International Journal for Quality in Health Care 2003; **Volume 15, Number 5**: pp. 407–412 10.1093/intqhc/mzg056

A reliability study of an instrument for measuring general practitioner consultation skills: the LIV-MAAS scale

IAN ENZER[1], JUDE ROBINSON[2], MAGGIE PEARSON[2], STUART BARTON[1] AND TOM WALLEY[1]

[1]Prescribing Research Group, Department of Pharmacology and Therapeutics, [2]Health and Community Care Research Unit, University of Liverpool, UK

Abstract

Objective. To evaluate the reliability of a new tool, the LIV-MAAS, in assessing consultation competence in UK general practice.

Design. These were pilot studies, with small numbers of participants. Videoed general practitioner (GP) consultations were analysed by trained lay and professional raters, using the LIV-MAAS. The inter-rater reliabilities were assessed. Four videos were assessed by five raters in a pilot study. After this, 71 consultations from eight doctors were assessed by sets of three raters.

Main measures. Inter-rater reliabilities and inter-consultation reliabilities.

Results. For the pilot study, the estimated inter-rater reliability ranged from 0.69 (one rater) to 0.91 (five raters). For the main study, the estimated inter-rater reliability for the LIV-MAAS checklist using two raters was 0.71, and using three raters it was 0.78. Mean differences in reliability within each series of nine consultations were 0.20 (three raters) and 0.42 (two raters).

Conclusions. As a measure of 'consultation competence', administered by trained raters (medical or lay) to real GP consultations, the LIV-MAAS instrument shows adequate reliability and stability but would benefit from considerable shortening. Further development of the LIV-MAAS and testing with larger samples are required.

Keywords: consultation competence, general practitioner assessment, inter-rater reliability, LIV-MAAS, MAAS-GP

Interviewing and consultation skills are key in the clinical practice of medicine [1]. In recent years, there have been attempts to make use of psychometric techniques and the development of specific instruments to measure the attributes of the consultation, both in the UK and internationally [2–7]. The expression 'consultation competence' is used in the UK literature to describe the range of skills a medical practitioner needs in a meeting with a patient [8–11].

Medical interviewing involves content and process [1,4]. The process requires communication skills to promote information flow and interpersonal skills to establish rapport with the patient. A recent 'functional' approach has three aspects: to collect information, to respond to the patient's emotions, and to educate and influence behaviour and implement treatment plans [3,12].

As part of a wider study of general practitioner (GP, family practitioner) consultation skills, we wished to use a reliable instrument to measure competence in the process of consultation. A literature search found one systematic review of such instruments [5], which recommended two 'as best fitting the criteria of reliability, validity and practicality': the Arizona Clinical Interview Rating Scale [7] (developed in the USA) and the MAAS-GP [4] (developed in The Netherlands, under a medical system much closer to that of the UK). Other tools also exist, including the recent UK Leicester Assessment Package (LAP) [8], but neither the MAAS-GP nor the LAP had been adequately tested with samples of real patients; for instance the MAAS-GP [4] was tested with simulated patients and medical student raters.

The MAAS-GP is based on the presence or absence of specified behaviours. A technical review identified several problems: MAAS-GP used an old scaling model with a very small group for item analysis, had a mixed item format, had some very low subscale reliabilities, and had limited norm data. Despite these shortcomings, the 68-item MAAS-GP has been claimed to best fit the criteria of reliability, validity, and practicality for evaluation instruments for medical interviewing skills [4,5]. We therefore chose to develop it with an additional 27 items added to include a patient perspective and UK focus. The new 95-item scale, the LIV-MAAS, is divided into six subscales and the full instrument can be reviewed on a website [13]. The content validity of this new tool has been demonstrated for the UK [14]. Like the MAAS-GP scale, it depends on a rater recording the presence or absence of

Address reprint requests to Professor T. Walley, Prescribing Research Group, Department of Pharmacology and Therapeutics, University of Liverpool, L69 3GF, UK. E-mail: twalley@liv.ac.uk

a behaviour. Despite this apparently simple dichotomy, there are few absolute measures, but instead an element of judgement on the part of the rater using the scale. Therefore, before it can be used as a research instrument, we needed to evaluate its reliability when used in videoed real GP–patient consultations and with lay or professional raters.

Methods

Design

Routine consultations between GPs and patients were video-recorded, with the consent of both and with ethics committee approval. The videos from each GP were then analysed using the LIV-MAAS by a team of trained raters, each working independently. The scores by each rater for each subscale were then compared. There were two studies, both of which could be regarded as pilot studies, but for ease of description we refer to the first as the 'pilot' and the second as the 'main' study. In the pilot study, five trained raters scored the same four consultations from two GPs independently (videos A1–A4), and in the second or main study, conducted some months later, a formal 'field' test of how the LIV-MAAS might actually be used was conducted, where a team of three raters independently scored nine videos from each of eight GPs (B1–B71).

Sample

As these were pilot studies, sample size was arbitrary. The four videos rated by all raters were not collected specifically for this project, but came from two GPs working in an urban practice, and were selected purposively from a wider bank of video consultations used for training medical students so as to cover a range of different types of consultation and conditions.

The videos for the field test were collected specifically for this study. The eight GPs (aged 35–60 years) were drawn from a pool of 15 in the north-west of England who had previously volunteered for participation in research, and were purposively selected so as to be broadly representative of UK general practice. Two were GP trainers, four were female, three described themselves as 'Asian', and five described themselves as 'white'. They came from single-handed and group practices, and from inner city, suburban, and rural areas.

Sequential patients attending each GP for routine consultations over 1 day were invited to participate until nine interviews had been recorded. The patients (aged 18 to >75 years) were 56% female and described themselves as: 'white' (90%), employed or in full-time education (41%), retired (24%), unemployed (13%), or 'other' (22%). Where there was more than one medical issue raised in a consultation, only the first rated was used in this analysis.

The instrument

The LIV-MAAS instrument consists of 95 items in six subscales [13]. The items are defined as explicit instructions, addressing the doctor's behaviour during the consultation and designed to reduce inter-assessor reliability. For instance, under 'exploration' (of the reason for the consultation), the topics are shown in Figure 1 and the doctor's display of the expected behaviour is recorded as either present or absent.

Most items have dichotomous answers, scoring behaviour as present (1) or absent (0). For the two trichotomously scored scales (communication skills and interpersonal skills), dichotomization made little difference (Spearman and Pearsonian correlations >0.94), so for ease of interpretation

	Present	Absent
1. Ask for the reason for encounter.		
2. Explores the emotional impact of the complaint /problem.		
3. Asks the patient to clarify why (s)he is presenting at this particular moment.		
4. Asks the patient to give his/her opinion on the cause of the problem.		
5. Asks how the complaint or problem is discussed within the family or primary group.		
6. Asks the patient of state what help (s)he desires.		
7. Asks how the patient has tried to solve the problem by him/herself.		
8. Explores the influence of the complaint on daily life		

Figure 1 An example of one scale on the LIV-MAAS (subscale 1 exploration of reasons for encounter).

Table 1 Pilot study: inter-rater reliabilities for all possible independent combinations of two to five raters

Number of raters	Video A1	Video A2	Video A3	Video A4	Mean
Single (95% CI)	0.76 (0.70–0.82)	0.73 (0.65–0.79)	0.80 (0.75–0.85)	0.47 (0.57–0.87)	0.69
Two (10 possible sets) (range)	0.86 (0.80–0.97)	0.84 (0.81–0.90)	0.90 (0.86–0.92)	0.63 (0.44–0.83)	0.81
Three (10 possible sets) (range)	0.90 (0.87–0.95)	0.89 (0.87–0.90)	0.93 (0.92–0.94)	0.72 (0.68–0.80)	0.86
Four (five possible sets) (range)	0.93 (0.91–0.94)	0.91 (0.91–0.92)	0.94 (all sets)	0.78 (0.75–0.80)	0.90
Five (95% CI)	0.94 (0.92–0.96)	0.93 (0.90–0.95)	0.95 (0.94–0.97)	0.82 (0.75–0.87)	0.91

CI, confidence interval.
Point reliabilities are all statistically significant ($P < 0.001$).

the middle and most positive responses were merged so that the possible range of score for the whole scale was 0–95.

Raters, training, and scoring procedures

The raters worked independently in assessing each video, but the same set of three raters assessed all video consultations for any given GP. Each set included one GP, and was drawn from a panel of two GPs and three lay observers.

The training of the raters involved working through item definitions in a handbook, with a group discussion on the goal of each item and the concepts. Each rater then reviewed and scored a standard series of videoed consultations against the LIV-MAAS checklist in a one-to-one tutorial with an investigator (S.B.). This took ~5 hours in total.

Computations

Inter-rater reliability was defined as the agreement of the raters on items and computed as an intra-class correlation coefficient from the independent raters' scores. Reliability is expressed as a decimal value between 0.00 and 1.00, with higher values indicating greater reliability. A desirable minimum of 0.7 is suggested [15].

Cohen's 'kappa' calculation checks agreement for categorical variables adjusted for chance [16,17], but there are problems concerning accuracy in interpretation (see [17], p. 347). Here several scales scored very low, no decision was made on a pass/no pass score, and few comparisons were available; hence 'kappa' was not used. A 'Generalizability Theory' approach exists for treating a judge as 'a representative of other potential judges' (see [18], p. 279). However, this is a pilot study with small numbers and lack of balance, and with 'nesting' of an individual GP's nine videoed consultations over the raters. Attention was therefore confined to reviewing the basic level of reliability and the effects on this of the number of raters. This was achieved in the first study stage by computing and reviewing all possible rater sets. As the level and pattern of inter-rater reliability appeared reasonable, the computations proceeded for the main independent three-rater scores on 71 consultations.

Inter-rater agreement was estimated by the SPSS® 10.0 Reliability Analysis intra-class correlation program [19]. For the pilot study, reliabilities were computed for all possible sets of raters. For the full set of five raters, the 95% confidence intervals could be computed by SPSS. For smaller sets of raters, only the range of point estimates could be given. For the main study, point estimates were generated directly by the SPSS program for the full three-rater panel. For this sample, two-rater reliabilities were also computed to check the degradation in reliability by a reduced panel. A by-product was an estimated reliability for a hypothetical single rater, which can be interpreted as a test-retest reliability. We also measured inter-consultation reliability, i.e. the consistency of a single GP's performance across all of his or her nine consultations.

Results

Pilot study

Table 1 summarizes the reliabilities for the 95-item scale for all possible combinations of two to five raters in the pilot study. Some raters failed to score all items. The mean scores (and ranges) for A1 to A3 were, respectively, 42 (36–56), 57 (49–70), and 51 (38–60), while the mean for video A4 was 22 (14–30). The average reliabilities for two raters on videos A1 to A4 were 0.86, 0.84, 0.90, and 0.63, respectively, and the average differences between pairs of raters were 0.07, 0.02 and 0.02 and 0.14. The estimated reliabilities increase from 0.69 for one rater to 0.91 for five, with decreasing increments for more raters.

Main study

Seventy-one consultations were analysed in total, and one recording was accidentally lost. The means and ranges of scores for each subscale and the total score are shown in Table 2. The inter-rater reliability results for the 95-item full scale, for the embedded 68 items from the original MAAS-GP scale, and for the additional 27 UK patient perspective items are shown in Table 3. The columns compare different numbers of raters.

For three raters, the average reliability of the 95-item scale is 0.76. Because of the different numbers of items, the value for the UK set of 27 items cannot be compared directly with that for the original 68-item MAAS-GP, but an estimate can

Table 2 Mean scores of LIV-MAAS and its subscales (and ranges) for the main sample

Subscale	No. of items (No. of additional UK set of items)	Mean score (range) $n = 71$
Exploring	8	0.9 (0–8)
History taking	23	2.6 (0–12)
Attitudes	8 (all UK)	1.0 (0–7)
Presenting solutions	17 (5 UK)	3.4 (0–12)
Structuring interview	19 (11 UK)	6.8 (0–14)
Interpersonal skills	12 (2 UK)	9.1 (0–12)
Communicative skills	8 (1 UK)	6.0 (0–8)
Totals (maximum possible score 95)	95 (27 UK)	29.9 (SEM 1.03)

SEM, standard error of the mean.

be made using the 'Spearman-Brown Prophecy' formula [18]. Standardized at 27 items, the original MAAS-GP items would have a reliability of 0.56 compared with 0.80 for the UK items. With 68 items, a scale using items similar to the UK items would have a reliability of 0.91 against the original MAAS-GP items of 0.76. The added UK items therefore have better reliability than the original MAAS-GP.

Inter-consultation variability within each set of one GP's nine videos marked by the same raters was high. Three raters gave a more stable assessment of reliability, as would be expected. Taking first the full 95-item scale, the average range in reliabilities was 0.42 (0.23–0.62) for two raters and 0.20 (0.09–0.30) for three raters. In the original 68-item scale, the average range was 0.55 (0.34–0.86) for two raters and 0.24 (0.16–0.40) for three. In the 27 additional items, the average range was 0.66 (0.36–0.97) for two raters and 0.35 (0.14–0.59) for three.

Discussion

In this study, inter-rater reliabilities using the LIV-MAAS were similar across raters, were reasonably consistent between each series of nine videoed consultations, and were consistent across GPs. In the pilot study, the estimated reliabilities increase from 0.69 for one rater to 0.91 for five. These compare well with the 'rule of thumb' level of 0.7 [15], while the figure for five raters perhaps approaches a ceiling value. Video A4 was clearly an outlier in the results: it related to a difficult psychomedical problem and may reflect a weakness in the ability of the scale to deal with such complex consultations. In the light of these broadly satisfactory results, we proceeded to a larger 'field' study. This confirmed the earlier results, although there was a slight decline of reliability scores from the earlier to the later studies (e.g. 0.81 for two raters down to 0.71), which may represent a decay in skills from the time of the original training.

These reliabilities are consistent with other instruments for two or three raters. Earlier reliability reports used different numbers of raters, formats, scoring systems, and coverage, and different calculation methods [2,3,9]. This may account for reliability estimates for pairs of raters ranging from 0.33 upwards, and for three raters up to 0.87. The LAP for students used artificial patients, and specimen problems, with two medical raters, gave a 0.82 predicted generalizability coefficient for eight consultations. For videoed real consultations, the LIV-MAAS 0.71 two-rater reliability compares well with this.

The inter-consultation variability was high and this stresses the need to evaluate a sufficient number of consultations by any one GP by a sufficient number of raters if a reliable assessment of that GP's performance is to be made. However, a preliminary analysis of a larger data set analysed using the LIV-MAAS suggests that the LIV-MAAS does discriminate between the performance of different GPs, using nine consultations per GP and three raters.

A criticism of current revalidation processes in the UK is that they are driven by doctors and allow only little lay involvement.

Table 3 Reliabilities for panels of single, two or three independent raters in the main sample for LIV-MAAS scale, MAAS-GP, and the added UK set of 27 items

	Single rater	Two raters	Three raters
LIV-MAAS (95 item)			
Averaged reliability[1] (ranges of mean correlations for each GP)	0.56 (NA)	0.71 (0.61–0.78)	0.78 (0.70–0.85)
Original MAAS (68 item)			
Averaged reliability[1] (ranges of mean correlations for each GP)	0.53 (NA)	0.67 (0.61–0.73)	0.76 (0.70–0.81)
UK set (27 item)			
Averaged reliability[1] (ranges of mean correlations for each GP)	0.61 (NA)	0.74 (0.58–0.75)	0.80 (0.68–0.89)

NA, not applicable; GP, general practitioner.
[1] Point reliabilities are all statistically significant ($P < 0.001$).

Consultation skills are a key part of revalidation for GPs: an important result here, therefore, is that the LIV-MAAS scale performed quite well when mixed lay and medical panels of assessors were used. This may open possibilities for its use in professional and lay evaluation of consultation skills.

Study weaknesses

The study has a number of weaknesses: most importantly we used small samples, both of GPs and of raters, and for a definitive evaluation far more of each would be required. We therefore regard both studies here as 'pilots' only. Furthermore, these small numbers lead to a design weakness, in that repeated consultations from each GP were evaluated by the same set of raters and this may have introduced problems of clustering, or a halo effect arising from an undefined rater–GP interaction. These may lead to inappropriately high reliabilities. It would have been better to have had many more GPs and raters, thereby allowing us to spread all of a GP's consultations across more raters.

In addition, the participating GPs were unrepresentative simply by their volunteering to participate in research, but preliminary work (unpublished) indicates that the instrument was able to discriminate between them, and would presumably be effective in GP groups less restricted in range. There was no analysis of consultations by patient characteristics or diagnosis: the consultation series for each GP were intended to be representative, but may have concealed systematic differences.

Finally, the LIV-MAAS is still a crude and underdeveloped instrument: it measures a process rather than actual communication, and not all of the process may be appropriate in any one situation. For instance, many items seem to contribute little in the UK context (see the low mean score in Table 1 of 29.9 points out of a possible 95). The very low scores for the 'exploring' and 'history taking' scales, where the doctor identifies himself and gathers background about the patient, show that this section may not be entirely relevant for British general practices, where patients are registered and known to their doctor over many years. This illustrates some of the difficulties in trying to apply a specific instrument to a more general real world health care setting: an instrument developed to assess the process of communication within the artificial situation of medical student consultations or consultations with actors as patients might not be a good indicator of true communication, or even of proper process, in real medical consultations. However, such instruments, even if they are derived in other countries and medical cultures, may form the basis of valid and reliable tools, as in the case presented here.

The instrument is also currently unsuitable for routine use since the average time for rating by each reviewer in the main study was ~30 minutes, i.e. three to four times longer than the consultation itself. This limits the practicality of the instrument and its applicability. We believe that it could be refined into a 40-item, four-subscale set by eliminating the individual items that are less discriminating between doctors; this might still usefully measure three postulated functions of the medical interview (diagnosis, meeting patients concerns, and education/treatment [11] plus the patient perspective), and still be adequately reliable. We continue to work to develop this.

Conclusions

When administered by trained raters (medical or lay) to real GP consultations, the LIV-MAAS provides an objective direct assessment of one aspect of consultation competence, and meets the criteria of validity [13] and reliability. However, it needs further evaluation in a larger sample of GPs and assessors. We also believe that its use on videoed consultations is acceptable to doctors and patients, and is feasible for research. Its potential as an educational instrument is untested. The LIV-MAAS is lengthy, and a shortened version might be suitable for assessment of consultation competence in revalidation of GPs.

Acknowledgements

We would like to thank the patients, raters, and GPs who participated in this study. The UK Department of Health funded this work as part of the Prescribing Research Initiative.

References

1. Swanson DB, Mayewsky RJ, Norsen L, Baran G, Mushlin AI. A psychometric study of measures of medical interviewing skills. In Proceedings of the 20th Annual Conference on Research in Medical Education. Washington DC: Association of American Medical Colleges, 1981, pp. 3–8.
2. Hess JW. Direct observation as a means of teaching and evaluating clinical skills. *J Med Edu* 1969; **44:** 934–938.
3. Hinz CF. A comparison of methods for evaluating medical student skill in relation to patients. *J Med Edu* 1966; **41:** 150–161.
4. Kraan HF, Crijnen AAF. *The Maastricht History-Taking and Advice Checklist: Studies of Instrumental Utility*. Amsterdam: Lundbeck, 1987.
5. Kraan HF, Crijnen AAM, van der Vleuten CPM, Imbos T. Evaluation instruments for medical interviewing skills. In: Lazare A, ed. *The Medical Interview: Clinical Care, Education and Research*, Chapter 39. New York: Springer, 1995.
6. McKinley RK, Fraser RC, van der Vleuten C, Hastings AM. Formative assessment of the consultation performance of medical students in the setting of general practice using a modified version of the Leicester assessment package. *Med Edu* 2000; **34:** 573–579.
7. Stillman PL, Brown DR, Sabers DL, Redfield DL. Construct validation of the Arizona Clinical Interview Rating Scale. *Educ Psychol Meas* 1977; **37:** 1031–1056.
8. 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.

9. 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:** 293–296.

10. McKinley RK, Fraser RC, Baker R. Model for directly assessing and improving clinical competence and performance in revalidation of clinicians. *Br Med J* 2001; **322:** 712–715.

11. Lazare A, Putnam SM, Lipkin M. Three functions of the medical interview. In: Lipkin M, Putnam SM, Lazare A, eds. *The Medical Interview: Clinical Care, Education and Research*, Chapter 1. New York: Springer, 1995.

12. Cohen-Cole SA. *The Medical Interview: The Three-Function Approach*. St Louis, MO: Mosby Year Book, 1991.

13. Robinson J, Walley T, Pearson M, Taylor D, Barton S. Measuring consultation skills in primary care in England: evaluation and development of content of the MAAS scale. *Br J Gen Pract* 2002; **52:** 889–894.

14. Prescribing Research Group (Liverpool, UK). *LIV-MAAS*. http://www.liv.ac.uk/prg/publications1.htm, 2002. (Accessed 5 February 2003.)

15. Kline P. *The Handbook of Psychological Testing*. London: Routledge, 1993.

16. Armitage P, Berry G. *Statistical Methods in Medical Research*, 3rd edition. Oxford: Blackwell Science Ltd, 1994.

17. Kelsey JL, Whittemore AW, Evans A, Thompson D. *Methods in Observational Epidemiology*, 2nd edition. New York: Oxford University Press, 1996.

18. Nunnally JM, Bernstein IH. *Psychometric Theory*. New York: McGraw-Hill Inc., 1994.

19. SPSS 10. Chicago, IL: SPSS, Inc.

Accepted for publication 14 April 2003