Consultation competence in general practice: testing the reliability of the Leicester assessment package

R C FRASER

R K MCKINLEY

H MULHOLLAND

SUMMARY

Background. *An acceptable assessment must be both valid and reliable; the face validity of the Leicester assessment package has already been established.*

Aim. *This study set out to test the reliability of the Leicester assessment package, and the factors influencing it, when used by multiple assessors to assess performance in general practice consultations.*

Method. *Six randomly selected course organizer assessors simultaneously used the package to conduct independent assessments of the performance of five doctors of widely varying abilities in consultation with six simulated patients. The scores allocated were subjected to generalizability analysis.*

Results. *The mean scores allocated for consultation performance of individual doctors ranged from 51% to 70%, with the lower scores being allocated to the less experienced doctors. Scores of each assessor across the cases were examined for internal consistency and five of the six assessors consistently scored the doctors with an alpha coefficient of the minimum accepted level of 0.80 or greater. The other assessor had a consistency of only 0.22. Measurements of consistency within cases between markers indicated that the first case produced unreliable results (alpha coefficient 0.25) but all other cases were scored consistently. Two independent assessors scoring eight consultations are the requisite numbers to achieve acceptable levels of reliability in a formal assessment process; seven consultations produce the minimum acceptable generalizability coefficient of 0.80 plus the first 'non-counting' consultation.*

Conclusion. *Required levels of reliability can be achieved when the package is used by multiple markers assessing the same consultations over a wide range of consultation performance. To achieve reliability only two hours of assessment time are required using the Leicester package compared with the previously regarded minimum of 32 hours. Although assessors can produce reliable scores with minimal training, intra-assessor reliability cannot be taken for granted and all assessors should be trained and calibrated before being sanctioned to conduct assessments, particularly for regulatory purposes. The Leicester assessment package has now been shown to be valid, reliable, feasible and easy to use in practice. It can, therefore, be recommended for use in both formative and summative assessment of consultation competence in general practice.*

Keywords: consultation skills; professional competence; peer review; assessment techniques.

R C Fraser, MD, FRCGP, professor of general practice and R K McKinley, MD, MRCP, MRCGP, senior lecturer in general practice, University of Leicester. H Mulholland, MA, PhD, senior lecturer in medical education, University of Dundee.

Submitted: 6 September 1993; accepted: 20 January 1994.

Introduction

IN adopting the report of its summative assessment working party in May 1993, the Joint Committee on Postgraduate Training for General Practice stated: 'A more objective process of assessment will be necessary to ensure a minimum standard of competence of all doctors entering the discipline of general practice.'[1] The committee identified an 'evaluation of clinical and consulting skills' as one of four proposed elements of assessment and regions were encouraged to experiment in developing their own assessment tools. No generally accepted system of clinical assessment currently exists.

Before any tool for the assessment of competence can be recommended for general use, whether for formative (educational) or summative (regulatory) purposes, it must first have been shown to be both valid and reliable. The face validity of the prioritized criteria in the Leicester assessment package[2] against which performance in general practice consultations can be judged, has already been established.[3] For an assessment system to be considered reliable, it must facilitate the production of comparable scores when used independently by different assessors (inter-assessor reliability). To safeguard the interests of those being assessed it is of particular importance that appropriate levels of inter-assessor reliability are achieved in regulatory assessments. In the assessment of consultation competence acceptable levels of reliability must be achieved for every consultation and the total score which is to be used as the basis for decisions in summative assessment must also be reliable across a number of different consultations.

The aim of this study was to test the reliability of the Leicester assessment package, and the factors influencing it, when used by multiple assessors to assess performance in general practice consultations.

Method

It was necessary to develop a methodology which could separate out the variance in scores specifically attributable to consulting doctors (subjects), assessors (markers) and patients (cases), and their interactions. To do this a method in which all assessors marked all doctors in all consultations was selected. The reliability of the scores produced were then tested using generalizability analysis.[4,5] This allows a determination of the required numbers of assessors and patients which would need to be used in a summative assessment for reliable scores to be produced. A full description of the educational and statistical principles underpinning the chosen methodology is given in Appendix 1.

Five doctors of varying seniority consulting with six simulated patients were observed and assessed by six assessors. The five doctors were a hospital doctor with no general practice experience, a second year vocational trainee, a third year vocational

trainee and two principals in general practice; three were men and two were women. They were chosen to test the reliability of the Leicester assessment package across a broad range of clinical competence.

The six assessors were course organizers chosen at random by an independent statistician from the 62 participants in the earlier validity study[3] who had indicated a willingness to participate in later studies. Four were from England, with one each from Scotland and Wales. Their task was to use the package to assess the consultation performance of the five doctors carrying out a series of consultations with the same six simulated patients. Each assessor thus observed a total of 30 consultations, over a period of three days.

The simulated patient scenarios spanned a range of clinical challenges which included acute and chronic conditions with a mixture of physical, psychological and social aspects:

- A 26-year-old woman presenting with acute backache.
- A 30-year-old woman presenting with tiredness and lack of energy.
- A 68-year-old man, a non-insulin dependent diabetic, presenting with pain and tingling in the legs.
- A 35-year-old woman presenting with a sore throat and a demand for antibiotics.
- A 41-year-old woman presenting with palpitations.
- A 56-year-old man presenting with chest pain.

The scenarios were devised in conjunction with colleagues in the departments of general practice at Leicester University and the Free University of Amsterdam, Netherlands. The simulated patients were trained and medical records created for all scenarios by colleagues in the Leicester department. All simulated patients were taught to portray the appropriate physical signs for their role.

The doctors were aware that the patients were simulated and that they had consented to appropriate physical examination on camera but they had no prior knowledge of the scenarios. All the doctors carried out a consultation with a different simulated patient a week before the formal assessment to familiarize them with the procedure.

Seven weeks before the assessments took place all assessors were provided with a complete copy of the Leicester assessment package and a 10-minute videotape as a 'user friendly' means of introducing it. Assessors were asked to familiarize themselves with its contents and to practise using the package with trainees and partners. On arrival in Leicester, a short briefing session was held to familiarize the assessors with the study protocol and their role in the study. Any difficulties encountered in using the package were also discussed.

The doctors were instructed: to consult as they would with real patients (no time limit was imposed); that each assessment would start when they were given the patient record; that they should summon the patients when they were ready (thus no cueing took place regarding consulting the patients' notes); and that they would be required to make a record of the consultation in the notes (a separate continuation card was provided for each doctor for this purpose).

All consultations were carried out and videotaped in a mock up consulting room in a studio in the audio-visual services department at Leicester University.

The assessors were positioned so that they could directly observe and hear all consultation events. At the end of each consultation photocopies of the notes made by the doctor were given to the assessors, who independently awarded marks to reflect their view of the performance of the doctor in each of the seven categories of consultation competence in the package and for overall consultation performance. At the end of each doctor's series of six consultations the assessors also allocated marks for the seven categories as well as for overall performance across all six consultations. All patients were presented to the doctors in the same order (as given above). The assessors were not made aware of the level of experience of the doctors.

Results

Table 1 shows the scores allocated by the assessors to reflect their judgement of the doctors' overall performances across all six consultations and the means of these scores. There was a wide range of mean scores from 51.3% to 70.2% with the lower scores being allocated to the less experienced doctors.

The alpha coefficient is a measure of internal consistency achieved in any assessment process. To achieve acceptable levels of internal consistency the alpha coefficient must be 0.80 or greater. In this instance alpha coefficients were calculated on the basis of overall scores for each case (consultation) from each assessor to determine the internal consistency of the scores allocated by markers, that is both intra- and inter-assessor reliability. It was found that five of the six assessors scored consistently with an alpha coefficient of 0.80 or above, that is scores were consistent across the six cases. The other assessor had an alpha coefficient of only 0.22. When consistency within cases between markers was examined it was found that the first case produced unreliable results with an alpha coefficient of 0.25, but that all other cases were scored consistently.

Generalizability analysis was applied to the data to predict the numbers of cases required to achieve acceptable levels of reliability using two independent assessors in a formal assessment process. In studies of clinical competence it is the accepted convention that a value of 0.80 or greater for the generalizability coefficient indicates an acceptable degree of reliability. For the Leicester assessment package two independent assessors scoring seven consultations would be adequate (Table 2). It should be noted that doubling the number of cases from seven to 14 increases the generalizability coefficient by only 0.05. A reduction in the number of cases to six would require the use of three assessors to reach an acceptable level of reliability (0.81).

Discussion

It has been demonstrated that required levels of reliability can be achieved when the Leicester assessment package is used by multiple markers in assessing the same consultations, that is the package produces inter-assessor reliability. Moreover, reliability is achieved when the package is used to assess the performance of doctors of widely varying abilities — an essential requirement if an assessment tool is to be used for regulatory purposes. There appear to be three reasons why this reliability has been achieved. First, in using the Leicester package all assessors are obliged to judge consultation performance against the same explicit and

Table 1. Percentage scores allocated to doctors by assessors to reflect their performance across all six consultations and the mean score for each doctor.

	% scores allocated by assessors						
Doctors	A	B	C	D	E	F	Mean
1	*56*	*54*	*54*	*51*	*48*	*45*	*51.3*
2	*57*	*59*	*57*	*56*	*57*	*53*	*56.5*
3	*75*	*72*	*64*	*65*	*57*	*65*	*66.3*
4	*70*	*74*	*63*	*67*	*65*	*63*	*67.0*
5	*76*	*70*	*58*	*69*	*71*	*77*	*70.2*

Table 2. Predicted generalizability coefficients with two independent assessors and varying numbers of cases.

Number of cases	Generalizability coefficient
2	0.68
3	0.72
4	0.75
5	0.76
6	0.79
7	0.80
8	0.82
10	0.83
12	0.84
14	0.85

validated criteria. This minimizes opportunities for assessors to be influenced in their judgements of performance by non-valid and/or idiosyncratic criteria.

Secondly, having reached the stage of considering how to convert their judgement of performance into marks, individual assessors using the package are required to refer to a set of guidelines for the allocation of such marks. These consist of a range of descriptions of performance, linked to scales of marks. This facilitates a more accurate 'calibration' of performance by markers and it is also likely to minimize the gap between 'hawks' and 'doves'.

Thirdly, although individual assessors are obliged to use a systematic and common approach to assessment when using the package, each assessor still needs to use his or her own clinical expertise in making his or her assessments, since the criteria are not case specific. The package thus allows assessors flexibility to adjust their construct of competence to match the particular clinical challenges posed in each consultation and to relate this to the particular consulting styles of the doctors who are being assessed. This method of scoring, known as limen referencing,[7] avoids the more rigid and mechanistic consequences of 'assessment by checklist', thus enabling new and completely unpredictable consultations to be more sensitively — and therefore more reliably — assessed. This also suggests that the Leicester package is likely to result in reliable scores in assessments involving real patients as well as simulations.

The results from this study have also shown that scores from the first case to be assessed were less reliable than those from later cases. This confirms the findings of Stillman and colleagues.[8] It is suggested that this may be due to subjects and/or markers not being at ease during the first consultation of the assessment process or to the nature of the case. Further research is required to determine the exact relationship between scores, case mix and order of consultation. Nevertheless, frequent use of the Leicester package in assessing doctors in consultation with real patients has repeatedly produced high levels of inter-assessor reliability with a wide case mix (unpublished results). On present evidence, however, scores from the first consultation of any regulatory assessment should be discarded in the interests of candidate equity.

Nevertheless, with simulated patients, using the Leicester package would require considerably less effort to achieve reliable results than has previously been reported.[5,6,9] In most studies, two assessors are accepted as the optimum number needed to be involved in assessment for regulatory purposes on the grounds of equity and feasibility. With the Leicester package, only eight patients (seven required for reliability plus the first 'non-counting' case), requiring an anticipated maximum of two hours of assessment time, are required compared with the 32 hours previously considered to be the minimum.[6] In this study, it was possible to create an appropriate range of clinical challenges to present to the consulting doctors. If real patients are to be used in any regulatory assessment process, that is where the clinical mix cannot be controlled, the numbers of patients required to produce both reliable and valid assessments of competence remain to be determined. This would be true whatever assessment instrument were used.

It is noteworthy that the assessors involved in this study had undergone minimal training in the use of the package. Nevertheless, acceptable levels of inter-assessor and intra-assessor reliability were achieved with the exception of one assessor. This indicates that intra-assessor reliability cannot be taken for granted. With more extensive training in the use of the package, however, it has been shown that assessors become better 'calibrated' without this affecting reliability.[10] It would seem essential, therefore, that all assessors should be trained and calibrated before being sanctioned to assess real candidates, particularly for regulatory purposes.

The Leicester assessment package has now been shown to be valid,[3] reliable, feasible and easy to use in practice.[10] The package can, therefore, be recommended for use in both formative and summative assessment of 'clinical and consulting skills'[1] in the setting of general practice.

Appendix 1. Educational and statistical principles underpinning the chosen methodology.

The complexity of clinical assessment procedures creates problems in both testing and determining their reliability. With a reliable assessment instrument, differences in scores should be caused by true differences between the subjects. This is only one of three sources of variance, however, in the assessment of clinical performance, the others being the influence of the markers and case specificity.

Markers can behave as 'hawks' (who mark low) and 'doves' (who mark high). Individual markers may also differ in their perceptions of the importance of various aspects of clinical competence. As a consequence, different markers reviewing the same performance may allocate marks differently, rendering the assessment process unreliable.

It has also been consistently demonstrated that subjects' ability to cope with one case does not necessarily correlate highly with their ability to cope with another.[6] To remedy the influence of case specificity on assessment scores, subjects need to be judged over a series of consultations, as this allows differences in scores between individual cases to be averaged out.

Figure 1a illustrates the sources of variance arising from these three main effects: subject (S), marker (M) and case (C). There are also three two-way interactions, S x M, C x S and C x M, and one three-way interactions, C x S x M, all of which may cause distortion of true scores. This illustrates the complexity of analysing the scores from assessments and can only be overcome by using the statistical technique of generalizability theory.[4,5] This enables the measurement of the size of each of three effects in a given assessment procedure and allowance to be made for them by predicting necessary changes in the design of the assessment (that is, determining the required numbers of markers and cases) which will result in truly reliable scores.

The assumptions on which generalizability theory rest are that the markers, cases and subjects are random samples drawn from the 'universes' of all possible markers, cases and subjects. A universe of markers might be all vocational training course organizers. A universe of cases might be all the possible patients presenting to a general practitioner and a universe of subjects might be all trainees at the end of their vocational training. Markers, cases and subjects become facets of the design. Figure 1a represents an ideal situation in which all the markers score all the candidates on all the cases. This is called a crossed design with which it is relatively simple to calculate the various contributions to variance.

This ideal situation is, of course, seldom possible in real assessments since variance owing to marker or case differences is usually 'nested' within the subject variance. Consequently the contribution to variance of each facet becomes difficult, or even impossible, to quantify. Figure 1b shows the components of variance when one assessor marks some of the subjects on one case, a different assessor marks the other subjects on the same case and different assessors mark other cases. It is not then possible to separate out the contribution to variance of the marker from that of the case, since each marks only one, or from the true differences between subjects since no one marks all of the subjects.

Difficulties arise in the assessment of a doctor's performance in real life

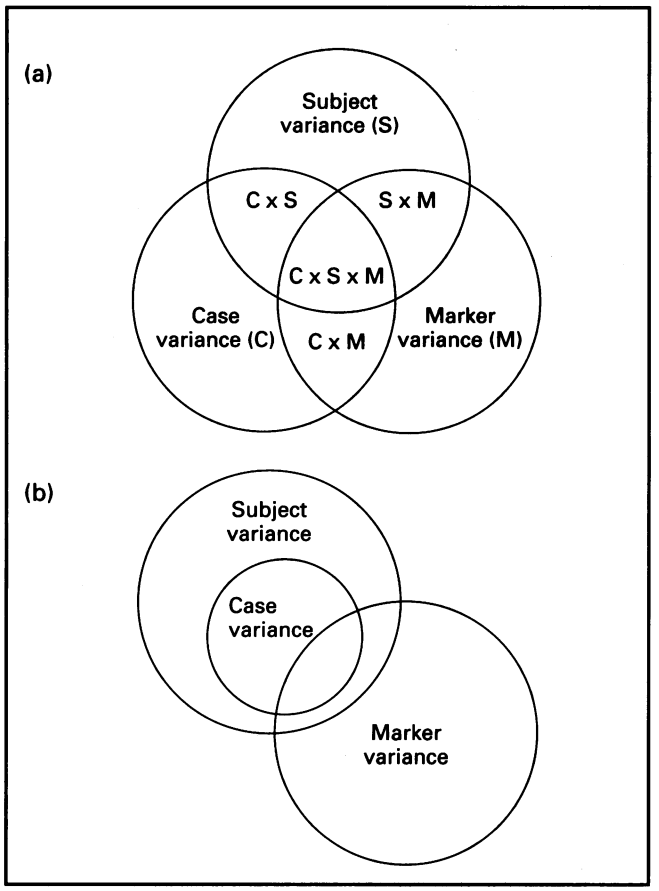

Figure 1. *(a) Crossed assessment design. (b) Nested assessment design.*

consultations because, unlike in a formal examination, no two subjects can see the same cases. The variance caused by case differences is therefore nested within the subject variance. Generalizability analysis can be carried out on such nested designs but the predicted components of variance for changes in the design cannot be so securely established as in a crossed design. In order to examine the effects of marker and case variance, it is necessary to create experimental conditions in which all subjects manage the same cases and are assessed by the same markers.

For the purposes of designing reliable assessments it is not necessary to use large numbers of subjects. Since the decisions needed are about the numbers of markers and cases, it is these factors rather than the subjects which are the focus of the analysis.

References

1. Joint Committee on Postgraduate Training for General Practice. *Report of summative assessment working party.* London: JCPTGP, 1993.
2. Fraser RC, McKinley RK, Mulholland H. Assessment of consultation competence in general practice: the Leicester assessment package. In: Harden RM, Hart IR, Mulholland H (eds). *Approaches to the assessment of clinical competence.* Part 1. Dundee: Centre for Medical Education, 1992.
3. Fraser RC, McKinley RK, Mulholland H. Consultation competence in general practice: establishing the face validity of prioritized criteria in the Leicester assessment package. *Br J Gen Pract* 1994; **44:** 109-113.
4. Brennan RL. *Elements of generalisability theory.* Iowa City, IA: American College Testing Program, 1983.
5. Colliver JA, Verhulst SJ, Williams RG, Norcini JJ. Reliability of performance on standardised patient cases: a comparison of consistency measures based on generalizability theory. *Teaching Learning Med* 1989; **1:** 31-37.
6. Van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardised patients: state of the art. *Teaching Learning Med* 1990; **2:** 58-76.
7. Black HD, Dockrell WB. *New developments in educational assessment.* Edinburgh: Scottish Council for Research in Education, 1988.
8. 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.
9. Swanson DB, Norcini JJ. Factors influencing reproducibility of tests using standardised patients. *Teaching Learning Med* 1989; **1:** 158-166.
10. Preston-Whyte ME, Fraser RC, McKinley RK. Teaching and assessment in the consultation: a workshop for general practice clinical teachers. *Med Teach* 1993; **15:** 205-210.

Acknowledgements

It is our pleasure to acknowledge the assistance provided by the following: the six course organizers who acted as assessors; Drs Elan Preston-Whyte, Paul Lazarus and Angela Lennox for training the simulated patients, creating patient records and contributing to the clinical scenarios; Dr Ale Gercama and Mr Gerrit van Staveren for contributing to the development of the clinical scenarios; and Dr Carol Jagger for drawing the sample. The study was supported by grants from the Department of Health and the Scientific Foundation Board of the Royal College of General Practitioners.

Address for correspondence

Professor R C Fraser, Department of General Practice, University of Leicester, Leicester General Hospital, Gwendolen Road, Leicester LE5 4PW.