

**Development and Application of Multi-Source Feedback  
for the  
Assessment of Consultant Surgeons.  
for the  
Specialty Advisory Board in General Surgery of the Royal College of Surgeons  
of Edinburgh.**

**Peter J. Driscoll**

**February 2008.**

Jolly defines assessment as “the measurement of an individual’s (or group’s) performance against external criteria” (Jolly, 1997). Although often seen as relating purely to examinations, assessment encompasses any educational process that records the development or progress of the learner and/or provides evidence of such progression within an educational programme or career (de Cossart & Fish 2005). Timely assessment is pleuripotent: it monitors and documents skills acquisition while identifying strengths and weaknesses to facilitate feedback; it allows the ranking of trainees for summative purposes, motivating both trainee and trainer; and it permits comparison of training programmes helping to maintain standards (Jolly, 1997).

Assessment in surgery has historically focussed on an individual’s knowledge or skills as compared with the curriculum laid down by the Royal Colleges. However, the qualities of a good surgeon also encompass personal values, attitudes and behaviours. Previously assessed both informally and subjectively by individual trainers, these areas were made explicit in the document ‘Good Medical Practice’ published by the General Medical Council in 2001 (General Medical Council , 2001). Their assessment and documentation has therefore become a priority.

Miller recognises four levels of assessment; two cognitive and, above these, two behavioural (Miller, 1990). The first level involves the gaining of knowledge (‘knows’) closely followed by the ability to theoretically apply that knowledge (‘knows how’). Further assessment requires the observation of an individual’s behaviour. The third level involves the demonstration of skills arising from the first two cognitive levels (‘shows how’) and includes assessment strategies such as assessment procedures, OSCEs, skills stations and clinical scenarios. All of these have a common factor; the assessee is aware that they are being assessed. However, here there is a problem in that an individual who knows he/she is being assessed may modify their behaviour in order to pass the test; this ‘Hawthorne Effect’ was first described in the assessment of workers at Hawthorne power plant in Chicago (Roethlisberger *et al* , 1939) and illustrates that demonstration does not equate to day-to-day behaviour. This is also well-described in medicine (Rethans *et al*, 1990; Rethans *et al*, 1991) The assessment of everyday behaviour is the fourth level of Miller’s assessment pyramid (‘does’) the implication is that it requires the assessment of individuals without their knowledge, raising issues of both ethics and employment law. However, the field of anthropology may offer an alternative approach in the form of *ethnography*, the study of a population or culture. Most frequently described in aboriginal tribes, the observer becomes integrated into the group whilst repeatedly making observations until eventually he/she is ignored by those being assessed. Everyday behaviour may now be observed in the absence of demonstration or influence of the observer (Atkinson *et al*, 2005). Thus, the assessment of day-to-day surgical practice must be repeated and ongoing so that the assessee learns to ignore it, ceasing demonstration and reverting to everyday practice, so allowing its observation and assessment. It is clearly impractical to shadow every surgeon in everything they do (the direct equivalent of ethnography). However, by incorporating assessments from individuals from every area of clinical practice, in the full knowledge of those being assessed, it is possible to obtain a fuller picture of their practice than would otherwise be possible. This forms the basis of multi-source feedback (MSF).

Multi-source feedback is also termed ‘360 degree’ assessment, a term which is perhaps more useful since it emphasises the need for assessments not just from colleagues but also from juniors, seniors, nurses and, in some cases, patients. The technique has been applied to the assessment of trainees by The Royal Australasian College of Physicians (Paget *et al*, 1996), The Royal College of Obstetricians and Gynaecologists in the UK and the Canadian Medical Licensing Authority (Hall *et al*,

1999) but individual psychometric properties remain undescribed. The Edinburgh Basic Surgical Trainee Assessment Form (EBSTAF) was developed by consensus (Baldwin *et al*, 1999) and is one of the few to be validated in its application (Paisley *et al*, 2001), but again relates exclusively to trainees.

First applied in industry, MSF has emerged as the dominant process for the assessment of professional attitudes and behaviours in the workplace, having been shown to be practical, reliable and reasonably valid. MSF will, however, never be totally safe from challenge since each assessment is, by its very nature, subjective. Furthermore, like any other assessment tool, MSF needs to be applied correctly for the resulting assessment to mean anything. It is therefore important to learn from other industries that have successfully applied these techniques. Wood *et al* recently reviewed published MSF systems in which they quote McCarthy & Garavan (McCarthy *et al*, 2001) by detailing six applications of MSF from industry: the identification of strengths and weaknesses of both individual and organisation; enhancing culture change; summative assessment of performance; evaluation of an individual's potential for selective purposes; enhancing teamwork by allowing members to comment; and identifying training needs for the system. Wood then goes on to relate these to the healthcare literature in which almost all the published studies aim to identify individuals with interpersonal problems on either a summative or formative basis. In doing so, they emphasise the importance of looking at the results of such assessments with care since if those identified as below standard are to suffer a disadvantage (such as loss of promotion or reduced pay) then the diagnosis must be certain. In contrast, if sub-standard individuals are to receive skilled feedback and/or additional training then one can afford to be less certain since although it may be unnecessary and waste time, it may also serve to safe-guard patients and potentially rescue poorer performers (Wood *et al*, 2006).

Akin to assessment methods developed previously, MSF makes accepted behaviours explicit; it thus drives behaviours. The descriptors therefore need to be sufficiently specific and wide reaching so as to describe the full range of desired behaviours since they will, effectively, act as the curriculum. The development of MSF systems is therefore time-consuming and highly skilled, calling upon expertise in psychology and behavioural sciences rather than knowledge of surgery itself. Many MSF systems have been described prior to and since the GMC's Good Medical Practice document (General Medical Council, 2001). However, as far back as 1975, Linn *et al* raised concerns as to what MSF was actually assessing. By factor analysis of their own 16-item four-point scale they showed that 40% of the total variance was due to what they termed an 'interpersonal or relationship factor'. They also highlight a second 'knowledge or skill factor' that would be best assessed by other means and yet was responsible for a further third of the variance (Linn *et al*, 1975). This has been repeatedly described since, whereby the same two factors have actually been assessed by MSF thus demonstrating the overwhelming 'halo effect' of interpersonal ability on the final outcome (Davidge *et al*, 1980; Dielman *et al*, 1980; Maxim *et al*, 1987; Risucci *et al*, 1989; Ramsey *et al*, 1993). The implication is that if you are a good 'people person' your failings may not be revealed by MSF assessment.

To date, no study has demonstrated a lasting relationship between excellence within an assessment and similarly outstanding performance in practice. However, the application of MSF in industry, where descriptors of performance are more developed, would suggest better evidence for the predictive validity of MSF over other strategies (Church, 2000). If we therefore accept that MSF best assesses interpersonal non-technical behaviours, the problem in surgery then becomes the lack of effective descriptors of good (or indeed bad) behaviour, making any observational study highly subjective. Again, lessons from high-reliability industries should be drawn upon

where failures in what have become known as 'non-technical skills' or 'humanistic factors'. Parallels have been repeatedly drawn between aviation and anaesthetics wherein periods of inactive monitoring (the flight or operation itself) are interspersed with highly skilled and performance-critical periods (take-off and landing, or induction and reversal). A review of airline crashes by the North American Space Administration (NASA) in 1979 concluded that 70% had resulted not from mechanical failures but from failures in interpersonal communication, teamwork, decision-making and leadership (Cooper *et al* , 1980). Airlines reacted to these findings by the development of behavioural marker systems for the observation of pilots' non-technical skills (NOTECHS - (Avermaete and Kruijzen 1998)) in parallel with Crew Resource Management (CRM) courses. These CRM courses accepted Reason's theory that human error is ubiquitous and inevitable and aimed not to make pilots error-free but instead provided them with countermeasures to avoid, trap and mitigate the consequences of error before disaster (Helmreich *et al*, 1999). CRM training was made mandatory in 1993 by the UK Civil Aviation Authority (CAA) for all UK pilots and has subsequently been adopted in other high-reliability environments including air-traffic control, the military, off-shore oil, fire services, and nuclear power (Flin *et al*, 2002). A similar review of medical critical incidents by Williamson in 1993 made similar conclusions regarding medical error, with 70-80% resulting from non-technical errors related to communication (Williamson *et al*, 1993). Drawing on the experience of the airlines, anaesthesia took the lead to develop Anaesthesia Crisis Resource Management (ACRM) courses (Howard *et al*, 1992; Gaba *et al* , 1994) and subsequently produced a taxonomy of behaviours made up of task management, teamworking, situation awareness and decision-making (ANTS – (Fletcher *et al* , 2001; Fletcher *et al*, 2003)). This differed quite markedly from that developed in airline pilots and served to illustrate what may have been intuitively obvious; anaesthetists were not pilots. Anaesthetists had their own anaesthesia-specific behaviours and therefore required their own anaesthesia-specific behavioural marker system. Surgeons are no different; the observation of surgeons' behaviours requires a surgeon-specific behavioural marker system and not the grafting of one developed within another environment, a practice that has been described and should be discouraged (Yule *et al*, 2006b). In this area, the Royal College of Surgeons of Edinburgh is in an advantageous position having been directly involved in the development and validation of just such a taxonomy of surgical non-technical skills (NOTSS – (Yule *et al*, 2006a)) developed during the intraoperative phase of an operation. This comprises 5 categories (situation awareness, decision making, task management, leadership and communication / teamwork) each subdivided into elements with illustrative 'good' and 'bad' behaviours. Yule has further demonstrated, perhaps more importantly, that such skills may respond to training (Yule *et al*, 2007). This last point is vital since there seems little point in documenting an individual's failings if nothing can be done about them, and it remains the subject of ongoing work.

MSF should therefore be constructed using well-researched and validated criteria. Furthermore, assessors must themselves be trained to use MSF prior to its application in order to maximise its reliability and therefore validity. This raises the issue of who should assess and how many assessors per assessment episode should be used.

The use of non-medical assessors was first described by Butterfield & Pearson who gave nurses the task of assessing the 'humanistic behaviours' of doctors. In doing so they illustrated distinct differences in opinion as to what qualities were desirable, leading them to question whether assessment of doctors by nurses was appropriate (Butterfield *et al*, 1990). Similar issues have since been raised in the assessment of doctors by patients, whose assessments may be unduly influenced by issues such as

the timeliness with which they see the doctor or whether they leave the consultation satisfied with the outcome. Such issues may be addressed by increasing the number of assessors and drawing from multiple disciplines, thus reducing the subjectivity and bias integral to single source assessment. However, this is logistically challenging and increases costs so that it is equally important not to use more assessors than are required. The evidence would suggest that the optimum number depends on the assessors themselves; whereas between 5 and 10 peers / colleagues may be required to get a representative result, this may need to be increased to 10 to 20 nurses and over 50 patients (Butterfield *et al*, 1991; Ramsey *et al*, 1993; Wenrich *et al*, 1993; Woolliscroft *et al*, 1994; Ramsey *et al*, 1996). However, the techniques used in the literature vary widely both in the training of the assessors and the focus of the assessment instruments.

It would appear that, in common with all other forms of assessment, the quality of MSF depends specifically on what is being assessed, who is doing the assessing and how they have been trained. The literature suggests MSF has great potential in the assessment of surgeons' behaviours, but only if applied correctly.

However, there is one further issue that must be discussed. MSF has repeatedly demonstrated mismatch with individuals' self-assessments, which tend to be higher (Mabe *et al*, 1982; Fletcher, 1999; Van der Heijden *et al*, 2004). It has been suggested that this illustrates a lack of insight that could be corrected by MSF as it identifies so-called 'blind spots'. However, in management at least, MSF from subordinates may be ignored (Bernadin *et al*, 1993), countering the beneficial effects, and even lead to ill-feeling. It is vital that the criteria examined by MSF are acceptable to those being assessed; if they feel the assessment to be irrelevant to their day-to-day practice, they are likely to ignore the outcome. This may be countered by the use of penalty if future assessments fail to show improvement (akin to the proverbial stick). However, if the assessee recognises the assessment to be relevant, he/she is likely to actively strive to improve (the proverbial carrot) with far better results. This has been repeatedly shown in industry (McEvoy *et al*, 1987; Fedor *et al*, 1989; Yuki *et al*, 1995; Wimer *et al*, 1998) but medicine has been slow to recognise the importance of assessee opinion on how they are assessed with only one paper to date looking specifically at this issue (Driscoll *et al*, 2003). This acceptance is vital to the shared understanding and development of a positive and supportive culture where non-technical professional behaviours are recognised as being as important as technical prowess, and where deficient individuals welcome the insight that MSF may afford in an effort to improve.

In conclusion, in developing MSF for the assessment of consultant surgeons we must be clear as to what it is we are assessing and how such assessments can legitimately be used subsequently. MSF must be explicit, simple and include trained assessors and the acceptance of those consultants to be assessed. Finally, development should be ongoing and integral to its application.

## References

1. Atkinson P, Pugsley L. 2005. Making sense of ethnography and medical education. *Medical Education* 39:228-34.
2. Avermaete J, Kruijsen E. 1998. NOTECHS. The evaluation of non-technical skills of multi-pilot aircrew in relation to the JAR-FCL requirements. Final Report NLR-CR-98443. Amsterdam, Netherlands.: National Aerospace Laboratory.;
3. Baldwin PJ, Paisley AM, Paterson-Brown S. 1999. Consultant surgeons' opinion of the skills required of basic surgical trainees. *Br J Surg* 86(8):1078-82.
4. Bernadin H, Dahmus S, Redmon G. 1993. Attitudes of first-line supervisors toward subordinate appraisals. *Human Resource Management* 32(2-3):315-24.
5. Butterfield OS, Mazzaferri EL. 1991. A new rating form for use by nurses in assessing residents' humanistic behaviour. *J Gen Intern Med* 6:155-61.
6. Butterfield OS, Pearson JA. 1990. Nurses in resident evaluation: a qualitative study of the participants' perspectives. *Evaluation and the Health Professions* 13:453-73.
7. Church AH. 2000. Do higher performing managers actually receive better ratings? a validation of multirater assessment methodology. *Consulting Psychology Journal: Practice and Research*(52):-99.
8. Cooper GE, White MD, Lauber JK. 1980. Resource Management on the Flight Deck: Proceedings of a NASA/Industry Workshop (NASA CP-2455). Moffett Field, CA: NASA - Ames Research Center.
9. Davidge AM, Hull AL. 1980. A system for the evaluation of medical students' clinical competence. *Journal of Medical Education* 55:65-7.
10. de Cossart I, Fish D. 2005. Assessment and its role in education for clinical practice: an overview. *Cultivating a thinking surgeon: new perspectives on clinical teaching, learning and assessment*. Shrewsbury: tfm Publishing; p 93-118.
11. Dielman TE, Hull AL, Davis WK. 1980. Psychometric properties of clinical performance ratings. *Evaluation and the Health Professions*(3):-103.
12. Driscoll PJ, Paisley AM, Paterson-Brown S. 2003. Trainees' opinions of the skills required of basic surgical trainees. *Am J Surg* 186(1):77-80.
13. Fedor D, Bettenhausen K. 1989. The impact of purpose, participant preconceptions and rating level on the acceptance of peer evaluations. *Group and Organisational Studies* 14(2):182-97.
14. Fletcher C. 1999. The implications of research on gender differences in self-assessment and 360 degree appraisal. *Human Resource Management* 9(1):39-46.

15. Fletcher G, Flin R, McGeorge P, Glavin RJ, Maran NJ, Patey R. 2001. Final Report: Development of a Behavioural Marker System for Anaesthetists Non-Technical Skills (ANTS). University of Aberdeen Grant Report for SCPMDE project reference: RDNES/991/C.
16. Fletcher G, Flin R, McGeorge P, Glavin R, Maran N, Patey R. 2003. Anaesthetists' Non-Technical Skills (ANTS): evaluation of a behavioural marker system. *Br J Anaesth* 90(5):580-8.
17. Flin R, O'Connor P, Mearns K. 2002. Crew resource management: improving safety in high reliability industries. *Team Performance Management* 8:68-78.
18. Gaba DM, Fish SK, Howard SK. 1994. *Crisis Management in Anaesthesiology*. New York: Churchill Livingstone.
19. General Medical Council. 2001. *Good Medical Practice*. 2 ed. London: General Medical Council.
20. Hall W, Violato C, Lewkonja R, Lockyer J, Fidler H, Toews J, Jennett P, Donoff M, Moores D. 1999. Assessment of physician performance in Alberta: the physician achievement review. *Can Med Assoc J* 161:52-7.
21. Helmreich RL, Merritt AC, Wilhelm JA. 1999. The evolution of Crew Resource Management training in commercial aviation. *Int J Aviat Psychol* 9(1):19-32.
22. Howard SK, Gaba DM, Fish KJ, Yang G, Sarnquist FH. 1992. Anesthesia crisis resource management training: teaching anesthesiologists to handle critical incidents. *Aviat Space Environ Med* 63(9):763-70.
23. Jolly B. 1997. Assessment and appraisal. *Medical Education* 31 Suppl 1:20-4.
24. Linn BS, Arostegui M, Zeppa R. 1975. Performance self assessment. *British Journal of Medical Education* 9:98-101.
25. Mabe P, West S. 1982. Validity of self-evaluation of ability: a review and meta-analysis. *Applied Psychology* 67:280-96.
26. Maxim BR, Dielman TE. 1987. Dimensionality, internal consistency and interrater reliability of clinical performance ratings. *Medical Education* 21:130-7.
27. McCarthy AM, Garavan TN. 2001. 360 degree feedback processes: performance improvement and employee career development. *Journal of European Industrial Training* 25(1):3-32.
28. McEvoy G, Buller P. 1987. User acceptance of peer appraisals in an industrial setting. *Personnel Psychology* 40:785-97.
29. Miller GE. 1990. The assessment of clinical skills/competence/performance. *Acad Med* 65(9 Suppl):S63-S67.
30. Paget NS, Newble DI, Saunders NA, Du J. 1996. Physician assessment pilot study for the Royal Australasian College of Physicians. *J Contin Educ Health Prof* 16:103-11.

31. Paisley AM, Baldwin PJ, Paterson-Brown S. 2001. Feasibility, reliability and validity of a new assessment form for use with basic surgical trainees. *Am J Surg* 182(1):24-9.
32. Ramsey PG, Carline JD, Blank LL, Wenrich MD. 1996. Feasibility of hospital-based use of peer ratings to evaluate the performances of practicing physicians. *Academic Medicine* 71(4):364-70.
33. Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. 1993. Use of peer ratings to evaluate physician performance. *JAMA* 269:1655-60.
34. Rethans JJ, Sturmans F, Drop R, van d, V, Hobus P. 1991. Does competence of general practitioners predict their performance? Comparison between examination setting and actual practice. *BMJ* 303(6814):1377-80.
35. Rethans JJ, van Leeuwen Y, Drop R, van d, V, Sturmans F. 1990. Competence and performance: two different concepts in the assessment of quality of medical care. *Fam Pract* 7(3):168-74.
36. Risucci DA, Tortolani AJ, Ward RJ. 1989. Ratings of surgical residents by self, supervisors and peers. *Surg Gynecol Obstet* 169(6):519-26.
37. Roethlisberger FJ, Dickson WJ. 1939. *Management and the Worker: An Account of a Research Program Conducted by Western Electric Company, Hawthorne Works, Chicago.* Cambridge, Massachusetts: Harvard University Press.
38. Van der Heijden BI, Nijhof AH. 2004. The value of subjectivity: problems and prospects for 360 degree appraisal systems. *International Journal of Resource Management* 15(3):493-511.
39. Wenrich MD, Carline JD, Giles LM, Ramsay P. 1993. Ratings of the performances of practising internists by hospital based registered nurses. *Academic Medicine* 68:680-7.
40. Williamson JA, Webb RK, Sellen A, Runciman WB, van der Walt JH. 1993. Human failure: an analysis of 2000 incident reports. *Anaesth Intensive Care* 21:678-83.
41. Wimer S, Nowack K. 1998. Thirteen common mistakes using 360 degree feedback. *Training & Development* 52(5):69-79.
42. Wood L, Hassell A, Whitehouse A, Bullock A, Wall D. 2006. A literature review of multi-source feedback systems within and without health services, leading to 10 tips for their successful design. *Medical Teacher* 28(7):e185-e191.
43. Woolliscroft JO, Howell JD, Patel BP, Swanson DB. 1994. Resident-patient interactions: the humanistic qualities of internal medicine residents assessed by patients, attending physicians, program supervisors, and nurses. *Academic Medicine* 69(3):216-24.
44. Yuki G, Lepsinger R. 1995. How to get the most of 360 degree feedback. *32* 12(45):50.

45. Yule S, Flin R, Paterson-Brown S, Maran N, Rowley D. 2006a. Development of a rating system for surgeons' non-technical skills. *Medical Education* 40:1098-104.
46. Yule S, Flin R, Paterson-Brown S, Maran N, Rowley D, Youngson G. 2007. Teaching surgeons about non-technical skills. *Surgeon* 5(2):86-9.
47. Yule S, Flin R, Paterson-Brown S, Maran N, Yule S, Flin R, Paterson-Brown S, Maran N. 2006b. Non-technical skills for surgeons in the operating room: a review of the literature. [Review] [72 refs]. *Surgery* 139(2):140-9.