

Assessment of general practitioners by video observation of communicative and medical performance in daily practice: issues of validity, reliability and feasibility

Paul Ram,¹ Richard Grol,¹ Jan Joost Rethans,¹ Berna Schouten,¹ Cees van der Vleuten² & Arnold Kester³

Objectives To develop a video assessment method for General Practitioners (GPs) by analysing issues of validity, reliability and feasibility of observation of videotaped regular consultations.

Design In a cross-sectional study consultations of 93 GPs were video recorded in the practice during 1 week. The GPs registered consultation and patient data in a log-book; 16 consultations per GP were selected using preset criteria. The quality of communicative and medical performance of these consultations was assessed by GP observers with a validated instrument. The validity of the procedure was evaluated by checking the content of each GP's sample using specific sample criteria. Selection bias was estimated by multiple regression analysis, with sample characteristics as independent variables and scores on communication and medical performance as dependent variables. The influence of observation on GPs and patients was assessed by a questionnaire. Generalizability theory was used to estimate reliability. Feasibility was assessed by conducting a questionnaire, by keeping accounts, and by checking the technical quality of the videotaped consultations.

Setting Universities of Nijmegen and Maastricht, The Netherlands.

Subjects General Practitioners (GPs).

Results The domain of general practice was well covered in the samples; content validity was satisfactory.

With regard to the sample characteristics, only the total duration of consultations appeared to correlate significantly with both the score on communication and the score on medical performance. A majority (71%) of GPs reported not being influenced by the observation, except in the first cases, and recognizing their usual daily performance in the videotaped consultations. An acceptable level of reliability was reached after 2.5 hours of observation, i.e. 12 cases by a single observer. The method was well accepted by both GPs and patients. The costs were £250 per GP.

Conclusions Video assessment of GPs in daily practice according to the procedures described is a valid and reliable method, one which is useful for education and quality improvement. There is a trade-off between feasibility on one hand and validity, reliability and credibility on the other hand. Compared to investments in observation methods in standardized settings, the costs of video observation of GPs' actual performance are acceptable.

Keywords *Clinical competence; cross-sectional studies; evaluation studies; feasibility studies; *physicians, family; physician-patient relationship; videotape recording.

Medical Education 1999;33:447-454

¹Centre for Quality Research, Universities of Maastricht and Nijmegen, The Netherlands, ²Department of Educational Development and Research, Maastricht University, The Netherlands, and ³Department of Methodology & Statistics, Maastricht University, The Netherlands

Correspondence: Paul Ram MD, General Practitioner, Department of General Practice, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands

Introduction

Assessment of the performance of general practitioners (GPs) in daily practice can play an essential role in improving the quality of care GPs provide to patients. It has been shown that self-ratings of performance usually produce overestimation of service delivery, and that self-selected postgraduate education has little effect on

changing practice behaviour of doctors; objective assessment is preferred.^{1,2} If deficiencies in practice performance could be identified objectively, postgraduate medical education and quality improvement activities could be focused on GPs' insight into their deficiencies and real educational needs in order to make them as effective as possible.^{3,4}

It is uncertain which assessment method(s) should be used. Which have the highest validity and reliability, and which are feasible on a broad scale? One can distinguish between direct and indirect methods of quality assessment.⁵ With the former, clinical activities are observed directly. Indirect methods, on the other hand, infer clinical competencies from written or oral examinations, from chart review, from assessment of referral letters or from data from insurance agencies. A direct method can be applied both in day-to-day practice and in a laboratory setting. Direct observation of actual performance in daily practice has the highest validity, since this method approximates the real professional world as closely as possible.⁶ Direct observation is possible by having standardized patients (SPs) or peer observers visit the practice or by placing video-equipment in the practice. Although used in daily practice, SPs present logistic difficulties.⁷ It is demanding for SPs to observe the complex processes involved in medical care and communication. Practice visitation by peer observers may disrupt real practice and requires considerable resource investments. On the contrary, video recordings of consultations can be judged by different trained observers without consultations being disrupted; there are, however, problems concerning validity, reliability and feasibility.

Validity might be reduced by the 'audience effect': both GPs and patients could be influenced by the video recording.⁸ Moreover, the logistics of performance testing often restrict the selection of consultations to be included into the assessment, thus incurring the risk that the sample may not be representative of actual patient care.⁹

Reliability might be influenced by the content specificity problem: for example, the predictive value of a mark, acquired with a case concerning cystitis to one dealing with angina pectoris turns out to be very low.¹⁰ Moreover, reliability could suffer from the unstandardized and complex problems as they are presented in day-to-day practice. The optimal management of a complex problem in a specific daily context may not be immediately apparent and experts may disagree in grading the quality of performance. A key issue yet to be determined is the duration of observation and the number and range of cases that need to be observed to give a reliable judgement of performance. Even under

standardized conditions the testing time required to achieve an acceptable level of reliability appears to be a limitation of such an assessment method.¹¹

As for feasibility, both ethical, practical and technical aspects may limit the use of video assessment. Videotaping all aspects of the complex care process, including physical examination, may be difficult, while acceptance by the target group may be hampered by fear of abuse of tapes.¹² Also, there are no reports on the cost of video assessment. While various problems can thus be expected in applying video assessment in the practice of GPs, videotapes of daily consultations are already being used successfully, for example in the United Kingdom in the selective assessment of GP trainees and in the Fellowship by Assessment of the Royal College of GPs.^{13,14} The latter has been revised recently to be more objective. However, issues concerning validity, reliability and feasibility need further investigation in order to support this method for both selective and educative purposes.

In order to investigate the problems mentioned above, a study was conducted focused on video assessment of both GPs' communication with patients and medical performance. The following questions were posed: how might consultations be identified and selected for a valid assessment; how reliable is video assessment considering the number of consultations and observers; how feasible is this method concerning technical aspects, logistics, costs and acceptance by GPs and patients?

Method

Subjects and procedure

In a cross-sectional study, consultations of general practitioners were videotaped during 1 week in their own practice. All general practitioners in the south of the Netherlands were invited by letter to participate in this study; 220 (25%) were willing to do so. For budgetary reasons 135 GPs were selected randomly, 100 of them to be videotaped, the other 35 to act as peer observers to score performance. Personal and professional characteristics of the participants (see Table 1) were collected and compared to the national population of GPs to determine representativeness (*t*-test and chi-square test) and confounding.¹⁵

A video-observation system, consisting of two cameras with built-in microphone, a monitor and a recorder, was installed by trained electricians in each practice. All consultations were recorded both in the consulting room and the examination room. The GP was responsible for switching from one camera to the

Table 1 Characteristics of study population ($n = 93$) and the population of Dutch General Practitioners ($n = 6549$) in percentages, except for age

	Study population (%)	Dutch GPs (%)
gender : male	88	85
sole practice	44	51
member of College of GPs	71	69
urban area	54	59
GP trainer/postgraduate trainer	55	n.a.*
age (mean, yrs)	43.1	44.3

*n.a. = not applicable.

other when necessary, and for recording patient and consultation data, such as the patient's age and gender, (number of) complaints and duration of consultation (see Table 2), into a logbook directly after the consultation. The receptionist informed patients about the video recording and asked permission. Consent was registered by giving the patient a coloured card (green = consent, red = no consent) to be handed to the GP at the start of the consultation. Video registrations were erased if patients revoked their permission.

Evaluation of validity

To achieve a satisfactory content validity, a blueprint for the practice video assessment was designed, based on prevalence of complaints and diseases in general practice and on a nationally accepted job description.¹⁶⁻¹⁹ With this blueprint 10 specific criteria were formulated for selecting a representative sample of consultations per doctor for the final assessment (see Table 2). The first criterion, the selection of sufficient different cases, received highest priority in order to solve the case-specificity problem. Using logbook data 20 consultations per GP, 16 for direct observation and four as reserve, were selected per GP. This number was expected to offer sufficient content validity and observation time for analysis. Since it is known that scores from the first cases to be assessed are less reliable than later ones, the first five cases were not included.²⁰ Validity was evaluated by checking each GP's sample with the specific sample criteria. Furthermore, the influence of differences in sample characteristics of each GP on the performance scores was measured by a multiple regression analysis, with performance scores as dependent variable and sample characteristics as independent variables. In this

Table 2 Sample criteria for selecting videotaped consultations to be included in data analyses and percentage of samples which met the criterion fully

sample criteria	samples in which criterion was fully met (%)
1. representativeness of ≥ 8 different ICPC chapters	100.0
2. the first 5 videotaped consultations not included	94.6
3. at least 8 cases with accepted guidelines	86.2
4. representativeness of 5 ICPC chapters with highest prevalence: D, K, L, R and S	71.3
5. at least 14 cases with duration ≥ 5 and ≤ 15 minutes/case	86.2
6. age distribution: youngest patient < 18 years and oldest patient > 65 years	84.0
7. gender heterogeneity: ≥ 6 and ≤ 10 female patients	72.3
8. at least 8 initial consultations, at least 4 follow-up consultations	70.2
9. maximum 2 cases with > 2 reasons for encounter	92.6
10. maximum 2 cases with only psychosocial problems	93.6

way the contribution of GPs' personal and professional characteristics to performance scores was also analysed. Influence of the observation on GPs and patients and feasibility aspects were assessed by a questionnaire for GPs, containing 20 structured questions (five-point Likert scale). GPs were asked to score questions about recognition of their 'usual working style', after they had viewed the selected consultations.

Evaluation of reliability

Consultations were assessed for communication and medical performance by using the MAAS-Global scoring list.²¹ This instrument, containing 12 items on the quality of communication with patients and four items on the quality of medical performance, is based on literature related to effective communication between GPs and patients and on guidelines for medical performance. Items are case-independent and global, but anchored with detailed criteria. They are scored on a seven-point Likert scale (Fig. 1). An extensive list of criteria is available.²²

GENERAL COMMUNICATION SKILLS

<i>Providing information</i>	0	1	2	3	4	5	6
announcement, categorizing							
in small amounts, concrete explanation							
comprehensible language							
inquires about reaction and comprehension							

0 = absent 1 = bad 2 = insufficient 3 = doubtful 4 = sufficient 5 = good 6 = excellent

Figure 1 MAAS-Global item 8. Information

Case scores were calculated by summing the item responses on the MAAS-Global for each case expressed as a percentage score. The mean score for the 16 cases, i.e. the sample, determined the total score of the participant GP. Scores were calculated for the two components: communication and medical performance, and for the overall consultation performance (communication score + medical performance score/2). To estimate reliability, generalizability theory was used. This theory uses ANOVA procedures and variance component estimation to estimate 'generalizability coefficients'.^{23,24} For part of the dataset – 25 GPs and 15 cases – cases were scored by pairs of independent peer-observers (raters). To avoid a halo-effect, i.e. the tendency to rate a GP high (or low) in all areas being evaluated in a session if he or she scores high (or low) in one area, a maximum of six consultations per GP was assessed by the same rater. Because different raters were used for different daily practice cases, a 'raters-nested-within-cases-within-persons' design was used, followed by variance component estimation.²⁵ This design allows a concurrent estimation of the reliability of raters and cases. From the variance components generalizability coefficients were estimated; this enables projection of the reliability from the actual number of cases used to other hypothetical test conditions with different sample sizes (see Table 4).

Evaluation of feasibility

Problems experienced by GPs and receptionists with the video installation, operating of the video equipment and the informed consent procedure were assessed by the 'feasibility and influence on GPs' questionnaire. Furthermore, feasibility was measured by checking the technical quality of the video taped consultations and particularly by checking the number of re-installations and spare consultations needed to complete the sample. Costs were calculated on the basis of accounts kept by all participants.

Results**Subjects and drop-out**

Of 100 GPs seven (7%) dropped out before video observation on the grounds that 'participation is too threatening' (1%) or that they were 'too busy to participate' (6%). With regard to professional and personal characteristics the study population ($n = 93$) was representative for Dutch GPs in all respects, except for average age (mean 43.1, SD 5.5, respectively, mean 44.3 years, SD 7.4; $t = 2.03$; d.f. = 92; $P < 0.05$, Table 1).

Validity

A minimum of 40 and a maximum of 95 consultations were recorded and registered in the logbooks by the participants. The number of videotaped consultations was related to the number of (patients per) surgeries during video recording. In the selection of 16 cases it was possible to meet criterion 1: in all samples (100%) eight cases from eight different chapters of the International Classification of Primary Care (ICPC)¹⁷ were included (Table 2). Thus, at least eight of all 17 ICPC chapters were represented in each sample. A majority of samples also met the remaining criteria. In a multiple regression analysis, with GPs' characteristics as independent variables and performance scores as dependent variables, only the characteristic 'sole practice' appeared to be a significant (negative) factor of influence on communication scores ($P < 0.05$). GPs' characteristics had no predictive value for scores on medical performance ($P < 0.05$).

Multiple regression analysis indicated a significant association between the total duration of the 16 consultations of each sample (mean 156.62, SD 19.62, minimum 100, maximum 200 minutes) with GPs' scores, on both communication and medical performance ($P < 0.001$). The longer the duration of 16

consultations, i.e. the sample, the higher the total score.

In the questionnaire, 29% of physicians reported they were influenced constantly by the video observation: 12% considered the effect positive (i.e. better consultations), while 17% believed the effect to be negative. Conversely, 42% of GPs thought they were not influenced at all while 29% considered that they were only influenced during the first cases. Since the first five cases have not been included in 95% of the samples, one may conclude that the majority of GPs felt they were not influenced by the observation. Moreover, 92% of the GPs recognized their usual style of performance, when they viewed the consultations after the analysis.

Reliability

The range in scores and the standard deviation in Table 3 indicate a broad distribution of scores, on both communication and medical performance. Table 4 shows generalizability coefficients as a function of the number of cases, the corresponding estimated testing time, and number of raters. Each entry represents a reliability estimate for the given sample sizes of cases and raters. An arbitrary 'benchmark' of 0.80 is usually

Table 3 MAAS-Global scores of GPs ($n = 93$) in percentages: mean, standard deviation (SD) and range

	mean (SD)	range
communication with patients	41.7 (7.4)	25.8–63.9
medical performance	62.0 (6.9)	47.1–84.1
overall performance	51.9 (6.6)	36.5–73.3

considered as a minimal acceptable value. This means that when an infinite number of similar tests is applied randomly to the same GPs with different cases in each test and different raters for each case, the average correlation expected between the scores of these tests would be >0.80 . Using two raters, this value is found after eight cases (i.e. 96 min) on the overall score, and on both communication and medical performance after 12 cases (i.e. 144 min). For one rater this level was reached after 12 and 16 cases, respectively.

Feasibility

In 3% of the installations, the video equipment had to be re-installed due to an erroneous operation by GPs; a further 5% of selected consultations had to be replaced by 'spare' consultations on account of inadequate sound or vision quality. Thus 92% of consultations were recorded adequately in the first instance, in both consulting rooms and examination rooms. Questionnaire data (response level 91%) showed that 82% of GPs considered the informed consent procedure clear; 85% of patients gave their permission to record. A minority of GPs (10%) considered video recording a burden for patients. Completion of the logbook took 92% of GPs less than 1 minute per case. Selection of consultations from the logbook by the researcher took 1 hour per participant. Peer-observers compared the logbook data with the content of video recordings and in 5% of cases logdata appeared to be incomplete.

Table 5 provides an estimate of costs for two different sample sizes of consultations. One video set (£800, to be depreciated within 4 years) was used by 14 GPs, which reduced costs to £15 per GP. Peer-observers were paid £15 per hour and the electrician was paid £10 per hour.

Table 4 Reliability: generalizability coefficients as a function of the number of cases and raters

sample size		performance					
		overall		communication		medical	
cases	minutes	1 rater	2 raters	1 rater	2 raters	1 rater	2 raters
4	48	0.58	0.68	0.55	0.65	0.52	0.63
8	96	0.74	0.81	0.71	0.79	0.68	0.78
12	144	0.81	0.87	0.79	0.85	0.76	0.84
16*	192	0.85	0.90	0.83	0.88	0.81	0.87
20	240	0.87	0.91	0.86	0.90	0.84	0.90

*actual sample size used.

Table 5 Costs in pounds sterling (£) per General Practitioner (4 cases = 48 minutes)

	12 cases	16 cases
equipment	15*	15*
videotapes	15*	15*
installation	50*	50*
one observer	75	100
administration	45	60
review logbook	20	25
raters training	30*	30*
total	£250	£295

*stable costs.

Discussion

To our knowledge this is the first study which systematically addresses issues of validity, reliability, and feasibility of a video procedure for assessing both communicative and medical performance of general practitioners (GPs) in daily practice. A blueprint including a procedure for sampling of consultations by using selection criteria has been developed. Generalizability, i.e. reliability considering number of cases and observers, and aspects of feasibility, especially costs, were taken into account. In this way, a valid and reliable assessment procedure has been established. Results on acceptance should be interpreted with care since our study population participated voluntarily and the assessment was educative, not summative.

Concerning validity, the results show clearly that the domain of general practice was covered extensively by the procedures followed and a satisfactory content validity was obtained. Eight different ICPC chapters were represented in each sample. Concerning this classification system, it is estimated that the use of other classification systems will lead to the same conclusions. In addition, sufficient initial and follow-up consultations with both elderly patients and children were included, which support the representativeness of daily practice care. Our sample criteria, formulated for the selection of 16 cases, were well met. The lack of control over the case selection process, one of the major disadvantages of an observation in practice,²⁶ was prevented by this procedure. Strict sample uniformity does not seem desirable, since a sample has to be representative for an individual GP. It should reflect differences between GPs with regard to characteristics of their patient population, workload, working style and their perception of tasks.^{27,28} The variation in total duration of consultations of each sample may reflect these differences. The positive relationship between the total duration of the samples and GPs' scores on communicative and medical

performance is in line with previous research.^{29,30} The negative relation between working single-handedly and high score of communication with patients may be explained by the contrasting difference with more partner practices, where collaboration must rely on communication and delegation of tasks within a team. Compared with selection of cases by candidates, case selection from the logbook by an external reviewer using preset criteria prevents sampling bias. The logistics of producing 90 consultations should not be underestimated. However, provided that the video-equipment has been installed and the informed consent procedure has been handled at the desk, GPs spend on average 1 minute per case to fill in the logbook and patient care can be continued normally. Moreover, continuous videotaping regular consultations during several days increases the chance to observe what a doctor is doing in his day-to-day practice; this might be different from what a doctor is capable of doing.³¹

Finally, with regard to validity, with two cameras actual performance was followed as closely as possible. Physical examination and concurrent communication are essential aspects of performance. Therefore, validity of video assessment in daily practice increases if equipment is able to record consultations in both consulting and examination rooms. Switching cameras by the GP intrudes on normal surgery routine; on the other hand, having others to operate equipment may entail high costs.³² An automatic switching system could solve this problem, which may decrease the audience effect.

We may conclude that the validity of this video assessment of GPs in daily practice is high.

The reliability results are encouraging, particularly considering the limitations in standardization of practice assessment. This is actually comparable or even better to test conditions using highly trained standardized patients in multiple station examinations at a 'skills laboratory'.^{11,33,34} These results may be explained by a consistency of GP's usual performance during video recording, confirmed by GPs recognizing their usual working style in the consultations. In addition, peer-observers were trained in four sessions in scoring videotaped regular consultations with the explicit criteria. Adding another observer to each case substantially improves reliability, suggesting considerable 'rater bias'. Therefore, efforts to improve reliability should be focused on rater training in scoring complex daily cases and on further development of nationally accepted guidelines for both communication and medical performance.³⁵

Finally, reliability (and validity) of this video assessment procedure could be enhanced by combining video observations with an inspection of medical records.

However, clinical data are often incomplete and inconsistent, which may raise other problems.³⁶⁻³⁸

Organization of our video assessment in daily practice proved to be feasible. Although 92% of all consultations were usable for data handling in first instance, this figure could even be higher with still better instructions to operators and GPs. Handling the informed consent procedure at the counter by the receptionists seems fair for the patients. Our figures of acceptability, similar to former studies,^{39,40} support this view, taking into account that physical examination was videotaped integrally. Our detailed informed consent procedure, the use of peer observers and the transport of the tapes by registered mail or couriers may have decreased patients' fear of abuse of tapes. For ethical reasons, in video assessment one should always demand that both GP participants and peer observers handle the tapes as 'medical secrets' and to erase the video registrations after reviewing, as we did. No copies should be made. Concerning ethical aspects, these videotapes should be considered as a part of patients' medical files just as are medical journals.

Concerning costs, compared with investments of resources in observation methods in medical education⁴¹ the costs were acceptable, and could even be reduced by frequent use of the same video equipment and by integrating rater-training in regular quality improvement activities or in GP-trainer activities. Acceptance of both participant GPs and peer observers was high due to the educative impact of a video assessment method, but participation was voluntary.

We conclude that we have developed a procedure for educative assessment of GPs by means of video observation in daily practice, taking into account issues of validity, reliability, feasibility and acceptability, which enables GPs to evaluate and to improve professional quality. There is a trade-off between content validity and feasibility. Further research to increase feasibility and acceptability by reducing the number of consultations to be observed, while maintaining content validity and reliability, is needed. A decision about pass-fail levels should be made if our procedure is used in selective assessment. Especially in high stakes assessments, credibility of the method is of major importance. In our opinion credibility and acceptability increases if the assessment method measures what it is supposed to measure in a reliable way.

References

- 1 McPhee SJ, Bird JA, Fordham D, Rodnick JE, Osborn EH. Promoting cancer preventing activities by primary care physicians. *JAMA* 1991;266:393-401.

- 2 Sibley JC, Spitzer WO, Rudnick KV, et al. Quality of care appraisal in primary care: a quantitative method. *Ann Intern Med* 1975;83:46-52.
- 3 Grol R, Wensing M, Jacobs A, Baker R. *Quality assurance in general practice. The state of the art in Europe*. Utrecht: Netherlands Huisartsen Genootschap; 1993.
- 4 Davis DA, Thomson MA, Oxman AD, Haynes RB. Evidence for the effectiveness of CME. A review of 50 randomized controlled trials. *JAMA* 1992;268:1111-7.
- 5 Rethans JJ, Westin S, Hays R. Methods for quality assessment in general practice. *Fam Pract* 1996;13:468-76.
- 6 van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ Theory Pract* 1996;1:41-67.
- 7 Rethans JJ, Sturmans F, Drop MJ, Van der Vleuten CPM. Assessment of performance in actual practice of general practitioners by use of standardized patients. *Br J Gen Pract* 1991;41:97-9.
- 8 Wakefield J. Direct observation. In: Neufeld VR, Norman GR, eds. *Assessing Clinical Competence*. New York: Springer Publishing Company; 1985:51-71.
- 9 Kane MT. The assessment of professional competence. *Eval Health Prof* 1992;15:163-82.
- 10 Elstein A, Shulman LS, Sprafka SA. *Medical problem solving: an analysis of clinical reasoning*. Cambridge Massachusetts: Harvard University Press; 1978.
- 11 Newble DI, Swanson DB. Psychometric characteristics of the objective structured clinical examination. *Med Educ* 1988;22:325-34.
- 12 Bain JE, Mackay NSD. Videotaping general practice consultations (letter to the editor). *BMJ* 1993;307:504.
- 13 Campbell LM, Howie JGR, Murray TS. Use of videotaped consultations in summative assessment of trainees in general practice. *Br J Gen Pract* 1995;45:137-41.
- 14 Anonymous. *Fellowship by assessment - criteria FBA8*. Vale of Trent faculty. Royal College of General Practitioners. 1 October 1996-30 September 1997.
- 15 Caulford PG, Lamb SB, Kaigas TB, Hanna E, Norman GR, Davis DA. Physician incompetence: specific problems and predictors. *Acad Med* 1994;69(Suppl.):16-20.
- 16 Metsemakers JFM, Höppener P, Knotterus JA. The registration network family practices: a computerized health information system in the Netherlands. *Br J Gen Pract* 1992;42:102-6.
- 17 Lamberts H, Wood MICPC. *International classification of primary care*. Oxford: Oxford University Press; 1987.
- 18 Lisdonk EH, van de Bosch WJHM van den, Huygen FJA, Lagro-Jansen ALM. *Ziekten in de Huisartspraktijk*. Utrecht, the Netherlands: Bunge; 1994.
- 19 Springer MP. *Basic job description for the general practitioner*. Utrecht: Dutch National Association of General Practitioners; 1983.
- 20 Stillman PL, Regan MB, Swanson DB, Haley HLA. Sequence effect in a multiple station examination using standardized patients. In: Hart IR, Harden RM, Des Marchais J, eds. *Current developments in assessing clinical competence*. Montreal, Canada: Can Heal Publications; 1992.

- 21 Van Thiel J, Van der Vleuten CPM, Kraan H. Assessment of medical interviewing skills: generalizability of scores using successive MAAS-versions. In: Harden RM, Hart IR, Mulholland H, eds. *Approaches to the assessment of clinical competence*. Dundee: Centre for Medical Education; 1992.
- 22 van Thiel J, van Dalen J, Ram P. *MAAS-globaal Criterialijst*, internal publication. Universiteit Maastricht, the Netherlands; 1995.
- 23 Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. *The dependability of behavioral measurements: theory of generalizability for scores and profiles*. New York: Wiley; 1972.
- 24 Brennan RL. *Elements of generalizability theory*. Iowa City, Iowa: American College Testing Publications; 1983.
- 25 Suen HK. *Principles of test theories*. Hillsdale, New Jersey: Lawrence Erlbaum Associates; 1990.
- 26 Hays RB, Jones BF, Adkins PB, McKain PJ. Analysis of videotaped consultations to certify competence. *Med J Aust* 1990;152:609-11.
- 27 Mookink HGA, Schellekens CMAM, Tielens VCL. Consultduur in de huisartspraktijk. *Huisarts Wet* 1993;36:285-90.
- 28 Bensing J. *Doctor-patient communication and the quality of care* (Dissertation). Rotterdam, the Netherlands: Erasmus University; 1991.
- 29 Howie JGR, Porter AMD, Heaney DJ, Hopton JL. Long to short consultation ratio: a proxy measure of quality of care for general practice. *Br J Gen Pract* 1991;41:48-54.
- 30 Howie JGR, Heaney DJ, Maxwell M. *Measuring quality in general practice*. Occasional paper 75. London: Royal College of General Practitioners; 1997.
- 31 Rethans JJ, Leeuwen van Y, Drop R, Sturmans F, van der Vleuten C. Performance and competence: two different constructs in the assessment of quality of medical care. *Fam Pract* 1990;7:168-74.
- 32 Pringle M, Stewart-Evans C. Does awareness of being video recorded affect doctor's consultation behaviour? *Br J Gen Pract* 1990;40:45-8.
- 33 Fraser RC, McKinley RK, Mulholland H. Consultation competence in general practice: testing the reliability of the Leicester assessment package. *Br J Gen Pract* 1994;44:293-6.
- 34 Norcini J, Swanson D. Factors influencing testing time, requirements for simulation-based measurements: do simulations ever yield reliable scores? *Teaching Learning Med* 1989;1:85-91.
- 35 Rutten GEH, Thomas S. *NHG Standaarden Voor de Huisarts*. Utrecht: Bunge/NHG; 1993.
- 36 Rethans JJ, Martin E, Metsemakers J. To what extent do clinical notes by general practitioners reflect actual performance? A study using simulated patients. *Br J Gen Pract* 1994;44:153-6.
- 37 Johnson N, Mant D, Jones L, Randall T. Use of computerised general practice data for population surveillance: comparative study of influenza data. *BMJ* 1991;302:766-8.
- 38 Houghton G. General practitioner reaccreditation: use of performance indicators. *Br J Gen Pract* 1995;45:677-81.
- 39 Shafir MS. Patient consent to observation. *Can Fam Phys* 1995;41:1367-72.
- 40 Campbell LM, Sullivan F, Murray TS. Videotaping of general practice consultations: effect on patient satisfaction. *BMJ* 1995;311:236.
- 41 Reznick RK, Smee S, Baumber JS, Cohen R, Rothman A, Blackmore D, Berard M. Guidelines for estimating the real cost of an Objective Structured Clinical Examination. *Acad Med* 1993;68:513-7.

Received 12 May 1998; editorial comments to authors 4 August 1998; accepted for publication 28 August 1998