general practice and underpinned by the RCGP curriculum, creates an opportunity to embed assessment into feedback, appraisal and learning.

Achieving reproducible assessments that allow inferences to be made about a GPRs performance in the workplace has been previously described as the “holy grail” of assessment. The results from this pilot demonstrate that the Multi-source Feedback tool and Patient Satisfaction Questionnaire can deliver the above. The overall reliability coefficient of both tools meet the standards for a high stakes assessment – in this case as part of the assessment for a CCT and the award of an MRCGP.

The blueprinting exercise demonstrates that using the MSF and PSQ tool in addition to the AKT and CSA, will cover the qualities required of a general practitioner outlined in Good Medical Practice, which has been mapped to the new RCGP curriculum.

The Video, SEA, Criterion Audit and Analysis of Referrals tools had lower overall reliability and therefore are unable to differentiate between GPRs in practice and to demonstrate the spectrum of performance between those excelling, those with average performance and those with performance difficulties. Psychometric analysis however shows that each tool is well designed with good or moderate/high internal consistency, and that with appropriate calibration or training is capable of achieving moderate/good levels of inter-rater reliability. Feedback from GPRs, trainers and assessors indicated that these tools were valuable and useful in the training environment.

It was recognised in piloting the six externally assessed tools, that maximum educational benefit required the expertise of the GP trainer using their feedback and appraisal skills. This is particularly the case in the MSF and PSQ tools where the skills of the GP trainer are required to ensure that insight and learning occur. The process of integration of assessment and learning is made explicit in the educational model (Appendix 3).

⋮

In addition, to determining the utility of the workplace based assessment tools, the national steering group was also concerned with the overall burden of assessment involved in WPBA. The group therefore produced a framework for assessment that took into account both utility and overall burden of assessment (Appendix 4). It is our view that the framework proposed is both practicable and acceptable to GPRs and GP trainers.

Using this framework will provide a measure of acquisition of the new RCGP curriculum competencies by GPRs that is both fair and defensible.

This pilot has tested tools with a GPR population, and while the results cannot be transferred to other groups, it seems likely that if tested in a less homogeneous population than the GPR cohort, that the results would be at least as impressive. This pilot work may therefore point to future directions for national and international re-licensure and re-accreditation of established doctors.

**Douglas J. Murphy**  
**David A. Bruce**  
**Kevin W. Eva**

**25 August 2006**

Unapproved

# Appendices

Appendix 1	Blueprinting Exercise
Appendix 2	Correlations
Appendix 3	Table 1 - Educational Model
Appendix 4	Table 2 - Proposed Assessment Framework
Appendix 5	Membership of the Steering Group

Unapproved

## Information

*Downloads of all tools/marking schedules are available from [www.dundee.ac.uk/gptraining](http://www.dundee.ac.uk/gptraining)*

Unapproved

APPENDIX 1

**Table 1: Mean effectiveness ratings assigned to each measurement instrument as a function of desired quality of physicians**

Measurement Instrument	Physician Quality								Mean (95% CI)
	Good Clinical Care	Maintaining Good Medical Practice	Relationship with Patients	Work with Colleagues	Probity	Health	Professionalism	Critical Thinking	
MCQ	<b>5.07</b>	5.05	2.01	1.97	2.37	1.62	3.03	4.43	3.20 (3.00–3.40)
OSCE	<b>5.55</b>	5.07	4.25	3.15	3.03	2.22	4.43	<b>4.95</b>	4.08 (3.88–4.28)
Video	<b>5.23</b>	4.45	<b>5.83</b>	2.78	2.90	2.52	<b>5.00</b>	<b>4.78</b>	4.19 (4.00–4.40)
SEA	4.75	4.92	3.67	4.45	3.83	2.63	4.48	<b>5.05</b>	4.22 (4.03–4.42)
Criterion Audit	4.02	4.63	2.40	3.90	3.42	1.78	3.67	<b>4.73</b>	3.57 (3.39–3.79)
MSF	4.85	4.72	5.00	<b>5.83</b>	<b>4.73</b>	<b>4.78</b>	<b>5.60</b>	<b>4.63</b>	5.02 (4.82–5.22)
Case Analysis	<b>5.37</b>	5.10	4.03	3.67	3.52	2.52	4.40	<b>5.15</b>	4.22 (4.02–4.42)
PSQ	4.07	3.70	<b>6.03</b>	2.92	3.28	3.07	4.75	2.95	3.85 (3.65–4.04)
Mean (95% CI)	4.86 (4.77–4.97)	4.71 (4.60–4.82)	4.16 (4.06–4.28)	3.58 (3.45–3.69)	3.39 (3.25–3.51)	2.64 (2.54–2.78)	4.42 (4.32–4.55)	4.58 (4.48–4.70)	4.05 (3.98–4.12)

Educationalists, trainers and registrars demonstrated broad agreement. Their opinion is represented by mean scores on a 1-7 scale (see Table One).

Tools perceived as testing qualities well are shaded and in bold font.

In general, qualities agreed to be assessed well by all tools including a machine marked paper (MCQ) and Clinical Skills Assessment (CSA) by Objective Structured Clinical Examination (OSCE), on average, included (1) good clinical care, (2) maintaining good medical practice, (3) relationships with patients, (4) professionalism, and (5) critical thinking. In contrast, qualities thought to be assessed less well by these tools, on average, included (1) ability to work with colleagues, (2) probity, and (3) health.

APPENDIX 2

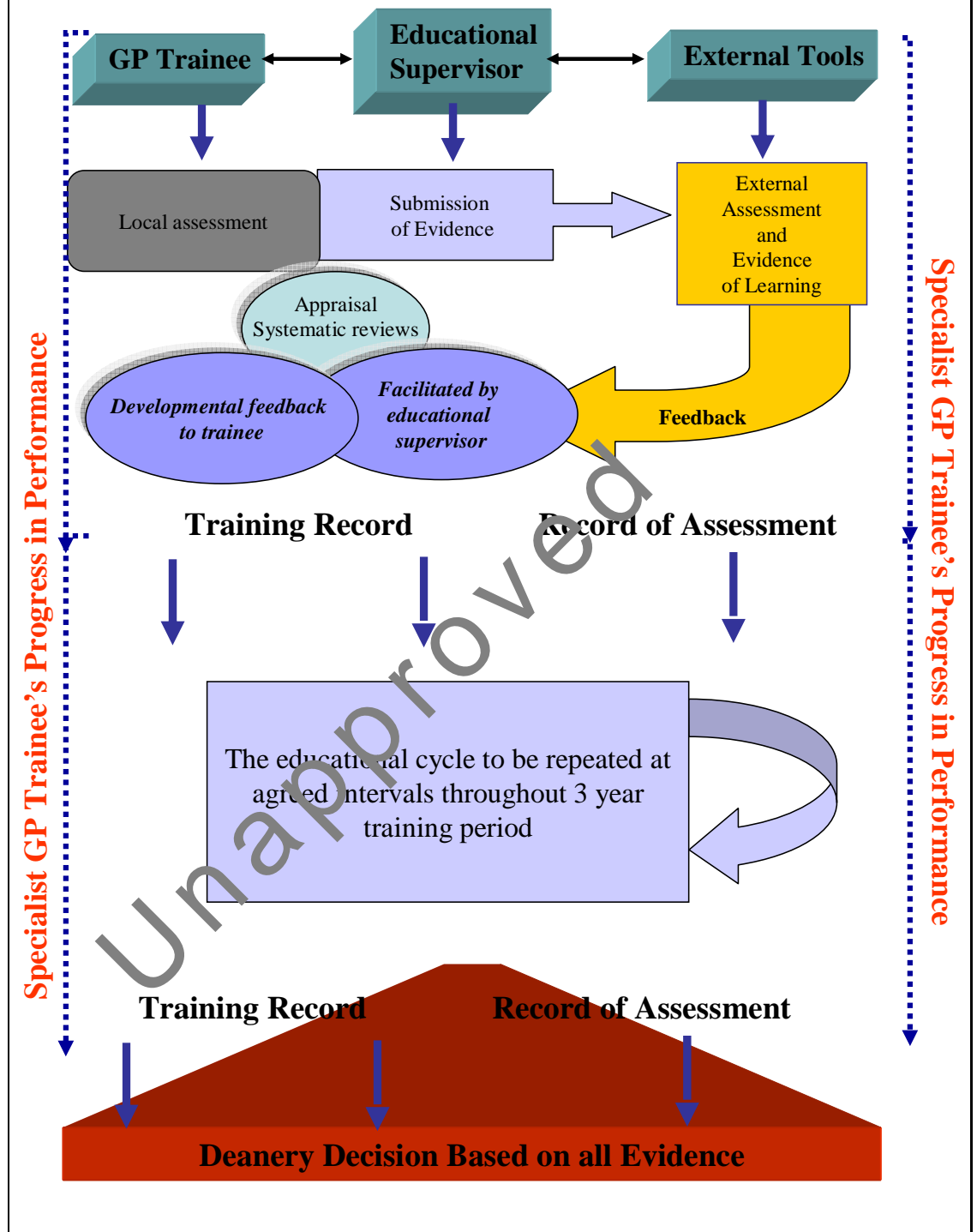
Correlation is significant at the 0.01 level (2-tailed). \*\*.

Correlations

		MN_REFER	M_PSQ_IT	M_PSQ_GR	MN_VIDEO	MMSF_CL1	MMSF_CL2	MMSF_NC1
MN_REFER	Pearson Correlation	1	.299	.299	.596	-.172	-.135	-.160
	Sig. (2-tailed)	.	.200	.201	.091	.443	.549	.525
	N	72	20	20	9	22	22	18
M_PSQ_IT	Pearson Correlation	.299	1	.922**	.009	-.071	-.120	.496
	Sig. (2-tailed)	.200	.	.000	.972	.845	.741	.101
	N	20	66	66	19	10	10	12
M_PSQ_GR	Pearson Correlation	.299	.922**	1	.015	-.050	-.109	.405
	Sig. (2-tailed)	.201	.000	.	.952	.891	.764	.191
	N	20	66	66	19	10	10	12
MN_VIDEO	Pearson Correlation	.596	.009	.015	1	.074	.040	-.560
	Sig. (2-tailed)	.091	.972	.952	.	.800	.891	.058
	N	9	19	19	30	14	14	12
MMSF_CL1	Pearson Correlation	-.172	-.071	-.050	.074	1	.921**	.605**
	Sig. (2-tailed)	.443	.845	.891	.800	.000	.000	.000
	N	22	10	10	14	16	46	37
MMSF_CL2	Pearson Correlation	-.135	-.120	-.109	.040	.921**	1	.628**
	Sig. (2-tailed)	.549	.741	.764	.891	.000	.	.000
	N	22	10	10	14	46	46	37
MMSF_NC1	Pearson Correlation	-.160	.496	.405	-.560	.605**	.628**	1
	Sig. (2-tailed)	.525	.101	.191	.058	.000	.000	.
	N	18	12	12	12	37	37	42



Appendix 3 -Table 1. The Educational Model of WPBA for nMRCGP



Unapproved

**Appendix 4 Table 2  
WPBA - PROPOSED ASSESSMENT FRAMEWORK**

	HOSPITAL Year 1	HOSPITAL Year 2	PRACTICE Year 3
<b>External Tools Core Framework WPBA UK CORE WPBA TOOLS</b>	<b>PSQ X 1</b>  <b>MSF X 2 clinical raters</b>	<b>PSQ Re-sit option</b>  <b>MSF Re-sit option</b>	<b>PSQ X1</b>  <b>MSF X 2 clinical &amp; non-clinical raters</b>
<b>Training Record</b> (additional assessment tool requirements)	CBD	CBD	CBD COT (video)
<b>Additional Teaching &amp; Training Resources</b>  (local feedback where applicable)	Possible locally decided application) DOPS (Possible locally decided application) mini CEX		DOPS      N/A GP context
			mini CEX
			SEA Referrals Audit

KEY      CBD      Case based discussion  
           COT      Consultation observation tool  
           CEX      Clinical evaluation exercise  
           DOPS     Direct observation procedural skills  
           MSF      Multi-source feedback (360 degree appraisal)  
           PSQ      Patient satisfaction questionnaire  
           SEA      Significant event analysis

.....

---

Appendix 5

## Membership of the Steering Group

Dr David Bruce (Chair) – Director of Postgraduate GP Education, East of Scotland Deanery

Dr Murray Lough – Assistant Director of Postgraduate GP Education, West of Scotland Deanery

Dr John Moy – Associate Advisor, South East Scotland Deanery

Dr Douglas Murphy – Associate Advisor, West of Scotland Deanery

Dr Gordon McLeay – Assistant Director of Postgraduate GP Education, East of Scotland Deanery

Dr Ronald McVicar – Assistant Director of Postgraduate GP Education, North of Scotland Deanery



COGPED

