BMJ Open

Assessing communication quality of consultations in primary care: initial reliability of the Global Consultation Rating Scale, based on the Calgary-Cambridge Guide to the Medical Interview

Jenni Burt,[1] Gary Abel,[1] Natasha Elmore,[1] John Campbell,[2] Martin Roland,[1] John Benson,[3] Jonathan Silverman[4]

To cite: Burt J, Abel G, Elmore N, *et al*. Assessing communication quality of consultations in primary care: initial reliability of the Global Consultation Rating Scale, based on the Calgary-Cambridge Guide to the Medical Interview. *BMJ Open* 2014;**4**: e004339. doi:10.1136/bmjopen-2013-004339

▸ Additional material is available. To view please visit the journal (http://dx.doi.org/10.1136/bmjopen-2013-004339).

Received 25 October 2013
Revised 29 January 2014
Accepted 13 February 2014

For numbered affiliations see end of article.

Correspondence to
Dr Jenni Burt;
jab35@medschl.cam.ac.uk

ABSTRACT

Objectives: To investigate initial reliability of the Global Consultation Rating Scale (GCRS: an instrument to assess the effectiveness of communication across an entire doctor–patient consultation, based on the Calgary-Cambridge guide to the medical interview), in simulated patient consultations.

Design: Multiple ratings of simulated general practitioner (GP)–patient consultations by trained GP evaluators.

Setting: UK primary care.

Participants: 21 GPs and six trained GP evaluators.

Outcome measures: GCRS score.

Methods: 6 GP raters used GCRS to rate randomly assigned video recordings of GP consultations with simulated patients. Each of the 42 consultations was rated separately by four raters. We considered whether a fixed difference between scores had the same meaning at all levels of performance. We then examined the reliability of GCRS using mixed linear regression models. We augmented our regression model to also examine whether there were systematic biases between the scores given by different raters and to look for possible order effects.

Results: Assessing the communication quality of individual consultations, GCRS achieved a reliability of 0.73 (95% CI 0.44 to 0.79) for two raters, 0.80 (0.54 to 0.85) for three and 0.85 (0.61 to 0.88) for four. We found an average difference of 1.65 (on a 0–10 scale) in the scores given by the least and most generous raters: adjusting for this evaluator bias increased reliability to 0.78 (0.53 to 0.83) for two raters; 0.85 (0.63 to 0.88) for three and 0.88 (0.69 to 0.91) for four. There were considerable order effects, with later consultations (after 15–20 ratings) receiving, on average, scores more than one point higher on a 0–10 scale.

Conclusions: GCRS shows good reliability with three raters assessing each consultation. We are currently developing the scale further by assessing a large sample of real-world consultations.

Strengths and limitations of this study

- The Global Consultation Rating Scale (GCRS) is based on the widely used Calgary-Cambridge guide to the medical interview, and is designed to evaluate a practitioner's communication skills across an entire consultation, linking the identification of potential training needs to an established approach to teaching communication skills.
- We considered evaluator bias and order effects to obtain a more robust assessment of the reliability of GCRS to evaluate communication competence within a particular consultation.
- A particular limitation is that our findings are based on the use of simulated patient consultations. This had an impact on our ability to assess the performance of GCRS to evaluate communication competence of individual doctors, rather than particular consultations. A full evaluation of the performance of GCRS requires the assessment of real-world consultations and we are undertaking this at present.

INTRODUCTION

During the past 30 years, an extensive research literature has defined the skills that enhance communication between doctor and patient. This evidence demonstrates the essential role that communication plays in high-quality healthcare by enabling more accurate, efficient and supportive interviews, by enhancing patient and professional experience and by improving health outcomes for patients. The use of specific communication skills has been shown to lead to improvements in symptom relief, in clinical outcomes and possibly in medicine adherence.[1–6] In light of these findings,

there has been increasing pressure from professional medical bodies to improve the training and evaluation of doctors in communication.[7–13]

In order to evaluate doctors' communication skills effectively, tools with solid theoretical grounding and good psychometric properties are required. Various rating scales exist to assess doctor–patient consultations, which vary widely in their setting, approach and in the published details of their psychometric properties.[14 15] Perhaps for these reasons, none have become standard to use within the National Health Service (NHS), in spite of National Institute for Health and Care Excellence (NICE) standards which require that "Patients experience effective interactions with staff who have demonstrated competency in relevant communication skills."[16] Recently, there has been a move towards domain, or global, marking schemes (awarding overall marks to groupings of items) rather than itemised checklists, the suggestion being that checklists may reward thoroughness rather than competence and work better for novices than for experts.[17] Global marking schemes may be more useful in postgraduate assessments, improving professional authenticity. We have, therefore, developed the Global Consultation Rating Scale (GCRS), based on the Calgary-Cambridge guide to the medical interview, to evaluate the communication effectiveness of an entire doctor–patient consultation, using the domain marking approach.

At present, there is a dearth of assessment tools that robustly measure the overall communication skills of an individual general practitioner (GP) in real-world practice. While a number of existing tools may be used to assess doctor–patient communication, their suitability to assess a doctor's overall communication skills in day-to-day practice irrespective of the content of the consultation is limited and they do not link specifically to educational material commonly used in the UK for subsequent communication skills development. GCRS differs from some alternative instruments, such as the MAAS-Global, in its aim of measuring communication skills only, irrespective of clinical content, to provide an assessment of doctors' generic communication skills and to thereby enable targeted communication teaching. For example, 4 of the 17 items in the MAAS-Global specifically assess medical content related to history, examination, diagnosis and management and other communication items are highly specific to particular content areas.[18] In comparison, the 12 global areas of GCRS include only communication process skills without content. Following the approach of the Calgary-Cambridge guide from which it is derived, GCRS takes the standpoint that, although the context of the interaction changes and the content of the communication varies, the process skills themselves remain the same and can be evaluated independently. This, together with domain rather than individual skill marking, enables the assessment of communication skills across a wide variety of consultations, especially helpful in real-world consultations where communication checklists cannot be specific and tailored for each case.

The Calgary-Cambridge guide to the medical interview[1 19–21] was developed by Silverman, Kurtz and Draper to delineate effective physician–patient communication skills and to provide an evidence-based structure for their analysis and teaching. Within the UK, over half of UK medical schools now use the Calgary-Cambridge approach in their communication skills programmes.[22] It has been widely translated and is used in the USA, Canada and Europe. It has been used to teach communication in general practice and specialist environments, at undergraduate and postgraduate levels.

Specific tools have been developed from the guide for the assessment of medical students, practising paediatricians, dentists, pharmacists and veterinary practitioners, as well as for specific components of the consultation such as explanation and planning in OSCE style examinations.[23–25] Before now however, there has been no validated method of using the Calgary-Cambridge consultation guide to assess complete consultations between qualified doctors and patients. This type of assessment is particularly important in postgraduate and continuing medical education in which the observation of whole consultations from real practice provides increased validity. In addition, for personal development and annual appraisal, a reliable validated assessment tool which also enables a specific link to targeted teaching of communication skills is particularly relevant. Our intention with GCRS is to develop an instrument capable of credibly evaluating a doctor's communication competence, identifying potential areas for improvement which could then be addressed directly with linked, tailored education, using the Calgary-Cambridge guide.

The aim of this study was to investigate the initial reliability of GCRS in simulated patient consultations such as those which might be used in training, as a precursor to its use with real patient consultations where GPs are assessed on their performance. To assess reliability, we asked five specific questions. These are detailed below, together with the reasons for their investigation:

A. Does a fixed difference between scores in GCRS have the same meaning at all levels of performance? If it does not, GCRS scores may not be useful for distinguishing between performance uniformly at all levels of performance, and could require transformation prior to analysis.
B. What is the reliability of GCRS in assessing individual consultations (with different numbers of raters per consultation)? One of two core questions: how consistently does GCRS perform in evaluating communication skills within a particular consultation, and how many raters are required to obtain performance estimates we are confident distinguish better from worse consultations?
C. What is the reliability of GCRS in assessing individual doctors' performance across a number of

consultations (with different numbers of raters and consultations per doctor)? The second core question: how many consultations, and how many raters, do we need to evaluate a particular doctors' consultation skills such that we can differentiate them from their peers?

D. Are some raters more generous than others in their assessments of consultations? Wide variation between the scores assigned by raters can lead to reduced reliability. Understanding whether systematic biases are present helps to inform whether to adjust reliability estimates for these or not.

E. Does the order in which a consultation is rated affect the score? Psychological experiments have shown that the order in which information is presented can influence the way in which that information is processed.[26] Sequential order biases may present themselves either as an overall increase or decrease in scores throughout a judging period; or as observable effects of implicit comparisons being made between the previous and current items being judged.[27 28] Thus, a GCRS rater may use norm-based rather than criterion-based referencing when assigning scores as they proceed through the consultations being evaluated.

METHODS

Trained GP raters watched video recordings of consultations between volunteer GPs and simulated patients and completed GCRS for each. We used videos from a previous study investigating the way in which GPs discussed taking statins to prevent cardiovascular disease with simulated patients trained to play one of two roles. The two roles differed in the extent of the actor's assertiveness in asking questions about proposed management. Both roles displayed sufficient cardiovascular risk to be eligible for statins according to current NICE recommendations. Actors were experienced in playing the role of simulated patients. They were provided with a detailed written role description, including notes on their intended style of response to questions. Actors rehearsed their roles before undertaking videotaped simulations with participant GPs. GPs (n=23) selected for recruitment to the original study varied in age, gender, length of time since qualification and nature of practice (location, size and involvement with dispensing or training). They were recruited from four primary care trusts across the East of England (Cambridge, Luton, Bedford and Peterborough). Each GP conducted two consultations in their practice (one with each simulated patient), furnished with the results of appropriate medical investigations for the simulated patient. The purpose of the consultation was, from the perspective of GP and patient, to discuss the possibility of starting statin medication. This generated a total of 46 recorded consultations. For this study, we excluded videos from two GPs: one had since become a trained GP GCRS evaluator, while the videos for the second were damaged (see online supplementary appendix 1 figure S1). This left 42 videoed consultations for assessment. All GPs gave their written consent for the re-use of their videos.

Global Consultation Rating Scale

The GCRS covers 12 domains from 'initiating the session' to 'closure' (see online supplementary appendix 3 for the full scale). Guidance is given within the text of the scale as to the nature of the skills that are assessed within each individual domain, which is given a score as follows: Not applicable (not scored)

0. Not done/poor
1. Adequate
2. Good

The use of a three-point scale, while narrow, (1) enables a clear focus on identifying the likely need for targeted training in that area and (2) reflects the need for a simple and easy-to-use scale suitable for use while observing a consultation. A total consultation score between 0 and 24 is obtained by summing the scores from the 12 domains. In the case where a domain is considered to be not applicable, scores are renormalised to be out of 24, for example, a score of 12 out of 22 would become a score of 13.1 (=12×24/22) out of 24 (NB: this was not required in this study).

GP raters

We recruited six GP raters experienced in teaching and assessing communication skills using the Calgary-Cambridge consultation guide within the School of Clinical Medicine, University of Cambridge. All attended a 2 h training session on the use of GCRS with JS, which included a specially created training video of consultations for evaluation. In training, particular attention was paid to the differences between 'good', 'adequate' and 'poor' communication behaviours, guided by the criterion referenced norms established by the Calgary-Cambridge guide. The aim was to establish a shared understanding of expected standards of behaviour across each domain.[29] Following training, each evaluator rated 28 videos. These were randomly assigned and provided in a random order for rating. Randomisation was performed with maximum cross over between raters to allow study of possible order effects (see online supplementary appendix for further details).

GP raters were requested to complete evaluations within 1 month of collecting the videos and were paid for their time. On receipt of ratings some missing domain scores were noted (19 of 2184, 0.87%). The five raters who had missed scores watched the corresponding videos again and filled in the missing sections only. Double data entry was conducted (NE, GA) for all ratings. For the four scores (0.20%), in which there was inconsistency, the original score sheets were consulted to obtain the correct score.

Statistical analysis

The overall aim of this work was to estimate the statistical reliability of GCRS as a tool to assess consultations or doctors. Statistical reliability is an index of how well better performance can be distinguished from worse performance, and estimates how much of the variation in scores is due to true variation in performance rather than to noise due to different raters rating the same consultation differently. A reliability of 1 indicates that all the variation in measured scores is due to true variation in performance, that is, that scores are perfectly reliable. A reliability of 0 indicates that all the variation in measured scores is due to statistical noise. Between these two extremes, a reliability of 0.8 is generally considered the minimum required for most applications.[30]

Does a fixed difference between scores in GCRS have the same meaning at all levels of performance?

One of the key assumptions made when calculating reliability is that measurement errors are independent of the true values. When this is not true a single reliability value cannot apply to all scores. Another way of thinking of this is that we require a fixed difference between two scores (eg, a two point difference) to have the same distinguishing quality across the full range of scores. For this to be true, the variability in raters' scores of the same consultation must be the same at all levels of performance. We checked this by plotting the SD of ratings for each consultation against the mean score for that consultation (a variation on the standard Bland-Altman plot, allowing for more than two ratings per consultation). We found that the variance was not the same across all mean scores, implying that, for raw scores, a fixed difference does not have the same meaning at all levels of performance. We, therefore, sought a transformation to stabilise the variance across all mean scores. The transformed data were used for all further analysis.

What is the reliability of GCRS for assessing single consultations?

Our experimental setup allowed us to distinguish between three different sources of variance:

1. differing performance between doctors
2. differing performance of the same doctor between consultations, and
3. differing evaluator scores of the same consultation

In order to calculate the crude reliability, we fitted a three-level linear regression model to reflect this, with no fixed effects and with random intercepts for consultation and doctor (ie, rating nested within consultation further nested within doctor). From such a model we can estimate the reliability that would be achieved for assessing single consultations with different numbers of raters (see online supplementary appendix). The same analysis was performed on the scores for each of the individual domain of GCRS.

What is the reliability of GCRS in assessing individual doctors' performance across a number of consultations?

Using the same approach, we can also estimate the reliability of GCRS for assessing doctor's performance using different numbers of raters to assess each doctor, and using different numbers of consultations per doctor (see online supplementary appendix).

Are some raters more generous than others in their assessments of consultations?

In order to establish whether there were systematic biases between the scores given by different raters, we augmented the model described above with fixed effects for raters. If present, biases between raters will increase the variation in scores, and in turn reduce the reliability of scores. The systematic biases between raters could be accounted for, and we estimated adjusted reliabilities after doing so.

Does the order in which a consultation is rated affect the score?

Finally, to investigate possible order effects we included the order of rating in the above model. To account for non-linear effects we used a restricted cubic spline with three knots. We excluded data from one evaluator in this analysis because they had not rated the consultations in the order requested.

CIs on all estimates were calculated using bias corrected bootstrapping with 1000 repetitions and resampling at the doctor level.

The approach outlined above falls somewhere between classical reliability studies in which only one source of variance is identified (eg, inter-rater reliability) and a generalisability theory approach.[31] However, due to the limited data available we feel the approach taken is the most appropriate, and further it allows a more nuanced investigation of order effects considering non-linear functions.

Statistical analysis was conducted using Stata V.11.2.

RESULTS

The distribution of mean scores for the 42 consultations assessed (untransformed on a 0–24 scale) is shown in figure 1A. The highest mean consultation score was 16.25 of 24 and the lowest 1.5.

Does a fixed difference in GCRS have the same meaning at all levels of performance?

Figure 1C shows the Bland-Altman type plot for the untransformed data. There was a clear trend of increasing SD of scores for each consultation with increasing mean score. This implies that there was a higher degree of agreement between raters at low scores than at the moderate scores (10–14) which form the upper end of our data set. We found that a transformation based on the logit function performed reasonably well at stabilising the variance (see online supplementary appendix

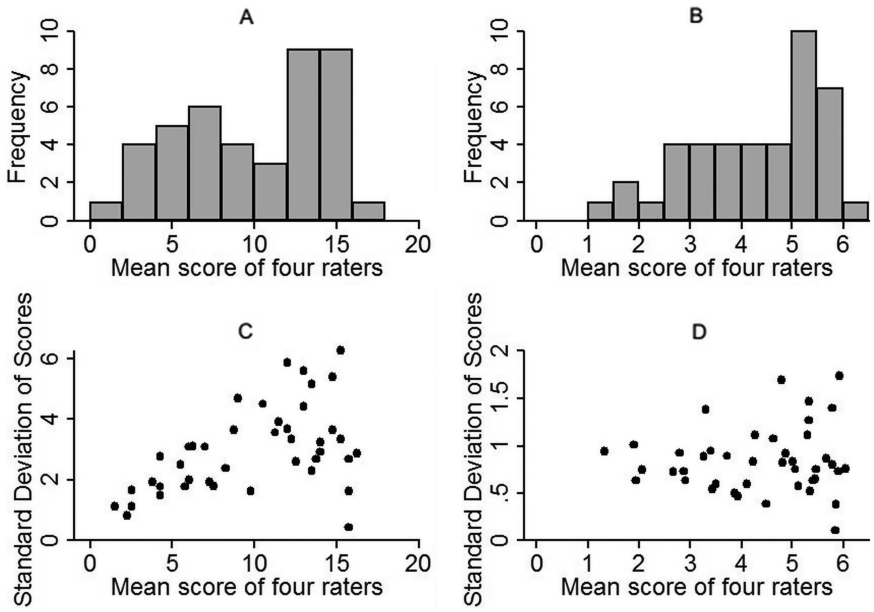

Figure 1 Histograms showing the distribution of mean consultation scores on the native (possible values 0 to 24) scale (A) and transformed (possible values 0 to 10) scale (B). Bland-Altman plot of consultation ratings shown on the native scale (C) and transformed scale (D).

for details and lookup table). The transformation has been constructed such that the transformed scores lie between 0 and 10. The distribution of the transformed scores is shown in figure 1B.

The resulting Bland-Altman plot of transformed data is shown in figure 1D in which there is little indication of a trend (note that the increase in spread of SDs is due to the possible values available and is not considered to be a major issue). All further results relate to the transformed data.

What is the reliability of GCRS in assessing single consultations, and in assessing individual doctors' performance?

The SDs for the three sources of variation estimated from the crude mixed model (with no adjustment for rater bias) are shown in table 1. The largest SD was that for between doctors, implying that this is where the largest variation is seen. The SD of scores of the same consultation by different raters was slightly smaller than that attributed to between doctors' performance. Finally, the estimates suggested that variation at the consultation level within individual doctors was essentially zero ($SD=1.03\times10^{-9}$). This finding is likely to be a function of our dataset. We do not present any reliability estimates for rating doctors here, and outline the reasons for this in the discussion. The reliability estimates for rating consultations for different numbers of raters are shown in table 2. In the crude model, the commonly used reliability thresholds of 0.7 (modest), 0.8 (acceptable) and 0.9 (excellent) were achieved using two, three and seven raters, respectively.[30] With four raters, as used in this study, we achieved a reliability of 0.85 (95% CI 0.61 to 0.88). Details of the distribution of scores and the reliabilities of individual domains are available in online supplementary appendix figure S2 and online supplementary appendix table S2. These indicate that four raters would be sufficient to provide a broad indication of domains where a doctor may have some performance issues.

Are some raters more generous than others in their assessments of consultations?

When we allowed for systematic bias between raters in our model we found that such bias was present (table 3). On an average, a difference of 1.65 (on the 0–10 scale for transformed data) was seen between the least and most generous raters. By adjusting for evaluator bias we increased reliability somewhat (table 2), and the number of raters needed to reach the 0.7, 0.8 and 0.9 thresholds became two, three and five, respectively.

Table 1 SDs estimated for the three sources of variation from a crude model and one adjusting for systematic bias between raters

	SD	
Source of variation	**Crude model**	**Model adjusted for evaluator bias**
Between doctors	1.21 (0.87, 1.38)	1.18 (0.87, 1.33)
Within doctors and between consultations	1.03×10^{-9} (7.25×10^{-13}, 1.95×10^{-9})	0.14 (0.00, 0.15)
Within consultations and between raters	1.03 (0.96, 1.16)	0.88 (0.82, 1.01)

Table 2 Crude and adjusted reliability for evaluating consultations for different numbers of raters using GCRS (transformed 0–10 data)

Number of raters	Crude reliability* (95% CI)	Reliability adjusted for evaluator bias* (95% CI)
1	0.58 (0.28 to 0.66)	0.65 (0.36 to 0.71)
2	0.73 (0.44 to 0.79)	0.78 (0.53 to 0.83)
3	0.80 (0.54 to 0.85)	0.85 (0.63 to 0.88)
4	0.85 (0.61 to 0.88)	0.88 (0.69 to 0.91)
5	0.87 (0.66 to 0.91)	0.90 (0.74 to 0.93)
6	0.89 (0.70 to 0.92)	0.92 (0.77 to 0.94)
7	0.91 (0.73 to 0.93)	0.93 (0.80 to 0.95)
8	0.92 (0.76 to 0.94)	0.94 (0.82 to 0.95)
9	0.93 (0.78 to 0.95)	0.94 (0.84 to 0.96)
10	0.93 (0.80 to 0.95)	0.95 (0.85 to 0.96)

*Calculated from the estimated SDs shown in table 1.
GCRS, Global Consultation Rating Scale.

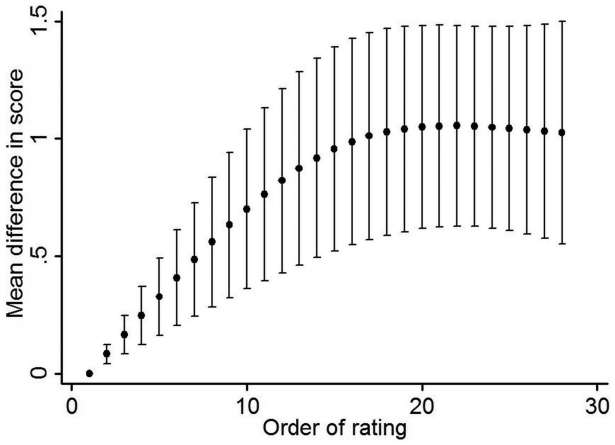

Figure 2 The effect of order of rating on transformed scores compared with the first rating performed. Dots indicate point estimates and bars show 95% CIs.

Does the order in which a consultation is rated affect the score?

Finally, we found evidence of considerable order effects, with raters giving higher scores, on average, as they progressed through the rating of consultations (figure 2). It appears that raters' scoring levelled out after performing around 15–20 ratings. Later consultations received, on average, scores more than one point higher on the 0–10 scale.

DISCUSSION

GCRS shows good reliability (>0.8) with three raters assessing each consultation, and modest reliability (>0.7) with two raters. Overall, consultations received low-to-moderate scores. This reflects previous findings with simulated patients, where it has been seen that participating doctors only attain about 40–60% of the guidelines or standards used for evaluation.[32] GCRS is designed to assess overall communication effectiveness of the entire doctor–patient consultation, encapsulating the quality of the interaction from the opening moments, through the gathering of information, provision of information, achieving shared understanding and shared decision-making, through to closure. It is a performance-based assessment (assessing what doctors actually do in professional practice) rather than a competence-based assessment (assessing what doctors can do in controlled representations of professional practice).[33] It is additionally a criterion-referenced measure; GCRS training course highlights the importance of assessing performance against the 'gold standard' outlined in the Calgary-Cambridge guide.

Table 3 Estimated biases between raters using GCRS (transformed 0–10 data)

Evaluator	Mean difference (95% CI)
1	Reference
2	−0.25 (−0.57 to 0.13)
3	−0.68 (−1.20 to −0.18)
4	0.97 (0.66 to 1.33)
5	−0.25 (−0.76 to 0.31)
6	0.49 (0.04 to 0.96)

GCRS, Global Consultation Rating Scale.

While GCRS was devised as a global assessment, doctors may be interested in knowing their performance in particular domains in order to most efficiently target training. For individual GCRS domains, reliability was broadly acceptable with four raters. Low reliability for two particular domains—non-verbal communication and closure—may be attributable to small between-consultation variance rather than to raters disagreeing with each other on these areas. There are two possible explanations: either that raters find it difficult to distinguish differences in doctors' behaviours on these items (reflecting inadequate training for raters in how to assess these domains, or challenges in capturing, eg, non-verbal behaviour) or that doctors perform comparably across consultations and compared with each other on these two domains, prompting raters to award consistently similar scores.

We found that a fixed difference between scores in GCRS did not have the same meaning at all levels of performance: untransformed scores (on a scale of 0 to 24) showed a higher degree of agreement between raters at low scores than at moderate scores. For this reason, analyses were performed on transformed scores. This has implications for the most suitable score to feedback to participants if, for example, GCRS is to be used in a training situation. Transformed scores may be intuitively more difficult for participants to understand, and we need to undertake further work on the acceptability of using transformed scores in assessments of an individual doctors' performance, and how best to calculate and present transformed scores for doctors and trainers.

While we found good reliability of GCRS in assessing the communication quality of individual consultations, comparison with existing instruments is difficult due to limited published psychometric data on assessing consultation (rather than doctor) quality. Interconsultation doctor reliability has been evaluated using the Four Habits Coding Scheme over 13 consultations (reliability of 0.72 with two raters),[34] and using the Liv-MAAS over nine consultations (reliability of 0.78 with three raters).[35] Evaluating the reliability of GCRS for assessing performance of individual doctors using different numbers of consultations will require more consultations per doctor, probably with greater subject variety, than we had in our dataset. We hope that further work on GCRS will enable us to estimate this in future.

We found consistent differences in scores assigned to consultations by the most and least generous raters. The Hawk/Dove phenomenon is well documented across a wide range of performance assessments, and can be addressed through training, through the use of more than one rater and through the use of post hoc statistical techniques.[36] All of these were employed in this study, and our finding of such variation highlights the importance of using pre-evaluation and postevaluation approaches in monitoring and acting upon differences between raters.[37]

We found evidence of considerable order effects. The use of multiple raters rating consultations in random order will tend to reduce order effects: sometimes a consultation will be rated early by an evaluator, and sometimes late; thus different orders for different raters average out. We have not been able to find other examples of the examination of this in GP consultation evaluation, but as previously stated, the influence of the sequential presentation of information on subsequent assessments of this information is a well-known phenomenon in the psychological literature.[26] Again, this is something which requires further work to assess how GCRS will perform in training situations.

The current study has a number of limitations. We included only a small number of GPs whose consultations had been recorded, derived from an earlier study, and only two similar scenarios per GP. These standardised scenarios do not reflect real-world consultations of variable nature and content, and we believe these are the reasons why we find little variation between consultations of the same doctor. We could not, therefore, assess how raters responded to different contexts: this is the focus of our next stage of work.

There are various sources of possible bias we did not examine due to sample size limitations. For example, contrast effect bias may be important in influencing rater behaviour, where, for example, viewing a good consultation after a series of poor consultations may lead to a substantial leap in scores assigned due to the contrast between them.

Feedback from raters showed that the assessment of consultations required significant concentration. Average consultation length was around 15 min: viewing each consultation and completing the rating scale means each evaluation can take around 20 min.

CONCLUSIONS

GCRS has good reliability (>0.8) for rating consultations if three raters are used. Systematic differences were observed between raters: adjusting for these further improves reliability of the scale. We are currently developing the scale further by assessing a large sample of consultations in a real-world setting. This will enable a more detailed examination of the ability of the scale to assess performance between consultations of the same doctor. Once further psychometric evaluation is completed, we envisage that GCRS has the capacity to provide a robust yet practical assessment tool for the evaluation of communication skills in everyday practice, linked to the Calgary-Cambridge training approach to target identified areas for improvement.

Author affiliations
[1]Cambridge Centre for Health Services Research, University of Cambridge, Cambridge, UK
[2]University of Exeter Medical School, University of Exeter, Exeter, UK
[3]Primary Care Unit, University of Cambridge, Cambridge, UK
[4]School of Clinical Medicine, University of Cambridge, Cambridge, UK

Acknowledgements The authors would like to thank all participating general practitioners (GPs) and GP evaluators for their assistance with this work. The authors also thank the two reviewers whose thoughtful feedback greatly improved this article.

Contributors JBu designed the study, contributed to the analysis and interpretation of data and drafted the article. GA designed the study, undertook the analysis and contributed to the interpretation of data and drafting of the final version of the article. NE undertook data collection, and contributed to the analysis, the interpretation of data and drafting of the final version of the article. JC and MR designed the study, contributed to the interpretation of data and critically revised the article. JBe designed the study, supervised data collection and contributed to the interpretation of data and drafting of the final version of the article. JS designed the study, contributed to the interpretation of data, critically revised the article and devised the Global Consultation Rating Scale. All authors conceived the study and approved the final version of the article.

Funding This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None.

Ethics approval Bromley Research Ethics Committee (REC ref: 12/LO/0421).

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

REFERENCES

1. Silverman J, Kurtz S, Draper J. *Skills for communicating with patients*. 3rd edn. Oxford: Radcliffe, 2013.
2. Makoul G. The interplay between education and research about patient-provider communication. *Patient Educ Couns* 2003;50:79–84.
3. Simpson M, Buckman R, Stewart M, *et al*. Doctor-patient communication: the Toronto consensus statement. *BMJ* 1991;303:1385–7.

4. Stewart M, Brown JB, Boon H, *et al*. Evidence on patient-doctor communication. *Cancer Prev Control* 1999;3:25–30.
5. Suchman AL. Research on patient-clinician relationships: celebrating success and identifying the next scope of work. *J Gen Intern Med* 2003;18:677–8.
6. von Fragstein M, Silverman J, Cushing A, *et al*. UK consensus statement on the content of communication curricula in undergraduate medical education. *Med Educ* 2008;42:1100–7.
7. Association of American Medical Colleges. Report 3: Contemporary Issues in Medicine: Communication in Medicine. Washington, DC: AAMC, 1999.
8. BMA. *Communication skills education for doctors: a discussion document*. London: BMJ, 2003.
9. Cowan DH, Laidlaw JC. Improvement of teaching and assessment of doctor-patient communication in Canadian medical schools. *J Cancer Educ* 1993;8:109–17.
10. Department of Health. *Medical schools: delivering the doctors of the future*. London: Department of Health, 2004.
11. General Medical Council. *Tomorrow's doctors: recommendations on undergraduate medical education*. London: GMC, 2009.
12. The Royal College of Physicians and Surgeons of Canada. Canadian Medical Education Directions for Specialists 2000 Project: Skills for the New Millennium: Report of the Societal Needs Working Group, 1996.
13. Workshop Planning Committee. Consensus statement from the workshop on teaching and assessment of communication in Canadian medical schools. *CMAJ* 1992;147:1149–50.
14. Boon H, Stewart M. Patient-physician communication assessment instruments: 1986 to 1996 in review. *Patient Educ Couns* 1998;35:161–76.
15. Schirmer JM, Mauksch L, Lang F, *et al*. Assessing communication competence: a review of current tools. *Fam Med* 2005;37:184–92.
16. NICE. Patient experience in adult NHS services. Quality Standards, QS15. 2012.
17. van der Vleuten CP, Schuwirth LW, Scheele F, *et al*. The assessment of professional competence: building blocks for theory development. *Best Pract Res Clin Obstet Gynaecol* 2010;24:703–19.
18. van Thiel J, Ram P, van Dalen J. *MAAS-Global manual*. Maastricht, Netherlands: University of Maastrich, 2000.
19. Kurtz S, Silverman J, Benson J, *et al*. Marrying content and process in clinical method teaching: enhancing the Calgary-Cambridge guides. *Acad Med* 2003;78:802–9.
20. Kurtz SM, Silverman J, Draper J. *Teaching and learning communication skills in medicine*. 2nd edn. Oxford, San Francisco: Radcliffe Medical, 2005.
21. Kurtz SM, Silverman JD. The Calgary-Cambridge referenced observation guides: an aid to defining the curriculum and organizing the teaching in communication training programmes. *Med Educ* 1996;30:83–9.
22. Gillard S, Benson J, Silverman J. Teaching and assessment of explanation and planning in medical schools in the United Kingdom: cross sectional questionnaire survey. *Med Teach* 2009;31:328–31.
23. Howells RJ, Davies HA, Silverman JD, *et al*. Assessment of doctors' consultation skills in the paediatric setting: the Paediatric Consultation Assessment Tool. *Arch Dis Child* 2010;95:323–9.
24. Radford A, Stockley P, Silverman J, *et al*. Development, teaching, and evaluation of a consultation structure model for use in veterinary education. *J Vet Med Educ* 2006;33:38–44.
25. Silverman J, Archer J, Gillard S, *et al*. Initial evaluation of EPSCALE, a rating scale that assesses the process of explanation and planning in the medical interview. *Patient Educ Couns* 2011;82:89–93.
26. Mussweiler T. Comparison processes in social judgments: mechanisms and consequences. *Psychol Rev* 2003;110:472–89.
27. Page L, Page K. Last shall be first: a field study of biases in sequential performance evaluation on the idol series. *J Econ Behav Organ* 2010;73:186.
28. Rotthoff KW. (Not Finding a) Sequential Order Bias in Elite Level Gymnastics, 2013.
29. Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. *Teach Learn Med* 2003;15:270–92.
30. Postgraduate Medical Education and Training Board. *Developing and maintaining an assessment system—a PMETB guide to good practice*. London: GMC, 2007.
31. Brennan RL. Generalizability Theory. *Educ Meas Issues Pract* 1992;11:27–34.
32. Rethans JJ, Sturmans F, Drop R, *et al*. Assessment of the performance of general practitioners by the use of standardized (simulated) patients. *Br J Gen Pract* 1991;41:97–9.
33. Rethans JJ, Norcini JJ, Baron-Maldonado M, *et al*. The relationship between competence and performance: implications for assessing practice performance. *Med Educ* 2002;36:901–9.
34. Krupat E, Frankel R, Stein T, *et al*. The Four Habits Coding Scheme: validation of an instrument to assess clinicians' communication behavior. *Patient Educ Couns* 2006;62:38–45.
35. Enzer I, Robinson J, Pearson M, *et al*. A reliability study of an instrument for measuring general practitioner consultation skills: the LIV-MAAS scale. *Int J Qual Health Care* 2003;15:407–12.
36. Harasym PH, Woloschuk W, Cunning L. Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. *Adv Health Sci Educ Theory Pract* 2008;13:617–32.
37. Bartman I, Roy M, Smee S. *Catching the hawks and doves: a method for identifying extreme examiners on objective structured clinical examinations*. Ottawa: Medical Council of Canada, 2011.