ORIGINAL ARTICLE

A theory-based instrument to evaluate team communication in the operating room: balancing measurement authenticity and reliability

Lorelei Lingard, Glenn Regehr, Sherry Espin, Sarah Whyte

Qual Saf Health Care 2006;**15**:422–426. doi: 10.1136/qshc.2005.015388

See end of article for authors' affiliations

Correspondence to: Dr Lorelei Lingard, Wilson Centre for Research in Education at University Health Network, University of Toronto, 200 Elizabeth Street, Eaton South 1–604, Toronto, Ontario, Canada M5G 2C4; lorelei.lingard@utoronto.ca

Accepted 13 August 2006

Background: Breakdown in communication among members of the healthcare team threatens the effective delivery of health services, and raises the risk of errors and adverse events.
Aim: To describe the process of developing an authentic, theory-based evaluation instrument that measures communication among members of the operating room team by documenting communication failures.
Methods: 25 procedures were viewed by 3 observers observing in pairs, and records of events on each communication failure observed were independently completed by each observer. Each record included the type and outcome of the failure (both selected from a checklist of options), as well as the time of occurrence and a description of the event. For each observer, records of events were compiled to create a profile for the procedure.
Results: At the level of identifying events in the procedure, mean inter-rater agreement was low (mean agreement across pairs 47.3%). However, inter-rater reliability regarding the total number of communication failures per procedure was reasonable (mean ICC across pairs 0.72). When observers recorded the same event, a strong concordance about the type of communication failure represented by the event was found.
Discussion: Reasonable inter-rater reliability was shown by the instrument in assessing the relative rate of communication failures displayed per procedure. The difficulties in identifying and interpreting individual communication events reflect the delicate balance between increased subtlety and increased error. Complex team communication does not readily reduce to mere observation of events; some level of interpretation is required to meaningfully account for communicative exchanges. Although such observer interpretation improves the subtlety and validity of the instrument, it necessarily introduces error, reducing reliability. Although we continue to work towards increasing the instrument's sensitivity at the level of individual categories, this study suggests that the instrument could be used to measure the effect of team communication intervention on overall failure rates at the level of procedure.

Effective communication in a team is not merely a professional courtesy,[1 2] but also a critical factor in ensuring the delivery of effective health services[3] and avoiding error and adverse events.[4–7] Communication has the power to either foster or threaten safe and effective healthcare. As recognition of this power has grown, so has the impetus to discern the features of strong and weak communication, and to develop methods for improving interprofessional communicative practice. Our recent research in the operating room has elaborated a theory of interprofessional communication in a team that describes tension catalysts, shows interpretive patterns and classifies recurrent failures.[8–10] This work suggests clear directions for educational interventions aimed at improving the status quo of operating room communication practices.[11]

Assessing the effectiveness of such interventions requires appropriate measures of communication in a team. The challenge in creating such measures is to provide analytical traction, while continuing to reflect the complex and evolving nature of communication in a team. Although performance measures of a team exist in healthcare, they have generally been designed with other objectives in mind. Thus, many of these measures treat communication as a fairly simple construct representing just one dimension in a global measure of healthcare team performance.[12–14] Others are more elaborated, but are constrained by their reliance on self-report.[15–18] Effective measures have been developed in other domains such as aviation[19]; however, their context sensitivity renders them difficult to transfer to disparate settings such as the operating room, which has larger teams, more diversity of disciplines, different spatial organisation and less predictability of communicative content.

To deal with this measurement need and to support education, patient safety and quality improvement efforts in the domain of communication among members of an operating room team, we developed a theory-based instrument that reflected the findings of our grounded theory research.[10] Among other findings, our work defined patterns of ''communication failure'' in the operating room, and classified these communication failures into four categories based on rhetorical theory:

- *Occasion*: Occasion included problems related to time and space. For instance, a common timing problem was the surgeon's post-incision question to the anaesthesiologist regarding administering antibiotics. In this case, the question is too late to be of maximum use as a reminder or prompt to deliver antibiotics, which ought to be given before incision.

Abbreviation: SEM, standard error of measurement

- *Content*: Content consisted of communicative exchanges that contained incomplete or inaccurate information, such as a nurse's inaccurate announcement that a patient was positive for hepatitis C.
- *Purpose*: Purpose included situations in which questions were asked by one team member, but not responded to by the team, prompting repeated and increasingly urgent requests.
- *Audience*: Audience reflected situations in which a communicative exchange excluded a key person. A recurrent instance of this failure was discussions between members of nursing and anaesthesia teams of how to position a patient on the table in the absence of a surgical representative, whose arrival would then necessitate a change in position to accommodate surgical needs.

Communication failures were frequent, occurring in 30% of procedurally relevant information exchanges. In addition to raising the potential for error, failures often had immediate negative consequences including delay, inefficiency and tension among the team members.

A measurement instrument was designed to capture the presence and type of communication failures being enacted in the operating room, as well as the frequency and nature of any consequences that arose from these communication failures. The purpose of this study was to deal with the feasibility and psychometric properties of the instrument in the operating room setting. This paper reports our progress towards this objective and our experience of trying to balance authenticity and reliability in an observational instrument for a complex communicative phenomenon.

METHODS

Development of the measurement tool

The tool was designed for use by a third-party observer in the operating room theatre, to provide a detailed record of each communication failure observed during the procedure. Each time a communication failure was observed, an ''Information exchange recording form'' was completed (fig 1).

Recognising that communication among members of an operating room team is characterised by a complicated set of overlapping, multifocal conversations that appear and reappear over time, we designed the form to balance ease of use and comprehensiveness of description. For each communication failure observed, the observer identified, using a checklist of options, the types of communication failure and any immediate consequences that arose. In addition, the observer recorded the time of the event, documented contextually relevant observation notes as required and identified continuity between related events that were separated in time by cross-referencing previous forms.

Early versions of the form were repeatedly vetted by four members of the operating room team, and β tested for usability by three members of the research team. Figure 1 presents the final version of the form.

Observer training

Three observers were trained to participate in the reliability testing process. Each had at least 1 year of experience as an observer in our research programme. One observer had clinical expertise as an operating room nurse educator. All three observers and the principle investigator trained on the instrument by applying it to observation transcripts from an earlier research study, making both individual and group decisions about which communication events in the transcript would constitute a failure, of what type, whether an outcome arose from this event and of what type. For this process, operational definitions and iconic examples of each failure category were developed from previously collected observational data to guide the coding. In addition, a ''quick reference'' of category definitions was also provided for use in the field. The training process was extensive. It involved lengthy discussions about different levels of interpretive effort in identifying an event as an error and decisions about ''how much'' interpretation was reasonable and reproducible among the raters. We determined that such interpretation was necessary to account for the complexity of many communicative exchanges, and their embeddedness in earlier exchanges and relationships. However, observers were encouraged to cite objective evidence from the transcript for their interpretations, and when this was not possible, to err on the side of not identifying an event as an error. When the interpretive observational process began to yield consistent consensus in the group, reliability testing of the instrument started in the operating room.

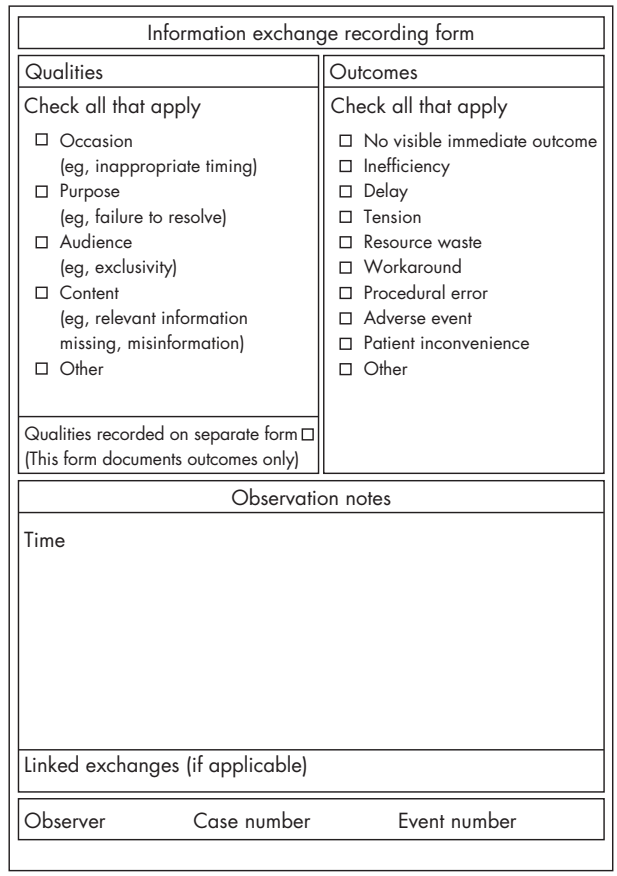

Figure 1 Evaluation form.

Procedure

Over a 3-month period (May–August 2003), 25 surgical patients were selected in general and vascular surgery at an academic health sciences centre. Ethical approval was obtained from the hospital's research ethics board, and observed members of the operating room consented to participate in the study. No operating room personnel declined to participate. Teams consisted of 5–8 members. Team membership varied for each observation session, except for two pairs of procedures (ie, two occasions when two observation sessions were conducted on the same day). Observed procedures ranged from the construction of an AV fistula to abdominal aortic aneurysmectomy, reflecting a range in complexity, case length and training situations. At least two observers attended each 1–4-h observation session. Observers functioned independently, positioning themselves

in separate locations in the operating room to allow for the broadest possible coverage of communication activities, to minimise the Hawthorne effect[20] and to maximise the independence of the observations.

After each observation session, data from the complete set of observation forms were collated and the following scores were generated for the procedure: the total number of communication failures, the frequency of each type of failure, and the frequency and types of consequences that arose. In addition, observers met to debrief about their application of the instrument to the team's communicative activities during the procedure. This helped to clarify discrepancies in the use of the tool; however, the completed forms from the procedure were not adjusted as a function of these debriefings.

Analysis

For each procedure observed, events recorded by the various observers were matched using the time recorded on the sheet. Each event that was recorded by at least one observer was given an event number and included in the analysis. When multiple observers recorded the same event, data from each observer were recorded for that event in the database. When an observer failed to record an event that had been recorded by the other observers, the data for that observer were identified as missing.

Analyses were separately performed for the presence and type of communication failure and for the presence and type of resulting consequences. As all three observers were not present for all procedures, it was necessary to analyse the data in a pairwise fashion, by calculating the inter-rater reliability separately for each of the three pairings of observers using the data from procedures that each pair jointly observed.

The data were assessed for inter-rater reliability in three ways. Firstly, at the level of individual events across all operative procedures seen by each pair of examiners, a simple percentage agreement was calculated as the number of events that were recorded by both observers as a proportion of the total number of events recorded by either observer.

Secondly, for each procedure, the total number of events recorded by each observer was calculated as summative statistics, representing the extent to which the procedure was ''failure rich'' or ''failure sparse''. The inter-rater reliability of this continuous measure for each procedure was calculated across all three observers with generalisability analysis (using a rater-by-procedure crossed design with missing data in cells where a rater was not present at the procedure).

Finally, to assess the inter-rater reliability for the classification of events into appropriate types, the subset of events that were identified by both observers across procedures was selected. As each event could be classified in multiple categories, each category was separately assessed for inter-rater reliability using Cohen's κ. Table 1 shows how the κ statistic was calculated separately for each pair of raters for each category, and the averages of these three values are represented in the last column.

RESULTS

The three observers viewed 18, 20 and 16 of the 25 procedures, with the number of jointly viewed procedures for the three pairings of observers being 13, 11 and 9, respectively.

Of the 25 procedures, 3 had no recorded communication failures and one outlier procedure had 15 recorded failures by one of the two raters. The remaining 21 procedures ranged between 1 and 9 failures during the observation period. The mean (standard deviation (SD)) number of failures per procedure for each of the three observers were 2.56 (2.19), 2.80 (2.17) and 3.00 (3.77), respectively. The mean (SD) number of failures per procedure averaged across observers within a procedure was 2.87 (2.68), suggesting reasonable variability in the number of failures across procedures. Table 1 gives a detailed summary of the mean frequency of each type of failure per procedure for each observer, which is presented in columns 2–4.

In recording individual communication failures in a procedure, there was relatively low agreement among observers about whether a communication failure had occurred. For example, of the 61 events recorded as a communication failure by either observer 1 or observer 2 during their 13 jointly observed procedures, only 23 (37%) events were identified by both observers as a failure. The equivalent index of agreement for the other two observer pairings was somewhat higher at 50% (9 jointly identified events among the 18 events recorded by either observer) and 56% (22 jointly identified events among the 39 events recorded by either observer).

Despite this relatively low agreement index at the level of individual occurrences, there was quite a reasonable inter-rater reliability in the total number of communication failures identified per procedure. The generalisability analysis produced variance components for rater, procedure and rater-by-procedure of −0.09 (assumed to be zero), 6.13 and 1.93, respectively, resulting in a calculated procedure-level single-rater generalisability coefficient of 0.76 (table 1, column 5) and a single-rater standard error of measurement (SEM) of 1.39 (as the rater variance is zero for this analysis, the relative and absolute SEM are identical). Using a D study analysis, this would suggest that the average of two independent raters' scores would produce a reliability of 0.86 (with a relative and absolute SEM of 0.98) and the average of three raters' scores would produce a reliability of 0.90 (with a relative and absolute SEM of 0.80).[i]

Further, although based on quite small numbers (23, 9 and 22 for the pairs), when both observers did record the same event as a communication failure, there was strong concordance about the type of communication failure represented by the event. As column 6 of table 1 shows, with the exception of the assignment of failures to the ''purpose'' category, the mean κ coefficients for each type of communication failure (calculated as the average of the κ coefficients for each of the three observer pairs) ranged from ''good'' (defined by Fleiss[21] as >0.60) to ''very good'' (defined by Fleiss[21] as >0.80). Again, with the exception of the categorisations of ''purpose'' failures, the κ values for the individual pairs ranged from 0.55 (labelled by Fleiss[21] as ''moderate'') to 0.91 (labelled by Fleiss[21] as ''excellent'').

The pattern of results was similar (but slightly lower) for the recording of outcomes arising from communication failures (table 1)—that is, the average index of observer agreement at the level of individual events was moderate to poor at 46%. At the level of the total number of outcomes per procedure, the generalisability analysis produced variance components for rater, procedure and rater-by-procedure of 0.001, 2.76 and 1.55, respectively, resulting in a single-rater reliability of 0.64 (with an absolute SEM of 1.25). Using a D study analysis, this would suggest the need for three observers per procedure to achieve a reliability of at least 0.80. The mean κ coefficients for each type of failure, where

[i]Recognising the relatively small sample size and the potential effect of a single outlier on the stability of the reliability estimates, the analyses were repeated with the 15-failure procedure removed. As the second rater, identified only nine failures for this procedure, the error variance was more inflated for this reason than the procedure variance. Therefore, the reliability estimates actually improved with the exclusion of this procedure (to 0.83 with an absolute SEM of 0.91). For completeness, the more conservative analyses with this procedure included are presented.

Table 1 Descriptive statistics and inter-rater reliability for the identification and classification of communication failures and their associated negative outcomes

	Observer 1	Observer 2	Observer 3	Reliability* (procedure)	Reliability† (event)
Number of procedures observed	18	20	16		
Mean number of failures perprocedure					
Total	3.00 (3.77)	2.80 (2.17)	2.56 (2.19)	0.76	
Occasion	0.78 (1.06)	1.10 (1.25)	0.81 (1.05)		0.83
Purpose	0.89 (1.13)	1.10 (1.25)	0.75 (1.00)		0.33
Audience	0.44 (0.71)	0.20 (0.41)	0.56 (0.73)		0.71
Content	1.72 (0.24)	1.75 (0.22)	1.38 (0.25)		0.70
Mean number of outcomes per procedure					
Total	2.64 (2.65)	2.06 (1.60)	2.31 (1.89)	0.64	
Inefficiency	1.43 (1.83)	1.18 (0.88)	1.31 (1.38)		0.81
Delay	0.71 (1.38)	0.12 (0.33)	0.31 (0.63)		0.64
Tension	1.36 (1.60)	1.12 (1.45)	0.92 (1.44)		0.67
Resource waste	0.14 (0.36)	0.12 (0.33)	0.15 (0.38)		—
Workaround					—
Procedural error					—
Adverse event					—
Patient inconvenience					—

*Single-rater generalisability coefficients for the total number of events recorded for each procedure.
†Mean of the three pairwise κ coefficients calculated for the type of event recorded, given that an event was recorded by both observers.

calculable, were good, to very good as seen in table 1, with the individual κ for specific pairs ranging from 0.2 to 1.0.

DISCUSSION

The development of a theory-based evaluation instrument that authentically captures the complex and subtle nature of communication among team members in the operating room has proved to be a challenge. The multifocal, overlapping and evolving nature of communication events—as well as the limitations imposed by the observers' physical location in the operating room—made it difficult for observers to exhaustively capture the full set of relevant events that occurred during each procedure. This sampling limitation could potentially be dealt with by using audio/video technology to standardise observers' vantage point. In our attempt to exploit this technology after this study, however, we found that video-mediated observations were problematic for two key reasons: (1) the quality of even expensive audio equipment was poor in the operating room environment where team members are masked; and (2) the reality of a few cameras and microphones in fixed positions can limit the observers' access to more subtle events of communication altogether.

Simple sampling limitations do not account for all disagreements among the observers. Rather, observers sometimes interpreted the same event differently, drawing our attention to two larger theoretical issues. Importantly, communication events are interconnected such that event B looks different depending on whether or not you witnessed event A. For instance, an interaction between members of nursing and anaesthesia teams about a surgery-relevant issue may seem to exclude the surgeon (ie, audience failure) if a previous interaction regarding this issue had not been witnessed. Similarly, an interaction between members of nursing and surgery teams for which all team members are within earshot may be interpreted as effective communication, until a later exchange shows that one team member did not hear the communication, or that the information exchanged was inaccurate. Observers who do not witness the second exchange will not record a failure.

Further, our decision to encourage the use of observers' judgement about whether events of communication constitute failures introduced a degree of measurement error into our study. For instance, we were not surprised that the lowest of the κ coefficients for type of communication failure was for purpose failures (0.33). Throughout our training period, it was evident that purpose failures required a greater degree of interpretation by observers than the other types of failure. This higher level of need for interpretation was partly owing to the fact that "purpose" is not "visible" in that way that, for example, a team member's absence from a relevant discussion is visible (audience failure). Further, the attribution of motivation inherent in judging the nature of a speaker's purpose and the extent to which it is resolved by ensuing oral and non-oral messages adds further interpretation to determining whether a purpose failure has occurred. By contrast, occasion failures had the highest κ coefficients (0.83), reflecting a lesser degree of interpretation required to determine whether communication was undesirably "late" or not. Even this category, however, was not interpretation free: a few instances of disagreement arose from one observer assessing a communication event as a timing failure when, in post-observation debriefings, another observer described witnessing the same event but determining that it was "late, but not late enough to be considered a failure".

This phenomenon illustrates a central measurement principle that the more finely grained the measurement, the more opportunity for error. Had we opted for less subtle data, we may have had improved reliability, but perhaps at the expense of ecological validity. For instance, we could have trained observers to count strictly "objective" events, such as the number of times that a circulating nurse receives a request for intraoperative equipment that requires retrieval from outside the room (which may constitute a "timing" error) or the number of team decision-making discussions that exclude at least one team member (which may constitute an "audience" error) or the number of times a question goes unanswered for a defined amount of time (which may constitute a "purpose" error). Although such events might be easier for individual observers to agree on, they represent a substantially less sophisticated account of communication in a team and are consequently of less value in assessing the precise effect of an intervention to improve communication. We would contend, therefore, that our decision to grapple with communication events individually and within an evolving social context of discourse, rather

than assigning them a priori meaning, strengthens the authenticity and ecological validity of our approach, even if it makes quantification and reliability a continuing challenge.

Finally, it is worth noting that this study used only three observers viewing (in shifting pairs) only 25 procedures. Although the effort to coordinate even this many paired observations was large, we acknowledge that from a psychometric perspective this represents a relatively small sample size. It is necessary, therefore, to recognise that the estimated parameters in this study, although theoretically unbiased by the small sample size, may be somewhat unstable, and replication on a large scale would provide additional support for our findings.

Notwithstanding these important limitations, the results are promising from the perspective of describing the overall quality of communication in a team over the course of a procedure. That is, this event-based tool, with proper training for observers, showed reasonable inter-rater reliability in assessing the relative rate of communication failures displayed per procedure, in classifying the type of communication failure being enacted and in identifying the consequences of that communication failure for a team functioning in an operating room. Owing to its ability to distinguish failure-rich from failure-sparse procedures, we are confident that the tool could be used to measure the effect of an intervention on communication in a team on overall failure rates at the level of procedure.

Our approach requires considerable training and sophistication of observers, and we continue to work towards increasing sensitivity at the level of individual events and categories. However, we are encouraged to pursue this line of inquiry, as it provides the opportunity of assessing communication among members of an operating room team not by single summative snapshots but rather by assembled records that can be used to construct a multifaceted communication ''profile'' over time. Our theoretical work in this domain has shown that communication in a team is rarely straightforwardly ''good'' or ''bad'', suggesting that measures in this domain need to be structured to pick up patterns that surface across a series of exchanges. The instrument we have developed advances the evaluation of team communication at the level of the procedure, by allowing us to acknowledge and represent these complexities rather than eliding them.

Authors' affiliations
L Lingard, G Regehr, S Whyte, Wilson Centre for Research in Education, University of Toronto, Eaton South, Toronto, Ontario, Canada
S Espin, Faculty of Nursing, Ryerson University, Toronto, Canada

Funding: This research was funded by the Canadian Institutes of Health Research (CIHR) and by the doctors of Ontario through the PSI Foundation. LL is supported by a CIHR New Investigator Award and as the BMO Financial Group Professor in Health Professions Education Research. GR is supported as the Richard and Elizabeth Currie Chair in Health Professions Education Research.

Competing interests: None declared.

REFERENCES

1 **Societal Needs Working Group**. CanMEDS 2000 project. Skills for the new millennium. *Ann R Coll Physicians Surg Can* 1996;**29**:206–16.
2 American Board of Internal Medicine. *Project professionalism*. Philadelphia, PA: American Board of Internal Medicine, 1995.
3 **Firth-Cozens J**. Multidisciplinary teamwork: the good, bad, and everything in between. *Qual Health Care* 2001;**10**:65–6.
4 **Kohn LT**, Corrigan JM, Donaldson MS, eds. *To err is human: building a safer health system*. Washington, DC: National Academy Press, 2000.
5 **Joint Commission on Accreditation of Healthcare Organizations (JCAHO)**. Sentinel event statistics: December 17, 2003. Oakbrook Terrace, IL: JCAHO, 2005http://www.jcaho.org/accredited+organizations/hospitals/sentinel+events/index.htm (accessed 13 Oct 2006).
6 **Gawande AA**, Zinner MJ, Studdert DM, *et al*. Analysis of errors reported by surgeons at three teaching hospitals. *Surgery* 2003;**133**:614–21.
7 **Sutcliffe KM**, Lewton E, Rosenthal MM. Communication failures: an insidious contributor to medical mishaps. *Acad Med* 2004;**79**:186–94.
8 **Lingard L**, Reznick R, Espin S, *et al*. Team communications in the operating room: talk patterns, sites of tension, and implications for novices. *Acad Med* 2002;**77**:37–42.
9 **Lingard L**, Reznick R, DeVito I, *et al*. Forming professional identities on the healthcare team: discursive constructions of the 'other' in the operating room. *Med Educ* 2002;**36**:728–34.
10 **Lingard L**, Espin S, Whyte S, *et al*. Communication failures in the operating room: an observational classification of recurrent types and outcomes. *Qual Saf Health Care* 2004;**13**:330–4.
11 **Lingard L**, Baker R, Espin S, *et al*. Getting teams to talk: development and pilot implementation of a checklist to promote safer operating room communication. *Qual Saf Health Care* 2005;**14**:340–6.
12 **Risser DT**, Simon R, Rice MM, *et al*. A structured teamwork system to reduce clinical errors. In: Spath PL, eds. *Error reduction in healthcare*. Chicago, IL: AHA Press, 2000.
13 **Carthey J**, de Leval M, Wright DJ, *et al*. Behavioral markers of surgical excellence. *Saf Sci* 2003;**41**:409–25.
14 **Young GJ**, Charns MP, Daley J, *et al*. Best practices for managing surgical services: the role of coordination. *Health Care Manage Rev* 1997;**22**:72–81.
15 **Shortell SM**, Rousseau DM, Gillies RR, *et al*. Organizational assessment in intensive care units (ICUs): construct development, reliability, and validity of the ICU nurse-physician questionnaire. *Med Care* 1991;**29**:709–26.
16 **Flin R**, Fletcher G, McGeorge P, *et al*. Anaesthetists' attitudes to teamwork and safety. *Anaesthesia* 2003;**58**:233–42.
17 **Pinto MB**. Gaining cooperation among members of hospital project teams. *Hosp Top* 1990;**68**:15–21.
18 **Hetherington RW**. The effects of formalization on departments of a multi-hospital system. *J Manag Stud* 1991;**28**:103–41.
19 **Helmreich RL**, Merritt AC. *Culture at work: national, organisational and professional influences*. Aldershot: Ashgate, 1998.
20 **Hammersley M**, Atkinson P. *Ethnography: principles in practice*, 2nd edn. New York: Routledge, 1995.
21 **Fleiss JL**. *Statistical methods for rates and proportions*, Vol 2. New York: Wiley, 1981.