LUCAS: a theoretically informed instrument to assess clinical communication in objective structured clinical examinations

Christopher D Huntley, Peter Salmon, Peter L Fisher, Ian Fletcher & Bridget Young

OBJECTIVES We reviewed papers describing the development of instruments for assessing clinical communication in undergraduate medical students. The instruments had important limitations: most lacked a theoretical basis, and their psychometric properties were often poor or inadequately investigated and reported. We therefore describe the development of a new instrument, the Liverpool Undergraduate Communication Assessment Scale (LUCAS), which is intended to overcome some of these limitations. We designed LUCAS to reflect the theory that communication is contextually dependent, inherently creative and cannot be fully described within a conceptual framework of discrete skills.

METHODS We investigated the preliminary psychometric properties of LUCAS in two studies. To assess construct and external validity, we examined correlations between examiners' LUCAS ratings and simulated patients' ratings of their relationships with students in Year 1 formative (n = 384) and summative (n = 347) objective structured clinical examination (OSCE) samples. Item–total correlations and item difficulty analyses were also performed. The dimensionality of LUCAS was examined by confirmatory factor analysis. We also assessed inter-rater reliability; four raters used LUCAS to rate 40 video-recorded encounters between Year 1 students and simulated patients.

RESULTS Simulated patient ratings correlated with examiner ratings across two OSCE datasets. All items correlated with the total score. Item difficulty showed LUCAS was able to discriminate between student performances. LUCAS had a two-dimensional factor structure: we labelled Factor 1 creative communication and Factor 2 procedural communication. The intraclass correlation coefficient was 0.73 (95% confidence interval 0.54–0.85), indicating acceptable reliability.

CONCLUSIONS We designed LUCAS to move the primary focus of examiners away from an assessment of students' enactment of behavioural skills to a judgement of how well students' communication met patients' needs. LUCAS demonstrated adequate reliability and validity. The instrument can be administered easily and efficiently and is therefore suitable for use in medical school examinations.

Medical Education 2012: ***46****: 267–276*
doi:10.1111/j.1365-2923.2011.04162.x

Department of Mental and Behavioural Health Sciences, University of Liverpool, Liverpool, UK

Correspondence: Dr Bridget Young, Department of Mental and Behavioural Health Sciences, University of Liverpool, Whelan Building, Brownlow Hill, Liverpool L69 3GB, UK. Tel: 00 44 151 794 5525; Fax: 00 44 151 794 5537; E-mail: Bridget.Young@liv.ac.uk

INTRODUCTION

Many medical schools now assess students' competence in clinical communication.[1] Assessment is typically by objective structured clinical examinations (OSCEs). Here, students consult with simulated patients (SPs) trained to follow standardised scenarios, while being rated by examiners or SPs according to standardised instruments. Many such instruments are now available. These must be efficient and robust if they are to be used in large, resource-limited undergraduate programmes and to assess students with fairness in high-stakes examinations. That is, they need to be brief and easily administered, to reflect theoretical understanding of the nature of clinical communication, and to measure communication validly and reliably.[2] In this paper, we review existing instruments before detailing theoretical and pragmatic considerations that underpinned the development of the Liverpool Undergraduate Communication Assessment Scale (LUCAS), which was intended to overcome some of the limitations of other instruments.

We reviewed instruments for assessing clinical communication in medical students. We searched the literature published in English between 1975 and 2009 in PubMed, Scopus and Web of Knowledge using combinations of the search terms 'clinical', 'communication', 'physician', 'doctor', 'patient', 'student', 'relations', 'interview', 'consultation', 'OSCE', 'medical', 'education' and 'undergraduate'. We also searched reference lists of relevant papers. We identified 14 instruments that met our inclusion criteria (Tables S1 and S2, online) and explored their theoretical background, validity and reliability.

Given the contested nature of communication and communication training,[3–7] it is surprising that the theoretical formulation of communication that informed these instruments was rarely explicit. Indeed, only one paper[8] detailed the theoretical conceptualisation – cognitive 'scripts'[9] – upon which the instrument was based. Nevertheless, the instruments varied widely in ways that implicitly suggested fundamentally different conceptualisations of communication. Some were to be completed by academic staff or practitioners, suggesting that good communication is objectively observable. Others were to be completed by SPs, suggesting that quality depends on receivers' experience. Most instruments sampled discrete skills. For example, an early iteration of one instrument[10] comprised 129 behavioural items, probably reflecting an unstated assumption that communication should be assessed by examining whether or not students' behavioural repertoire encompasses discrete skills that are presumed necessary to meet patients' needs. By contrast, another[11] contained only four global items and an overall item, probably reflecting an assumption that competent communication cannot be conceptualised as discrete skills, and that its assessment requires a holistic approach.

Consistent with the lack of theoretical specification, only limited evidence of the validity of the instruments has been published. An instrument's content validity might be demonstrated via the generation of its test items.[12] However, the source of items was often unclear. Some studies reported feedback from researchers or other experts,[10,13,14] doctors,[13] SPs[13] or medical students[13] in the course of an instrument's development, but none appeared to be guided by explicit conceptualisations of the nature of clinical communication or informed by evidence from patients. Other approaches to validity provided further clues to originators' assumptions. Support for construct validity refers to the relationship of scores to evidence of relevant theoretical constructs[12] and several studies have validated examiners' ratings against ratings by other groups, including unspecified communication experts[15] or SPs,[8] again reflecting different opinions about whether communication quality resides in observable skills or recipients' experience. Support for construct validity has also been claimed in contexts in which senior students who received communication training obtained higher scores than junior students who did not,[11,14–17] presumably reflecting the belief that communication quality depends on explicit teaching. No evidence of predictive validity demonstrated by, for example, indications that scores predict aspects of clinical performance that should theoretically depend on good communication has been shown for any instrument.

The factorial structure of the instruments has only rarely been reported as evidence of their underlying constructs. In factor analysis of one instrument,[18] communication (items loading on this factor included 'discussed options', 'answered questions') was distinguished from 'interpersonal skills' (e.g. 'was friendly', 'showed interest in the patient as a person'), implying that communication is multifaceted.[18,19] Widespread use of Cronbach's alpha to indicate internal consistency implies that instruments are measuring single constructs.[17,18] Further support for construct validity has come from studies investigating convergent and discriminant validity (respectively: a measure's correlation with others with which

theoretically it should be related and, conversely, a measure's independence from others to which it should be unrelated).[12] Correlations with other communication assessment instruments were reported.[8,10,13,14,18,20] One instrument demonstrated discriminant validity in showing no association with general academic aptitude.[17] However, these findings were not accompanied by an explicit theoretical rationale for the selection of comparator instruments. No study reported evidence for criterion validity (evidence of concurrent validation or predictive validation), principally because of the lack of a suitable criterion or accepted reference measure.

More attention has been paid to reliability,[21] although this too has sometimes been disregarded.[22,23] Amongst the 10 papers that did report reliability, the diverse statistics used limit the comparability of the instruments.[10,11,13–19,24] Four papers[13,15,16,24] reported bivariate correlations between two raters, but such correlations only estimate reliability for one pair of raters and overestimate inter-rater agreement because they disregard additive bias (when one rater consistently scores more highly than the other). Product–moment correlations have ranged from 0.49 to 0.97. Intraclass correlation coefficients (ICCs) can be used; these describe the variance explained by differences between students' real scores as a proportion of total variance, including variance ascribed to raters, as well as error. This statistic can accommodate more than two raters. However, there are different versions of the ICC, each with different assumptions of generalisability to the population of raters and varying sensitivity to systematic differences between raters.[25] More frequently, studies have reported generalisability coefficients. Conceptually, these are an extension of the ICC methodology but, given sufficient observations per student, they can distinguish the contribution of multiple sources of error such as those relating to the station, day, examiner and SP. Reported generalisability coefficients have ranged from 0.59 to 0.72, suggesting moderate to good reliability. However, no study reported confidence intervals (CIs) for any of these reliability estimates.

The conceptual and psychometric limitations of existing instruments contrast with the high-stakes nature of the assessments for which they can be used. Therefore, despite the proliferation of instruments, there is a need for one that is explicitly informed by a theoretical conceptualisation of clinical communication, and possesses demonstrable validity and reliability. Here, we describe the background to the design and psychometric testing of a new clinical communication assessment instrument.

Construction of the LUCAS

The Liverpool Undergraduate Communication Assessment Scale is a 10-item rating instrument intended for use by examiners in OSCEs. Here, we describe its development and the rationales for the studies we used to investigate its psychometric properties. Our theoretical context assumed that skilled communication is only loosely related to communication skills.[4–6] The meaning of most communication – in clinical settings, as in daily life – is subjectively defined and therefore varies according to the context. For example, what a student says will have different effects on different patients depending on the patients' particular needs and expectations[4–6,26] and thus a communication component that one patient regards as caring may be perceived negatively by another. It is rarely possible to point to a discrete skill that has a consistent meaning, or to attribute specific outcomes to discrete communication skills. Communication is therefore inherently creative and cannot be developed simply by learning skills and rules for combining and displaying them. A few communication behaviours, such as those conveying disrespect, are likely to disrupt clinical care in almost any context and need to be detected, but, given its individual and contextual specificity, communication should be assessed holistically and flexibly.[27] Therefore, rather than making technical judgements about whether predefined skills are displayed, examiners need to make aesthetic judgements about whether communication 'works'. We further theorised that poor communication rarely reflects the absence of skills in students' behavioural repertoires and more often reflects students' underlying knowledge, emotional processes or values.[4,5]

LUCAS reflects this reasoning in several ways. Firstly, we wanted to measure whether students communicated in ways that addressed patients' needs, not whether they had particular skills in their behavioural repertoire. Our instrument's guidance for examiners explicitly directs them to consider the impact – on patients in general – of students' communication, rather than merely scoring students' behaviour in isolation. (Appendices S1 and S2, online, show LUCAS and the accompanying scoring descriptors, respectively.) Linked to this, most items are global and require examiners to use judgement and empathy with the patient in assessing communication. In principle, SPs may be in the best position to judge communication in this way. However, governance

considerations constrain the use of non-university staff in these high-stakes examinations. Moreover, quality of communication extends beyond 'customer evaluation' to encompass ethical issues such as equity and the requirement to address patients' needs, even when these differ from wants, and we reasoned that examiners are probably better placed than SPs to judge these issues. Finally, it was crucial that the instrument be feasible for use in an OSCE context, in which scoring must be completed in only 30 seconds.

Secondly, the quality of communication is only loosely related to what can be observed. Particularly in the absence of behaviours that most observers would recognise as damaging, the quality of communication reflects the perspective of the receiver or observer and it is unrealistic for an examiner to aim to rate quality objectively on a continuous scale. Therefore, we designed the instrument to identify problematic rather than excellent communication: that is, communication that prevents the patient from establishing a sense of relationship. This aim was operationalised using the response options Competent, Borderline and Unacceptable for items C–H.

Thirdly, to identify the relatively few but clinically important behaviours that our and others' inductive research exploring patients' experiences of communication[28–30] has indicated would prevent most patients from developing a sense of relationship with their doctor (e.g. disrespect, false reassurance, inaccurate information), we designed two professional behaviour items (items I and J). Our assumption of the dichotomous nature of such communication informed our decision to use only two scoring options, Competent or Unacceptable, for these items. Their enhanced scoring weighting reflects the adverse impact on patients of unprofessional behaviour and so penalises students who communicate in these ways. The instrument also includes two specific behaviours, introduction and identity check, which are regarded as essential.

Given this background, the validity of LUCAS can be demonstrated by relationships between the patient's own experience of the consultation and the examiner's rating.[31] Therefore, if the items pertain to qualities that are indeed necessary for a clinical relationship, examiner scores should positively correlate with the SP's sense of relationship. This is a form of convergent validity whereby SP ratings of student performance are taken as proxies for real patient ratings. Thus, we compared examiner scores with SP ratings. Corresponding to the longstanding construction of clinical relationships as comprising two dimensions – affiliation and authority[32,33] – we asked SPs to rate students according to their sense of the student's warmth and caring, and their degree of confidence in the student. Furthermore, users need to know the factorial structure of an instrument as a guide to whether it is measuring distinct components of communication and therefore we explored LUCAS's dimensionality. Finally, given that LUCAS is intended for use in a range of settings with a range of examiners, reliability should be calculated in a way that quantifies generalisability to the population of examiners.

In summary, LUCAS is derived from a theory which conceptualises clinical communication as inherently creative and subjective. In order to establish preliminary psychometric properties, we conducted two studies examining validity and inter-rater reliability, respectively.

METHODS

Study 1: validity

Procedure

After receiving ethical approval from the university's research ethics committee, we drew on data from two cohorts of Year 1 medical students attending formative (n = 384) and summative (n = 347) communication OSCEs to examine the validity of LUCAS. The mean age of the students at the start of the academic year was 19.0 years (standard deviation [SD] = 1.61 years) and 53.7% were female. Each OSCE consisted of four stations. Formative scenarios for these stations referred to Crohn's disease, dizzy spells, high blood pressure and chesty cough. Summative scenarios involved deep vein thrombosis, irritable bowel syndrome, chronic fatigue syndrome and glue ear. Students had a maximum of 5 minutes at each station. For each student, ratings for each item were averaged across the stations to provide a single score. Examiners had previous experience as OSCE examiners and were clinicians, social science or clinical academic research staff, or doctoral students of clinical communication. They were given written information on the OSCE process including the assessment instrument and associated descriptors. In the absence of suitable criteria for assessing validity, we focused on support for construct validity. Firstly, we examined correlations of items with two ratings by SPs, which addressed two fundamental aspects of clinical relationships: affiliation ('I felt this person was warm and caring'), and authority ('I felt

confident in this person').[32,33] Each was rated on a 5-point scale ranging from 'Strongly disagree' to 'Strongly agree'. Item–total correlations were calculated to determine if individual LUCAS items were consistent with the rest of the scale. We also performed an item difficulty analysis whereby we calculated, for each item, the proportion scored correctly. Finally, we examined the factor structure as described below.

Data analysis

Product–moment correlations were calculated between examiner and SP ratings, with significance set at $p < 0.01$ to constrain type I errors. To explore the factor structure, we performed separate principal components analyses of each dataset, deciding with the aid of a scree test the number of components to retain for oblique rotation. Loadings of ≥ 0.25 were regarded as potentially indicating significant relationships.

We calculated corrected item–total correlations, regarding those of < 0.30 as potentially indicating items that might be removed. Item difficulty was defined according to the proportion of students scoring the item correctly (i.e. 'competent'). We expected most items to score > 0.7, indicating that examiners considered that the majority of students' performances did not preclude or damage the doctor–patient relationship.

We tested the suggested model by separate confirmatory factor analyses of data for each cohort using maximum likelihood estimation applied to the covariance matrix. Variances of independent latent variables were fixed so that all path coefficients were free to vary. Alternative models were tested in each dataset, as described below.

Path coefficients indicate the strength of association between variables and were tested individually for significance at $p < 0.05$. In addition, we tested the overall fit of the model. Because there is no single adequate index of fit, several indicators were calculated.[34] Chi-squared tests establish whether the model differs significantly from the observed data; in principle, a non-significant chi-squared statistic indicates an excellent fit, although in practice non-significance is rarely found. The comparative fit index (CFI) provides a practical index of fit that is insensitive to sample size. The standardised root mean square residual (SRMSR) indicates the discrepancy between observed and predicted covariances. Finally, the root mean square error of approximation (RMSEA) describes the degree of lack of fit in relation to the number of degrees of freedom (d.f.) in the model. Fit is commonly regarded as confirmed when the CFI is > 0.90; more stringent recommendations for excellent fit require the CFI to be ≥ 0.95, SRMSR to be ≤ 0.05 or RMSEA to be ≤ 0.05.[35] Correlational and principal components analyses were performed using SPSS Version 16 (SPSS, Inc., Chicago, IL, USA). Confirmatory factor analysis was performed using EQS Version 6.1 (Multivariate Software, Inc., Encino, CA, USA).[36]

Study 2: inter-rater reliability

Procedure

We randomly selected 40 interactions from a pool of 88 video-recorded student–SP interactions drawn from a Year 1 formative clinical communication OSCE. We provided students with verbal and written information about the study and informed them that their participation was voluntary. The university's research ethics committee granted approval and participating students gave written consent. Interactions were usually 4–5 minutes in length and comprised two different scenarios: (i) an information-gathering scenario depicting a patient who was experiencing dizzy spells, and (ii) an information-gathering and providing scenario depicting a patient with high blood pressure. We recruited the first four OSCE examiners who responded to our request for volunteer raters. Each had examined previously. They included a social worker studying for a PhD, a postgraduate researcher in clinical communication, a public health doctor and a psychologist from the university clinical communication team.

Raters viewed the videos privately on laptop computers with external speakers and were asked to rate communication as they would in a 'real' OSCE. Each received assessment sheets and accompanying descriptors of assessment domains. To approximate OSCE conditions, raters watched videos in blocks of 10, rating each as it was shown, and were unable to pause or rewind the videos.

Data analysis

Statistical analysis was conducted using SPSS Version 17.0.2 (SPSS, Inc.). To quantify reliability among examiners' total LUCAS scores, we calculated an ICC. To enable generalisation to the population of examiners, we used the second class of ICC (ICC2), in which students and raters were treated as random effects. Given that systematic variability attributable to

raters can be a major source of unreliability, we calculated the absolute-agreement ICC. Furthermore, as a typical OSCE will involve more than one examiner, we report the average measures ICC; this represents an average index across raters. Other instruments have reported reliability using generalisability (an extension of ICC methodology) ranging from 0.59 to 0.72, although without CIs. Therefore, we regarded ICC values of ≥ 0.70 as acceptable by comparison with available instruments.

RESULTS

Study 1: validity

In both datasets, each SP rating correlated with most examiner ratings (Table 1), although correlations with 'Professional verbal conduct' and 'Greeting' and 'Identity check' were generally low and, in one cohort, were non-significant.

Corrected item–total correlations are shown in Table S3. All items correlated with the total score, except item B (Identity check), which did not correlate in the summative dataset. This item was retained for two reasons: (i) it correlated with the scale total in the formative dataset, and (ii) it is considered an essential component of clinical interaction. Results of the item difficulty analysis indicate that students score more poorly on items E (Questions), F (Empathy), G (Clarification) and H (Consulting style) in both formative and summative datasets compared with other items (Table S4). The percentages scored as competent ranged from 67.58% to 96.76% across the items.

In the principal components analysis, the scree test suggested two factors, explaining 59% and 51% of the variance in the formative and summative cohorts, respectively. Factor loadings are presented in Table 2. Similar loadings were found for each cohort. Cronbach's α-values for Factor 1 were 0.78 and 0.69 for formative and summative data, respectively; Cronbach's α-values for Factor 2 were 0.86 and 0.79.

In the confirmatory analysis, we therefore first tested a two-factor model based on these results (Fig. 1). Goodness-of-fit indices satisfied the criteria for excellent fit with data for both the formative assessment ($\chi^2_{(34)} = 45.75$, $p < 0.001$, CFI = 0.99, SRMSR = 0.002, RMSEA = 0.05 [95% CI 0.00–0.051]) and the summative assessment ($\chi^2_{(34)} = 68.54$, $p < 0.001$, CFI = 0.97, SRMSR = 0.002, RMSEA = 0.05 [95% CI 0.034–0.07]). We explored a one-factor model in each dataset, in which all variables loaded on this factor, but this slightly worsened the fit (formative assessment: $\chi^2_{(35)} = 128.65$, $p < 0.0001$, CFI = 0.95, SRMSR = 0.003, RMSEA = 0.08 [95% CI 0.06–0.10]; summative assessment: $\chi^2_{(35)} = 131.00$, $p < 0.0001$, CFI = 0.90,

Table 1 Correlations between items on the Liverpool Undergraduate Communication Assessment Scale (LUCAS) and simulated patient (SP) ratings from Year 1 formative (n = 384) and summative (n = 347) clinical communication objective structured clinical examinations

	Formative SP ratings		Summative SP ratings	
LUCAS items (examiner ratings)	**Affiliation**	**Authority**	**Affiliation**	**Authority**
A Greeting and introduction	0.16*	0.26*	0.12*	0.08
B Identity check	0.26*	0.32*	0.17*	0.17*
C Audibility and clarity of speech	0.52*	0.60*	0.29*	0.37*
D Non-verbal behaviour	0.45*	0.52*	0.40*	0.51*
E Questions, prompts or explanations	0.54*	0.59*	0.52*	0.55*
F Empathy and responsiveness	0.67*	0.58*	0.57*	0.49*
G Clarification and summarising	0.48*	0.51*	0.26*	0.35*
H Consulting style and organisation	0.53*	0.60*	0.47*	0.54*
I Professional behaviour	0.46*	0.48*	0.35*	0.37*
J Professional spoken conduct	0.23*	0.29*	0.25*	0.26*

* $p < 0.01$
Affiliation = 'I felt this person was warm and caring'
Authority = 'I felt confident in this person'

Table 2 Rotated loadings of items on the Liverpool Undergraduate Communication Assessment Scale (LUCAS) (structure matrix) for Year 1 formative and summative objective structured clinical examination samples

	Factor loadings			
	Formative		Summative	
LUCAS items	Factor 1	Factor 2	Factor 1	Factor 2
A Greeting and introduction	0.09	0.93	– 0.03	0.60
B Identity check	0.41	0.32	0.08	0.67
C Audibility and clarity of speech	0.76	0.28	0.45	0.52
D Non-verbal behaviour	0.67	0.35	0.61	0.50
E Questions, prompts or explanations	0.81	0.14	0.82	0.14
F Empathy and responsiveness	0.79	–0.06	0.82	0.05
G Clarification and summarising	0.75	0.12	0.60	0.09
H Consulting style and organisation	0.81	0.20	0.79	0.22
I Professional behaviour	0.74	0.29	0.44	0.46
J Professional spoken conduct	0.56	0.25	0.23	0.56

Factor 1 = creative communication; Factor 2 = procedural communication

SRMSR = 0.003, RMSEA = 0.09 [95% CIs 0.07–0.11]). The factors were, however, highly correlated. Indicators of Factor 1 were Questioning, Empathy, Clarification and Consulting style; indicators of Factor 2 were Greeting, Identity check, Audibility and Non-verbal behaviour. All path coefficients were significant. Therefore, the raters had assessed two correlated components of communication, of which one concerned empathy and consulting style including the use of questions and clarification, and the other concerned professional and non-verbal aspects. We labelled Factor 1 'creative communication' and Factor 2 'procedural communication'. Reflecting the intercorrelation between these factors, and the utility of a single score for assessment purposes, we calculated Cronbach's α-values for the combined scale. These were 0.88 and 0.81 for formative and summative data, respectively, indicating that the scale could be used to provide a single score.

Study 2: inter-rater reliability

We used the total score when calculating reliability. The ICC2 (A, 4) was 0.73 (95% CI 0.54–0.85). This is a class 2 ICC in which all subjects (students) were evaluated by all raters using absolute agreement, average measures. This indicates fair to good reliability. As a comparison, the two-way random ICC using consistency agreement with average measures was 0.78 (95% CI 0.65–0.88).

DISCUSSION

We have described a new clinical communication assessment instrument which implements a conceptualisation of clinical communication as an endeavour that is inherently creative and which aims to assess communication holistically. The instrument demonstrated acceptable reliability and validity.

Turning first to theory, we and others have argued that communication cannot be fully described within a conceptual framework of discrete skills that universally embody 'good' communication.[3–7,26,37] Because every clinical situation is unique and practitioners need to take account of patients' individuality, we reasoned that communication is essentially creative and should be assessed holistically and flexibly. Our instrument therefore assesses communication in a way that recognises that it is contextually dependent and that meeting patients' needs requires originality rather than the observance of invariant 'rules'. It requires examiners to make judgements about how students' communication affects patients and to identify those students whose communication is likely to prevent patients from forming a sense of relationship with them. Nevertheless, like all instruments, LUCAS reflects several pragmatic compromises. In particular, it would probably have failed to gain acceptance within our medical school's skills-based curriculum if it had

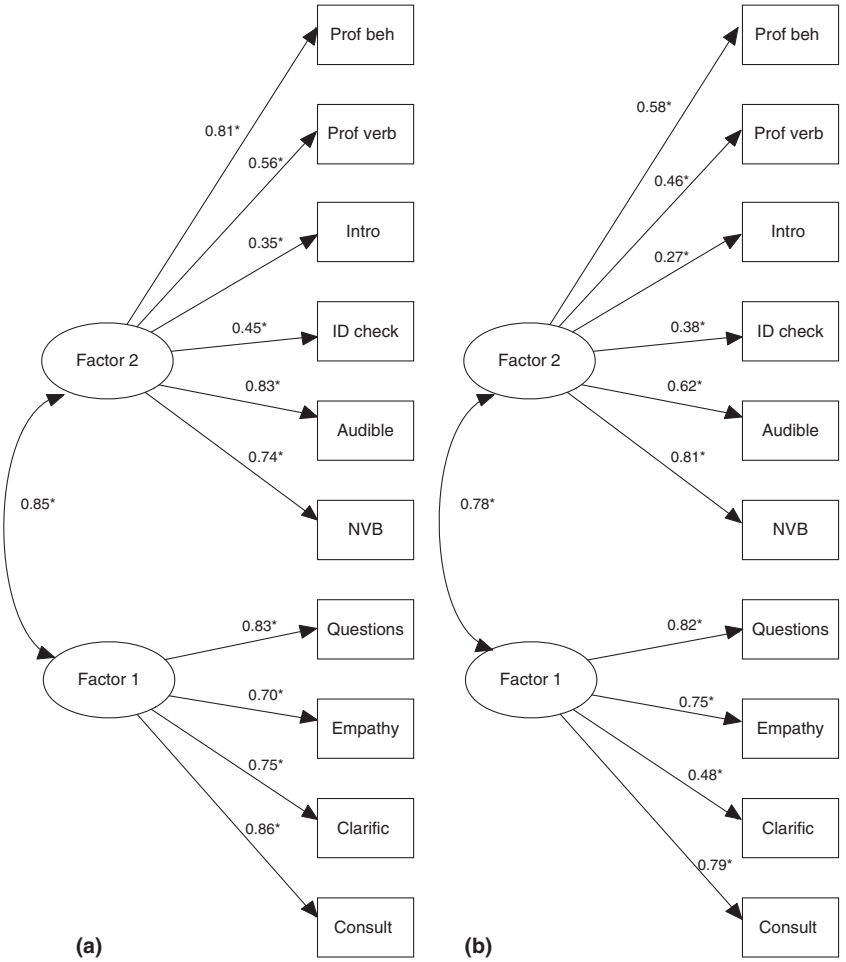

Figure 1 Results of confirmatory factor analyses based on Year 1 (a) formative (n = 384) and (b) summative (n = 347) clinical communication objective structured clinical examination data. Single-headed arrows signify causal paths; double-headed arrows signify covariance. All other paths are significant. Standardised path coefficients (β) are shown. All paths are significant. Rectangles represent observed variables (i.e. items on the Liverpool Undergraduate Communication Assessment Scale [LUCAS]); ovals represent latent variables (i.e. the underlying aspects of communication that LUCAS items are detecting). For goodness of fit, see text. Prof beh = Professional behaviour; Prof verb = Professional spoken/verbal conduct; Intro = Greeting and introduction; ID Check = Identity check; Audible = Audibility and clarity of speech; NVB = Non-verbal behaviour; Questions = Questions, prompts or explanations; Empathy = Empathy and responsiveness; Clarific = Clarification and summarising; Consult = Consulting style and organisation

diverged radically from this orthodoxy. It therefore contains items that those working within a skills framework will recognise. Students clearly need feedback that helps them to label, distinguish and reflect on different elements of communication and thus our instrument distinguishes students' competence in particular domains. Nevertheless, LUCAS is intended to move the primary focus of examiners away from the issue of whether or not students deploy specific skills to a concern with how students use skills and with the likely impact on patients. Further research scrutinising how examiners make their judgements will be important to test how successfully it achieves this aim. This will include exploring whether examiners are able to empathise with the situation portrayed in the assessment and to judge whether communication 'works', rather than simply attending to whether or not students display particular behaviours.

Comparison with other instruments is problematic because of heterogeneity in the reporting of reliability estimates; our review illustrates how a different picture can emerge when less stringent indices are used. Nevertheless, LUCAS demonstrated fair-to-good reliability, with an ICC of 0.73, and appears suitable for use in OSCEs in which pragmatic considerations such as testing time are important. Factor analysis provided support for the construct validity of LUCAS, with a two-factor solution providing excellent fit. Factor 1 (e.g. Questions, Empathy)

concerned more creative aspects of communication, whereas Factor 2 (e.g. Identity check, Audibility) concerned procedural and constrained behaviours. This supports our expectation of a separation between the procedural and the more creative components of communication.[6]

Our study has several limitations. Firstly, it is the first examination of the psychometric properties of LUCAS. Independent replication of these results is required in other institutions and contexts. Secondly, samples for both Studies 1 and 2 were drawn from a Year 1 cohort. Clinical encounters in Year 1 communication OSCEs are relatively short at 5 minutes. Further research is necessary to investigate the psychometric properties of LUCAS in higher-stakes OSCEs and across more complex and lengthier clinical encounters. Thirdly, validity in assessment arguably arises from the characteristics of examiners as much as from the properties of the scale.[37–39] Further research is necessary to investigate how examiners make their judgements and whether they are able to empathise with the patient's situation. Linked to this, although we recruited from the pool of typical University of Liverpool examiners, the backgrounds of our four raters differed, which may have contributed to unreliability among them. Reliability may have been greater if we had drawn from a more restricted cohort, as others have.[8,14] Further studies are required to investigate other sources of examiner heterogeneity, such as professional background and experience. Finally, raters in other studies received substantial amounts of training.[16] The raters in this study were not formally trained, although they had previously examined on OSCEs and were given the same information pack used to brief OSCE examiners. Further research is necessary to determine whether training improves the reliability of examiners. Recruiting a larger pool of examiners would also reduce the CIs around the point estimate.

In conclusion, this paper provides initial support for LUCAS in assessing medical undergraduates. We have set out LUCAS's conceptual basis and preliminary evidence of its psychometric properties. Additional studies will be necessary to explore the reliability and validity of LUCAS and its utility in assessing medical students in a way that reconciles the need for formal evaluation with the creative and holistic nature of communication.

Contributors: CDH, BY and PLF conceived the idea for the studies, which were then developed and designed by all authors. CDH led the implementation and all authors oversaw the conduct of the studies. CDH collected data for Study 1 and Study 2. Analysis for Study 1 was conducted by CDH with significant input from IF. Analysis for Study 2 was conducted by CDH and PS. Results were reviewed and interpreted by all authors. CDH led the drafting of the paper with significant input from BY and PS. PLF and IF critically reviewed draft revisions, suggested amendments and revised sections of the manuscript. CDH, BY and PS had full access to the data and take responsibility for data integrity and the accuracy of the data analysis. All authors commented on and approved the final manuscript. BY is the study guarantor.

Acknowledgements: none.
Funding: none.
Conflicts of interest: none.
Ethical approval: this study was approved by the Medical School Ethics Committee, University of Liverpool.

REFERENCES

1 Von Fragstein M, Silverman J, Cushing A, Quilligan S, Salisbury H, Wiskin C. UK consensus statement on the content of communication curricula in undergraduate medical education. *Med Educ* 2008;**42**:1100–7.
2 American Educational Research Association. *Standards for Educational and Psychological Testing*. Washington, DC: AERA 1999.
3 Plum A. Communication as skill – a critique and alternative proposal. *J Humanistic Psychol* 1981;**21**:3–19.
4 Salmon P, Young B. Core assumptions and research opportunities in clinical communication. *Patient Educ Couns* 2005;**58**:225–34.
5 Salmon P, Young B. Dependence and caring in clinical communication: the relevance of attachment and other theories. *Patient Educ Couns* 2009;**74**:331–8.
6 Salmon P, Young B. Creativity in clinical communication: from communication skills to skilled communication. *Med Educ* 2011;**45**:217–26.
7 Zoppi K, Epstein RM. Is communication a skill? Communication behaviours and being in relation. *Fam Med* 2002;**34**:319–24.
8 Humphris GM, Kaney S. The Liverpool brief assessment system for communication skills in the making of doctors. *Adv Health Sci Educ* 2001;**6**:69–80.
9 Kinderman P, Humphris G. Clinical communication skills teaching: the role of cognitive schemata. *Med Educ* 1995;**29**:436–42.
10 Kraan H, Crijnen A. *The Maastricht History-Taking and Advice Checklist*. Amsterdam: Lundbeck 1987.
11 Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. *Med Educ* 2003;**37**:1012–6.
12 Anastasi A. *Psychological Testing*. New York, NY: MacMillan Publishing 1976.
13 Makoul G. The SEGUE Framework for teaching and assessing communication skills. *Patient Educ Couns* 2001;**45**:23–34.
14 Thiel J, Kraan HF, van der Vleuten CPM. Reliability and feasibility of measuring medical interviewing skills: the

revised Maastricht History-Taking and Advice Checklist. *Med Educ* 1991;**25**:224–9.
15 Lang F, McCord R, Harvill L, Anderson DS. Communication assessment using the Common Ground Instrument: psychometric properties. *Fam Med* 2004;**36**:189–98.
16 Evans BJ, Stanley RO, Coman GJ, Sinnott V. Measuring medical students' communication skills: development and evaluation of an interview rating scale. *Psychol Health* 1992;**6**:213–25.
17 Stillman PL, Brown DR, Redfield DL, Sabers DL. Construct validation of the Arizona Clinical Interview Rating Scale. *Educ Psychol Meas* 1977;**37**:1031–8.
18 Yudkowsky R, Downing SM, Sandlow LJ. Developing an institution-based assessment of resident communication and interpersonal skills. *Acad Med* 2006;**81**:1115–22.
19 Iramaneerat C, Myford CM, Yudkowsky R, Lowenstein T. Evaluating the effectiveness of rating instruments for a communication skills assessment of medical residents. *Adv Health Sci Educ Theory Pract* 2009;**14**:575–94.
20 Cohen DS, Colliver JA, Marcy MS, Fried ED, Swartz MH. Psychometric properties of a standardised patient checklist and rating-scale form used to assess interpersonal and communication skills. *Acad Med* 1996;**71** (Suppl):87–9.
21 Boon H, Stewart M. Patient–physician communication assessment instruments: 1986 to 1996 in review. *Patient Educ Couns* 1998;**35**:161–76.
22 Kurtz S, Silverman J, Benson J, Draper J. Marrying content and process in clinical method teaching: enhancing the Calgary–Cambridge Guides. *Acad Med* 2003;**78**:802–9.
23 Kurtz SM, Silverman JD. The Calgary–Cambridge Referenced Observation Guides: an aid to defining the curriculum and organising the teaching in communication training programmes. *Med Educ* 1996;**30**:83–9.
24 Krupat E, Frankel R, Stein T, Irish J. The four habits coding scheme: validation of an instrument to assess clinicians' communication behaviour. *Patient Educ Couns* 2006;**62**:38–45.
25 Shrout PE, Fleiss JL. Intraclass correlations – uses in assessing rater reliability. *Psychol Bull* 1979;**86**:420–8.
26 Egener B, Cole-Kelly K. Satisfying the patient, but failing the test. *Acad Med* 2004;**79**:508–10.
27 Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. *Acad Med* 1999;**74**:1129–34.
28 Epstein RM. Making communication research matter: what do patients notice, what do patients want, and what do patients need? *Patient Educ Couns* 2006;**60**:272–8.
29 Thomsen DK, Pedersen AF, Johansen MB, Jensen AB, Zachariae R. Breast cancer patients' narratives about positive and negative communication experiences. *Acta Oncol* 2007;**46**:900–8.
30 Wright EB, Holcombe C, Salmon P. Doctors' communication of trust, care, and respect in breast cancer: qualitative study. *BMJ* 2004;**328**:864–7.
31 Mazor KM, Ockene JK, Rogers HJ, Carlin MM, Quirk ME. The relationship between checklist scores on a communication OSCE and analogue patients' perceptions of communication. *Adv Health Sci Educ* 2005;**10**:37–51.
32 Kiesler DJ, Auerbach SM. Integrating measurement of control and affiliation in studies of physician–patient interaction: the interpersonal circumplex. *Soc Sci Med* 2003;**57**:1707–22.
33 Ong LML, De Haes JCJM, Hoos AM, Lammes FB. Doctor–patient communication: a review of the literature. *Soc Sci Med* 1995;**40**:903–18.
34 Hu L-T, Bentler PM. Cut-off criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct Equ Modeling* 1999;**6**:1–55.
35 Byrne BM. *Structural Equation Modeling with AMOS*. London: Lawrence Erlbaum 2001.
36 Bentler PM. *EQS: Structural Equations Program Manual*. Encino, CA: Multivariate Software, Inc. 1995.
37 Skelton JR. *Language and Clinical Communication: This Bright Babylon*. Oxford: Radcliffe Publishing 2008.
38 Eisner EW. Artistry in education. *Scand J Educ Res* 2003;**47**:373–84.
39 Eisner EW. What can education learn from the arts about the practice of education? *Int J Educ Arts* 2004;**5**:1–12.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Table S1. Communication assessment instruments for use by examiners, including details of their design, implementation, reliability and validity.

Table S2. Communication assessment instruments for use by simulated patients, including details of their design, implementation, reliability and validity.

Table S3. Corrected item–total correlations for Year 1 formative (n = 384) and summative (n = 347) samples.

Table S4. Item difficulty indices for Year 1 formative and summative objective structured clinical examination samples, showing the percentage of responses scored as 'competent' for each item.

Appendix S1. Liverpool Undergraduate Communication Assessment Scale (LUCAS).

Appendix S2. Scoring descriptors for Liverpool Undergraduate Communication Assessment Scale (LUCAS).

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than for missing material) should be directed to the corresponding author for the article.

Received 6 April 2011; editorial comments to authors 10 June 2011; accepted for publication 25 August 2011