Title	Regulatory assessment of the consultation competence of Family Physicians in Hong Kong
Author(s)	Fraser, RC; Lee, RSY; Yiu, YK; Lam, CLK; McKinley, RK; Van der Vleuten, CPM
Citation	Hong Kong Practitioner, 2004, v. 26 n. 1, p. 5-15
Issued Date	2004
URL	http://hdl.handle.net/10722/53400
Rights	Creative Commons: Attribution 3.0 Hong Kong License

Original Article ***The Hong Kong Practitioner*** VOLUME 26 / January 2004

Regulatory assessment of the consultation competence of Family Physicians in Hong Kong

R C Fraser，R S Y Lee 李兆妍，Y K Yiu 姚玉筠，C L K Lam 林露娟，R K McKinley，C P M Van der Vleuten

Summary

Objective*: To evaluate the Consultation Skills Assessment (CSA) component of the Exit Assessment of the Higher Vocational Training Programme of the Hong Kong College of Family Physicians with particular referes' owence to content validity and reliability.*

Design: *An observational study in which candidates were directly observed and independently assessed by three assessors in the candidatn practice setting during which they were expected to consult with six unselected and consecutive patients within two hours of consulting time.*

Subjects: *Eighty-one candidates, 476 patients and 26 assessors (one external).*

Main outcome measures: *Content validity and reliability (contributions to variance and generalisability) of the overall process.*

Results: *Between 1997 and 2003, 81 clinical assessments were carried out. Internal assessors conducted a range of 1-19 assessments and the external assessor was present at 59 assessments (78.7%). The pass rate per CSA diet varied from 25-*

摘要

目的：以評估香港家庭醫學院的高級在職訓練課程終期考核中的臨床技巧部份(CSA)內容的有效性和可信度。

設計：研究以觀察形式方法進行。三名考官於考生的診療室內直接並且獨立地觀察，考生在二小時內連續 診治六名不經選擇的病人的表現。

對象：81位考生，476位病人，26個考官，其中一名來自海外。

測量內容：整個流程內容的有效性和可信度。

結果：在1997和2003之間，共有81個臨床的評估。本地考官參與了1-19個評估，而海外的考官則參與了59次評核(78.7%)。合格率分別爲 25-100%。

分數爲65%（合格分數）或以下爲46.9%而70%以上的則爲8.6%。

病人呈現的臨床挑戰包括以下七個範圍以供考官評核，醫生和病人的關係(100%評估)，病者的治療(100%)，解決問題的能力(100%)，滙集病史的能力(99.8%)，紀錄保存(99.6%)，身體檢查(97.2%)和疾病的預防(82.8%)。

88.9%的情況下每組不同考官給予考生分數皆於五個百分點裡面。在1997至2003

100%.

Overall 46.9% of candidates were allocated marks below 65% (the pass mark). 8.6% of candidates were allocated marks above 70%.

The clinical challenges presented by patients were judged by all assessors to be sufficient to enable consultation performance to be assessed across the seven LAP consultation categories as follows: Behaviour and relationship with patients (100% of consultations), Patient management (100%), Problem solving (100%), Interviewing/History taking (99.8%), Record keeping (99.6%), Physical examination (97.2%) and Anticipatory care (82.8%).

The scores independently allocated by paired assessors to any individual candidate were within five percentage points on 88.9% of occasions. Between 1997 and 2003, reliability co-efficients (G) of 0.66 and 0.73 were achieved with two or three assessors respectively assessing six consultations. The corresponding figures for 2001-2003 were 0.71 and 0.76 respectively. Thirty percent of variance was attributed to variance between candidates, 11% associated with the cases (patient challenges) and 38% the confounded effect of the case by assessors nested within candidates plus all other non-explained variance.

Conclusion: *The CSA achieves high content validity and authenticity as it uses direct observation of performance, formally validated and explicit criteria against which performance is judged, and real patient*

之間，可信係數(G)於二或三名考官的考察 6個診症情形下分別爲0.66和0.73；相對應2001-2003間分別是0.71和0.76。百分之三十的差異因爲考生的不同，11%之間的差異歸於 病人的分別，而38%則因爲考官與考生構成的因素和未知的原因有關。

結論：高臨床技巧部份(CSA)能達成高度的有效性和確實性。因爲它使用了直接觀察的方法、認可和清析的標準和真正的病人。它差一點達成評核管制公認的可信係數 0.8。要達成此目標，需要使用三個考官9次診症或者二個考官和 14次診症。這樣會造成安排上相當多的困難。利用一些策略性的方法可以減少由於考官的因素而引起得分方面的偏差。

主要詞彙：評核管制，臨床技巧，家庭醫生，可信度和有效性

HK Pract **2004;26:5-15**

challenges. It fails narrowly to achieve the recognised threshold in regulatory assessments of a reliability co-efficient of 0.8. To do so, the CSA would need to use three assessors and nine consultations, or two assessors and 14 consultations which would pose considerable logistical difficulties. Potential strategies to reduce the distorting impact on scores by the assessors have been identified.

Keywords: *Regulatory assessment; consultation competence; family physicians; reliability and validity*

R C Fraser, CBE, MD, FRCGP, FHKCFP
Professor of General Practice,
University of Leicester, UK,
External Assessor for Exit Assessment, The Hong Kong College of Family Physicians.

R S Y Lee, MBBS(HK), MPH, FHKCFP, FHKAM(Family Medicine)
Chairman,
Exit Assessment Committee, The Hong Kong College of Family Physicians.

Y K Yiu, MBBS(HK), FHKCFP, FRACGP, FHKAM(Family Medicine)
C L K Lam, MBBS(HK), FRCGP, HKAM(Family Medicine)
Past Chairmen,
Exit Assessment Committee, The Hong Kong College of Family Physicians.

R K McKinley, BSc, MD, MRCP, FRCGP
Senior Lecturer in General Practice and Director,
Clinical Consultation Research and Development Unit, University of Leicester.

C P M Van der Vleuten, PhD
Professor,
Educational Development and Research, University of Maastricht, The Netherlands.

Correspondence to :
Professor R C Fraser,
Clinical Division of General Practice and Primary Health Care, University of Leicester,
Leicester General Hospital, Leicester LE5 4PW, U.K.

Introduction

Since the core activity in clinical practice is the consultation between doctor and patient,[1] it follows that the focal point of any assessment of clinical competence must be the systematic observation and analysis of the performance of a clinician in the consultation.[2] It is for this reason that a Consultation Skills Assessment (CSA) became one of the three components (along with a Practice Assessment and a Clinical Audit Report) of the Exit Assessment (EA) of the Higher Vocational Training Programme of the Hong Kong College of Family Physicians (HKCFP).

To pass the EA, which was instituted in 1997, candidates must achieve a mark of 65% in all three components each of which is assessed separately. Although a regulatory assessment, all candidates were "volunteers" who had previously passed the Conjoint Fellowship Examination of the HKCFP and the Royal Australian College of General Practitioners. Those who passed the EA also became eligible for election to Fellowship of the Hong Kong Academy of Medicine. We now report an observational study whose aim was to evaluate the CSA with particular reference to content validity and reliability.

Methods

The procedure for the Hong Kong CSA is laid down by the Exit Assessment Committee of the HKCFP and all candidates and assessors are required to follow it.

The CSA consists of direct and independent observation by three assessors in the candidates' own practice setting during which candidates are expected to consult with six unselected and consecutive patients within a period of two hours of consulting time. Patients were required to give written consent. Twenty-five internal assessors and one external assessor (RCF) were involved.

Whenever possible - and with permission of the patient - assessors directly observed candidates conducting physical examinations and/or using instruments and personally verified claimed physical findings as appropriate. The assessment tool used was the Leicester Assessment Package (LAP).[3,4] The LAP is an integrated assessment tool which contains seven prioritised categories of consultation competence and 39 component competences (see **Box 1**), and it also has descriptive criteria to assist assessors in allocation of marks (see **Box 2**). Use of the LAP requires a consulting doctor at various stages in the consultation to answer certain questions to enable the assessors to know the reasoning which underpins his/her actions (see **Box 3**). One of the paired internal assessors asked the questions and timed the consultations.

Box 1: Competences assessed in the Consultation Skills Assessment

1. **Interviewing/history taking (Relative weighting: 20%)**
   Introduces self to patients; puts patients at ease; allows patients to elaborate presenting problem fully; listens attentively; seeks clarification of words used by patients as appropriate; phrases questions simply and clearly; uses silence appropriately; recognises patients' verbal and non-verbal cues; identifies patients' reasons for consultation; elicits relevant and specific information from patients and/or their records to help distinguish between working diagnoses; considers physical, social and psychological factors as appropriate; exhibits well-organised approach to information gathering.

2. **Physical examination (Relative weighting: 10%)**
   Performs examination and elicits physical signs correctly and sensitively; uses the instruments commonly used in the relevant clinical setting in a competent and sensitive manner.

3. **Patient management (Relative weighting: 20%)**
   Formulates management plans appropriate to findings and circumstances in collaboration with patients; makes discriminating use of investigations, referral and drug therapy; is prepared to use time appropriately; demonstrates understanding of the importance of reassurance and explanation and uses clear and understandable language; checks patients' level of understanding; arranges appropriate follow-up; attempts to modify help-seeking behaviour of patients as appropriate.

4. **Problem solving (Relative weighting: 20%)**
   Generates appropriate working diagnoses or identifies problem(s) depending on circumstances; seeks relevant and discriminating physical signs to help confirm or refute working diagnoses; correctly interprets and applies information obtained from patient records, history, physical examination and investigation; is capable of applying knowledge of basic, behavioural and clinical sciences to the identification, management and solution of patients' problems; is capable of recognising limits of personal competence and acting accordingly.

5. **Behaviour and relationship with patients (Relative weighting: 10%)**
   Maintains friendly but professional relationship with patients with due regard to the ethics of medical practice; conveys sensitivity to the needs of patients; demonstrates an awareness that the patient's attitude to the doctor (and vice versa) affects management and achievement of levels of co-operation and compliance.

6. **Anticipatory care (Relative weighting: 10%)**
   Acts on appropriate opportunities for health promotion and disease prevention; provides sufficient explanation to patients for preventive initiatives taken; sensitively attempts to enlist the co-operation of patients to promote change to healthier lifestyles.

7. **Record keeping (Relative weighting: 10%)**
   Makes accurate, legible and appropriate record of every doctor-patient contact and referral. The minimum information recorded should include date of consultation, relevant history and examination findings, any measurement carried out (e.g. BP, peak flow, weight, etc.), the diagnosis/problem (preferably 'boxed'), outline of management plan, investigations ordered and follow-up arrangements. If a prescription is issued, the name(s) of drug(s), dose, quantity provided and special precautions intimated to the patient should be recorded.

Box 2: Criteria for the allocation of marks

Marks	Descriptor of performance

85% or above	Consistently demonstrates mastery of all components:the criterion performance.
75% - 84%	Consistently demonstrates mastery of most components and capability in all.
65% - 74%	Consistently demonstrates capability in almost all components to a high standard and a satisfactory standard in all. Duration of most consultations appropriate.
55% - 64%	Demonstrates capability in most components to a satisfactory standard: demonstrates minor omissions and/or defects in some components.
45% - 54%	Demonstrates inadequacies in several components but no major omissions or defects.
44% or below	Demonstrates several major omissions and/or serious defects; clearly unacceptable standard overall.

Box 3: Questions to be asked of candidates

1. At the end of initial history taking (Candidate to inform assessor):
   - What are your diagnostic hypotheses at this stage?
   - Why have you erected these hypotheses?
   - What physical examination do you intend to carry out, and why?
2. After physical examination:
   - What did you find on examination of the patient?
   - How have these findings affected your thoughts?
3. After the patient has left:
   - Why did you choose your management plan?

In the week preceding the assessment each candidate received written instructions regarding the procedure to be followed. This was reinforced verbally immediately prior to the scheduled assessment time. A candidate was first required to give the assessors a brief verbal summary of each patient's past medical history to include the date and reason for the last consultation and to state whether the current consultation was a planned follow up or not. Since the consultations were conducted mainly in Cantonese one of the internal assessors provided the external assessor with a written account in English of the key interchanges between doctor and patient as they occurred. In order to maintain the independence of judgment of the external assessor, internal assessors were not permitted to provide opinions. The LAP questions were asked at the appropriate time (see **Box 3**) but no comments were allowed and assessors were prohibited from entering into discussion with candidates. At the end of history taking, the candidate provided the assessors with a short summary (in English) of the patient's presenting problem(s).

According to the procedures outlined, the assessors then independently allocated marks to reflect the candidate's performance in every LAP consultation category challenged in individual consultations including record keeping. The sum of the consultation category marks

represented the global performance in individual consultations. All marks for individual consultations had to be allocated before the next consultation began. If a consultation category was not challenged in a particular consultation the denominator was altered appropriately. For example, if anticipatory care was not challenged the denominator became 90 instead of 100 in that consultation.

Once the required number of consultations had been observed the allocated marks were transferred to a master mark sheet. Assessors then awarded final marks to reflect overall performance in all seven categories of competence in turn. Since assessors were required to take account of the nature and difficulty of the clinical challenges presented, the overall mark did not automatically represent the average of marks awarded for individual consultations. The final counting marks were those allocated by the two internal assessors except when one allocated a "fail" mark (i.e. below 65%) and the other a pass mark. When this occurred, the mark of the designated third assessor (usually the external assessor) became the second counting mark to determine the fate of the candidate since successful candidates had to achieve a mark of 65+% from at least two assessors in order to pass.

The internal assessors were all experienced clinicians and senior Fellows of the HKCFP and all had received formal training from RCF on the use of the LAP for regulatory purposes. All had been formally appointed as assessors by the HKCFP. Immediately prior to every diet of the EA most assessors also attended a formal briefing session. The external assessor had the ultimate responsibility for ensuring the quality and equity of the overall assessment process.

To estimate reliability, a generalisability analysis[5] was carried out with candidates: persons (P), consultations (C) and assessors (A) as factors and sources of variance. Since real (i.e. unstandardised) patients were used, different candidates were assessed dealing with differing sorts of patient challenges. Thus, case variance was nested within candidate variance (C:P). Furthermore, the same assessors were used across all cases for an individual candidate, but different assessors were allocated to different candidates (A:P). Variance components were estimated using this design and subsequently generalisability co-efficients were computed for a number of different samples of assessors and cases. (The complexity of clinical assessment procedures creates problems in both testing and determining their reliability. Although further explanation is provided in the Discussion, readers interested in more details of the educational and statistical principles underpinning the complex methodology involved can consult Fraser *et al Br J Gen Pract* 1994;44:293-296.)

Content validity was investigated by determining the extent to which the LAP categories of consultation competence were sufficiently challenged by the real patients encountered to enable the assessors to arrive at judgments of actual clinical performance of candidates.

Results

Between 1997 and 2003, 81 clinical assessments were conducted involving a total of 476 patients. In 72 assessments, six consultations were completed; in 80 assessments, five consultations were completed and in one assessment, only four consultations were completed within the permitted limit of two hours of consultation time. The assessments were performed by 25 internal assessors and one external assessor (RCF). The internal assessors conducted a range of 1-19 assessments: 12 assessors participated in five or less and six in 11 or more. Three assessors were present at 75 assessments, the third assessor was the external assessor at 59 assessments (78.7%) and on six occasions, only two (internal) assessors were present. It was decided to conduct all further analysis on the subsample of 80 candidates with 5 complete consultations rated by the counting assessors.

The mean global consultation score was 64.9% (range 53.0-75.8%, standard deviation 3.844). Thirty-eight candidates (46.9%) were allocated scores below 65% by the counting assessors and seven candidates (8.6%) were allocated scores above 70%. The pass rate per CSA diet varied from 25%-100% as it was a criterion-referenced assessment. The scores independently and individually allocated by the two paired counting assessors were identical in 12 assessments (14.8%), within two percentage points of each other in 66.7% of assessments and within five percentage points in 88.9%. The maximum difference in scores allocated to an individual candidate by the counting assessors was 11.5 percentage points. The scores allocated by the designated third assessor "counted" on 30 occasions (40%); on 21 occasions the external assessor was the counting assessor.

Table 1 sets out the contribution of each source of variance to the reliability of the marks allocated. Almost one third (30%) of all variance was attributed to the variance between candidates (P), while only 11% of variance was associated with the cases. The largest contribution (38%) was the confounded effect of the case by assessors within candidates plus all non-explained residual variance.

Table 1: Contribution of each source of variance to the reliability of the marks allocated[1]

Source of variance[2]	Estimated variance component	Standard error	Contribution (%) to variance
P	8.06	2.12	30%
C:P	3.10	0.80	11%
A:P	5.56	1.24	21%
CA:P	10.27	0.84	38%

[1] *Based on scores allocated by the two counting assessors*

[2] *P = Candidates*

C:P = Consultations nested within candidates

A:P = Assessors nested within candidates

CA:P = Consultations x assessors nested within candidates

Table 2 shows the generalisability coefficients as a function of the number of cases and assessors using the same assessors' rating across all cases for a single candidate. Using six cases, generalisability coefficients of 0.66 and 0.73 were achieved with two and three assessors respectively. Across the several years of administration slight differences were noted (not reported in **Table 2**). The corresponding figures for the 38 candidates from 1997 to 2000 and the 37 candidates from 2001 to 2003 were 0.65/0.73 and 0.71/0.76. Based on the 2001-2003 figures, to achieve a rating of 0.80 would require three assessors observing nine consultations or two assessors and 14 consultations.

Table 2: Generalisability coefficients
As a function of the number of consultations (cases) and assessors using the same assessors rating across all consultations for a single candidate

Number of consultations	Number of assessors		
	1	**2**	**3**
3	0.45	0.59	0.67
4	0.48	0.63	0.70
5	0.49	0.65	0.72
6	0.51	0.66	0.73
7	0.52	0.67	0.74
10	0.54	0.69	0.76
20	0.56	0.72	0.79

The clinical challenges presented by patients were judged by all assessors to be sufficient to enable performance to be assessed across the seven consultation categories as follows: Behaviour and relationship with patients (100% of consultations), Patient management (100%), Problem solving (100%), Interviewing/History taking (99.8%), Record keeping (99.6%), Physical examination (97.2%) and Anticipatory care (82.8%).

Each assessment consumed approximately 10.5 hours of assessor time: two hours per assessor of direct observation plus approximately one and one half hours per assessor for scrutinising the medical record,

arriving at, collating and documenting component and final mark allocations. To this has to be added travelling time to and from the candidates' health centres. Assessor training for the CSA entailed a compulsory initial seminar lasting approximately four hours and a collective briefing prior to every examination diet. Most assessors also took advantage of annual single "refresher" training seminars.

Discussion

"In high stakes (regulatory) assessments, credibility of the method is of major importance".[6] Primarily, this means that the assessment must be both valid and reliable,[7] although compromises usually have to be made in the "real world" on the grounds of feasibility.[7]

For an assessment to be valid it must measure what it is supposed to measure. In the context of the CSA, this means that the criteria against which consultation competence is judged should be professionally important and relevant, and the nature of the clinical challenges encountered should be suitable, reflecting day-to-day practice as closely as possible.[7] Furthermore, the assessment process must allow any assessor to directly observe the consultation performance of the candidate at all times.[8] The CSA in Hong Kong satisfies all these validity criteria.

When the decision was taken by the HKCFP in 1996 to adopt the LAP as the assessment tool for the CSA, the LAP criteria of consultation competence had been formally validated in the UK[9] but not in Hong Kong. Although the LAP criteria of consultation competence proved acceptable in Hong Kong,[10] a formal test of their validity in the context of Hong Kong family practice was not conducted until 2003. It is gratifying to report that the results closely replicated those in the UK.[11]

It is generally recognised that direct observation of actual performance in daily practice with real patients is the most authentic and valid approach to the assessment of consultation competence as it most closely represents real clinical practice.[6,12,13] Indeed, a systematic review[8] of published articles (1966-2001) on the validity and reliability of measures of clinical competence of physicians, medical students and residents identified the direct observation of trainees "in real-life situations" as the most valid form of assessment. Nevertheless, unlike the CSA, "few assessments observe trainees in real-life situations".[8]

Although using real patients is highly authentic, it has the potential to compromise the content validity of the clinical challenge.[14] Perhaps the optimum method for overcoming this potential problem is the use of a blueprint with specific criteria for selecting a representative sample of

patient challenges.[6] The approach used in the CSA was to record the proportion of consultations in which each of the seven LAP categories of consultation competence was deemed by the assessors to have been sufficiently challenged for them to make a judgement of the consulting doctor's performance. This condition was satisfied in over 98% of consultations for five of the seven consultation categories. In respect of the other two categories, it is well recognised and accepted that a physical examination does not need to be conducted in every consultation, and that relevant anticipatory care opportunities do not exist in all consultations. Nevertheless, it cannot be guaranteed that every candidate in the CSA encountered a totally representative sample of clinical challenges or identical degrees of difficulty. However, assessors were required to take account of the nature and difficulty of the clinical challenges in making their judgements (see **Methods**). Although the extent to which this occurred was not quantified, the difficulty of the cases has been factored into the generalisability analysis (see **below**).

"Reliability is defined as the extent to which a result reflects all possible measurements of the same construct".[5] In the CSA the "construct of interest" is the candidate's consultation competence. The analysis of variance components facilitates measurement of all the possible sources of error in the assessment process (see **Table 1**). With a reliable assessment instrument, differences in scores should reflect true differences between candidates (P). The other sources of variance in any assessment of consultation performance are the influence of the assessors (A) and the nature and difficulty of the clinical challenges (C), i.e. case specificity. Combinations of any of these sources of variance can cause distortion of true scores.

Determining true variations in candidate performance is the principal function of a regulatory assessment in order to accurately identify candidates who deserve to pass or fail. In the CSA nearly one third (30%) of all the variance to the reliability of marks allocated can be attributed to the real differences in performance between candidates. Accordingly, the CSA succeeded satisfactorily in this objective. The influence of the nature and difficulty of the patient challenge (C) and the effects of the interaction between candidates across cases (C:P) was relatively small (11%). Although both effects cannot be disentangled because of the nesting of cases within candidates, the reported influence of case specificity on performance is frequently much larger.[7] Indeed, it is accepted wisdom that "professional behaviour is highly dependent upon the nature and details of the problem being faced".[5] On the other hand, the assessor contribution to variance is considerable; both nested within candidates (A:P=21%) and in the confounded influence of the case by assessors within candidates (plus all other non-explained residual variance) (38%). Overall, the relative contribution of these variance components indicates that the CSA method is able to discriminate reasonably well in the assessment of

consultation competence of candidates, but probably requires some further sampling of cases and a reduction in the overall number of assessors used. Unfortunately, comparing variances across studies with different designs is extremely difficult if not impossible. Nevertheless, the small differences in scores independently allocated to individual candidates by counting assessors was virtually identical to those achieved in assessments of clinical competence using the Objective Structured Long Examination Record (OSLER).[15,16]

Although "the relative size of the separate sources of variation (variance) provide rich information in their own right, they can also be combined using equations .... to express the extent to which the result reflects all possible measurements of the construct of interest. The result is a fraction between zero and one called the generalisability coefficient (G). It integrates the discriminating ability of the test and the reproducibility of the result. Essentially, it provides a measure of how confident you can be that any differences detected between assessees are real differences .... simply because it takes account of all possible sources of error at the same time".[5] By mathematically modelling G in different hypothetical scenarios it is possible to estimate G with different numbers of observers or cases (analogous to the power calculation in an intervention trial).[5]

By convention, a regulatory assessment should achieve a reliability coefficient (G) of 0.8. Overall, between 1997 and 2003 the CSA narrowly failed to replicate that level as it achieved G values of 0.66 and 0.73 using six cases with two and three examiners respectively (see **Table 2**), although there was a marginal improvement in the latter three years compared to the first four years. The extent to which this would have disadvantaged or been "unfair" to individual candidates or contributed to the failure rate of 47% is impossible to quantify although there is no evidence of systematic bias. Nevertheless, to achieve a rating of 0.80 would require three assessors observing nine consultations or two assessors and 14 consultations. Both these options would pose considerable logistical difficulties for the assessors who are all volunteers and busy practising family physicians.

Ways do need to be found, however, to reduce the distortion of true scores caused by the assessors. For perfectly understandable reasons, many assessors in the CSA are infrequent participants in the assessment process because of other professional commitments. Indeed, almost half the internal assessors have participated in five or less assessments. Combined with the currently modest programme of preparation this provides insufficient opportunities for assessors to become fully familiar with the assessment tool and its application in practice. On the other hand, the potential option of restricting the CSA assessors to a small highly-trained group may not be practical because of competing professional pressures. A more practicable approach might be to provide potential assessors with initial training and then

encourage them not only to use the assessment tool for educational purposes in the intervals between regulatory diets of the CSA but also to participate in a minimum number of clinical assessments per annum. This would make assessors more familiar with the content and application of the assessment tool, which has been demonstrated to reduce subjectivity and the potential for bias in assessors.[17] In what is a criterion-referenced assessment this would also facilitate the acquisition by the assessors of a shared set of standards.

Conclusion

The Consultation Skills Assessment of the Exit Assessment of the Hong Kong College of Family Physicians is a valid approach to the assessment of consultation competence. However, the CSA marginally fails to reach the recognised threshold for reliability in regulatory assessments. Practicable strategies have been outlined to reduce the distorting impact of the assessors on the performance scores of candidates.

Acknowledgements

We thank the assessors for their enthusiastic contributions to the CSA and Ron Hoogenboon for carrying out some of the analysis.

Key messages

1. High stakes (regulatory) assessments must be valid, reliable and feasible.

2. In the assessment of consultation competence it is generally recognised that direct observation in real-life situations is the most valid method.

3. There are few reports of directly observed assessments of consultation competence in real-life situations.

4. This study demonstrates that the direct observation and assessment of the global consultation competence of family physicians in Hong Kong by two independent assessors using the Leicester Assessment Package is a valid approach and is feasible to conduct.

5. Although it marginally fails to reach the recognised threshold for reliability in regulatory assessments practicable strategies can be implemented to improve the reliability of the assessment.

References

1. Spence J. The need for understanding the individual as part of the training and function of doctors and nurses. In: *National Association for Mental Health,* Ed. *The Purpose and Practice of Medicine*. Oxford: Oxford University Press, 1960;271-280.

2. Hager P, Gonczi A, Athanasou J. General issues about assessment of competence. *Assessment and Evaluation in Higher Education* 1994;19:3-15.

3. Fraser RC, McKinley RK, Mulholland H. Assessment of consultation competence in general practice: the Leicester Assessment Package. In: Harden RM, Hart IR, Mulholland H, (eds). In: *Approaches to the Assessment of Clinical Competence. Part 1*. Dundee: Centre for Medical Education, 1992;192-198.

4. Fraser RC. *Assessment of Consultation Competence*. *The Leicester Assessment Package*. 2nd edition. Glaxo Medical Fellowship, Macclesfield, 1994.

5. Crossley J, Davies H, Humphris G, *et al*. Generalisability: a key to unlock professional assessment. *Med Educ* 2002;36:972-978.

6. Ram P, Grol R, Rethans JJ, *et al*. Assessment of general practitioners by video observation of communicative and medical performance in daily practice: issues of validity, reliability and feasibility. *Med Educ* 1999;33:447-454.

7. Van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. *Advances in Health Sciences Education* 1996;1:41-67.

8. Epstein RM, Hundert EM. Defining and assessing professional competence. *JAMA* 2002;287(2):226-235.

9. Fraser RC, McKinley RK, Mulholland H. Consultation competence in general practice: establishing the face validity of prioritised criteria in the Leicester assessment package. *Br J Gen Pract* 1994;44:109-113.

10. Fraser RC. Consultation strengths and weaknesses of Hong Kong family physicians: an external view. *HK Pract* 1999;21:561-562.

11. Lau J, Fraser RC, Lam CLK. Establishing the content validity in Hong Kong of the prioritised criteria of consultation competence in the Leicester Assessment Package. *HK Pract* 2003;25:596-602.

12. Campbell LM, Howie JGR, Murray TS. Use of videotaped consultations in summative assessment of trainees in general practice. *Br J Gen Pract* 1995;45:137-141.

13. Norcini J. The death of the long case? *BMJ* 2002;324:408-409.

14. Caulford PG, Lamb SB, Kaigas TB, *et al*. Physician incompetence: specific problems and predictors. *Acad Med* 1994;69Suppl(10):16-18.

15. Gleeson F. AMEE Medical Education Guide No 9. Assessment of clinical competence using the Objective Structured Long Examination Record (OSLER). *Med Teacher* 1997;19:1:7-14.

16. Gleeson F. Personal communication. 1994.

17. Kane MT. The assessment of professional competence. *Evaluation and the*

Health Professions 1992;15:163-182.
