Reproducibility of clinical performance assessment in practice using incognito standardized patients

Simone Gorter,[1] *Jan-Joost Rethans,*[2] *Désirée van der Heijde,*[1] *Albert Scherpbier,*[3] *Harry Houben,*[4] *Cees van der Vleuten*[5] *& Sjef van der Linden*[1]

Background The reproducibility of authentic assessment methods has been investigated for objective structured clinical examinations (OSCEs) and video assessment in general practice, but not for assessment with incognito standardized patients.

Purpose To investigate the reproducibility of assessment with incognito standardized patients.

Methods A total of 27 Dutch rheumatologists in 16 hospitals were each visited by 8 incognito standardized patients presenting with different rheumatological disorders. After each visit, the standardized patient completed a case-specific checklist containing items on medical history, physical examination and management. Over a 20-month period, 254 incognito visits took place, of which 201 were first visits. The standardized patient was detected by the rheumatologist in 2 cases only. These encounters were not included in the analysis. Generalizability theory was used to investigate the reproducibility of the assessment.

Results One fifth of the variance can be attributed to variation between rheumatologists. The largest variance is due to the variation in difficulty among cases. A reproducible assessment requires 3 hours of testing time (6 cases) if it is obtained through a norm-referenced interpretation of scores and 7 hours of testing time (14 cases) if it is obtained through an absolute interpretation of scores.

Conclusion The reproducibility of performance assessment in clinical practice by incognito standardized patients is similar to that of other authentic measurements for the assessment of clinical competence and performance.

Keywords Education, medical, postgraduate/*methods; clinical competence/*standards; educational assessment; reproducibility of results; rheumatology/standards; Netherlands.

Medical Education 2002;36:827–832

Introduction

Research and development concerning assessment of clinical competence have been characterized by increasing efforts to design methods of authentic assessment.[1] The success of the objective structured clinical examination (OSCE) and its widespread use are due to the combination of authenticity and standardized test-taking conditions.[2] The OSCE, however, provides an indirect measure of performance only, because it involves *simulated* clinical encounters and represents an abstraction from the real clinical situation.[3] It still requires extrapolation from test conditions to those of the real world.[4] In terms of Miller's competence pyramid, the OSCE does not bridge the gap between 'showing how' under artificial test conditions and actually 'doing' in daily clinical practice.[5] There is evidence of discrepancies between 'competence' ('what a doctor is capable of doing') and 'performance' ('what a doctor does in daily practice').[6] Apparently, even the authentic OSCE requires extrapolation from test to real world conditions, thereby detracting from the validity of the measurement. Less extrapolation would be necessary if assessment could be performed by unobtrusive or unnoticed direct observation and measurement of doctors' performances in daily practice.

[1]Division of Rheumatology, Department of Internal Medicine, University Hospital Maastricht, The Netherlands
[2]Skillslab, Institute for Medical Education, Faculty of Medicine, Maastricht University, The Netherlands
[3]Institute for Medical Education, Faculty of Medicine, Maastricht University, The Netherlands
[4]Department of Rheumatology, Atrium Medical Centre Heerlen, The Netherlands
[5]Department of Educational Development and Research, Faculty of Medicine, Maastricht University, The Netherlands

Correspondence: Dr Jan-Joost Rethans, Skillslab, Institute for Medical Education, Faculty of Medicine, Maastricht University, PO Box: 616, 6200 MD Maastricht, The Netherlands. Tel.: 00 31 43 388 1790; Fax: 00 31 43 388 4127; E-mail: j.rethans@sk.unimaas.nl

Key learning points

Assessment of clinical performance by incognito standardized patients is a highly authentic form of performance assessment.

Assessment of doctors' performance in daily clinical practice by incognito standardized patients yields equal reproducibility to other authentic measures of competence and performance, such as OSCEs.

Three hours of testing time (6 cases) are needed for a reproducible assessment with a norm-referenced interpretation of scores.

Seven hours of testing time (14 cases) are required for a reproducible assessment with an absolute score interpretation.

The potential difficulty in assessment in daily clinical practice lies in the inherent non-standardized nature of test conditions. Patient case-mix poses an almost insurmountable difficulty with assessment in actual clinical practice. It has been identified as one of the reasons for the hopeless unreliability of any conventional global rating by supervisors who evaluate actual clinical performance.[7,8] The lesson to be learned from all psychometric evaluations of clinical competence measurements is that sufficient sampling is necessary to achieve reproducible scores.[9] It has consistently been demonstrated that the clinical context (case content/patient) is a dominant factor which affects the reproducibility of the measurement. This means that any type of measurement will require a large sample across clinical contexts. Other factors that affect reproducibility, such as examiner variability or variability caused by different patients playing the same role, are either less influential or can be neutralized through effective sampling in an efficient test design.[10] Even the 'objective' OSCE requires a large sample of stations, i.e. at least 4 hours of testing time, before minimally reproducible scores are obtained.[11] We therefore estimated that, with an appropriate sampling strategy, authentic assessment in clinical practice might well yield reproducible measurements.[12] This is supported by recent evidence from a study by Ram and colleagues.[13] In a study of video assessment of a random sample of 16 taped, non-standardized consultations in general practice, Ram *et al.* obtained reproducibility coefficients of about 0·80. Each consultation was rated by one rater and a different observer. Testing time was approximately 3 hours.

Another direct method for assessing performance in clinical practice is the incognito standardized patient method.[14,15] This method makes use of standardized patients (SPs) who consult a doctor as if they were a real patient. Although the doctor is informed that an SP may show up at some time, he or she does not know when to expect them. The SPs are extensively trained to portray their clinical scenario in a highly standardized way. They have also been trained to score the doctor's performance on a checklist after the encounter. Incognito SP-based assessment provides a highly authentic evaluation of performance in daily practice. The standardization of the SP role may suggest some degree of artificiality, but studies have shown that doctors are unable to distinguish SPs from real patients.[16] Not surprisingly, studies using incognito SPs typically report very low detection rates.[17,18,20]

In this study we investigated the reproducibility of incognito SP-based assessment. So far, studies on the reproducibility of this method have focused on SP consistency or SP accuracy in role-playing.[21,22] To our knowledge, there are no published studies reporting an analysis of the overall reproducibility of incognito SP-based assessment involving repeated visits by different SPs portraying different cases. In light of the above discussion on the reproducibility of authentic methods, and the encouraging results of the video assessment method, we set out to determine the reproducibility of the incognito SP-based method. This study reports an analysis of a data set obtained from 22 Dutch rheumatologists, each of whom saw 8 different incognito SPs presenting with different clinical conditions. Each case was presented by one of 2 SPs who were specially trained to portray a specific case.

Method

Subjects

Out of a total of 127 Dutch rheumatologists, 116 were asked to participate. The remaining 11 rheumatologists were involved in the development of the study. They either participated in the development of the checklists for rating the visits or in the preparatory training sessions with SPs. A total of 57 (49%) of the invited rheumatologists gave written consent. For logistical reasons, 27 rheumatologists, spread throughout the Netherlands, were selected for participation in the study. Consent was obtained from the boards of the hospitals where the participating rheumatologists practised. A total of 22 of the 27 rheumatologists were each visited by 8 incognito SPs. Details of rheumatologists' characteristics have been published elsewhere.[19]

Instruments

Eight patient roles were developed (Table 1). Each case was based on a real patient who had presented in an outpatient ward of a university hospital. Sixteen SPs were recruited and each role was portrayed by 2 SPs. Some SPs were real patients who presented their own stable rheumatological condition. All SPs were extensively trained to play their role consistently and to score the rheumatologist's performance accurately on a case-specific checklist. The face validity of the presentations of the cases by the SPs was judged to be sufficient by 4 rheumatologists. The SPs practised checklist scoring using methods described elsewhere[20] until an acceptable level of accuracy (85% agreement) was achieved. Specific checklists were developed for each case by a panel of 11 rheumatologists. The number of items varied from 50 to 75. The items asked for information on the medical content of the visit. Each list included items on history taking, physical examination and management. The panel also identified key items considered to be essential components of a rheumatologist's consultation on a particular case. On average, the key items accounted for 55% of all items on a checklist. The SPs completed the checklist immediately after the consultation.

Procedure

Each role was played by 2 SPs. The SPs were randomly allocated to the rheumatologists. No rheumatologist was visited by the same SP twice, unless an appointment had been made for a follow-up visit. General practitioners from the areas of the hospitals attended by the SPs participated in the study by writing referral letters to the rheumatologists for the SPs. The participating rheumatologists were asked to complete a detection form if they suspected a patient of being an SP. We assumed validity as long as the SPs remained undetected. Arrangements were made with the participating hospitals to ensure that SPs remained under cover. Results of additional investigations ordered for SPs by the rheumatologist, such as lab tests and reports of radiological investigations, as well as real radiographs were simulated and care was taken that the rheumatologist received these results in the manner customary for that hospital. A detailed description of the study methods and the feasibility of introducing incognito SPs in secondary care has been published elsewhere.[20] In 2 cases, SPs were unmasked. In both cases this was due to administrative mistakes rather than non-authentic patient portrayal. Appointments for follow-up visits were made in 53 consultations. The patient concerned kept the appointment and completed one checklist for both visits. In this way a total of 254 visits took place over the period between July 1998 and February 2000. It was estimated through logbook analysis that a consultation took about 30 minutes.

Table 1 Mean percentage scores and standard deviations of the rheumatologists' ($n = 22$) performances in each of the 8 cases

	Case	
	Mean	Standard deviation
Lateral epicondylitis	49	16
Fibromyalgia	63	12
Ankylosing spondylitis	54	10
Polymyalgia rheumatica	73	10
Rheumatoid arthritis	34	9·4
Haemochromatosis arthropathy	65	17
Psoriatic arthritis	52	8·5
Osteoporosis	46	9·8

Statistics/analysis

For each case, the number of actions performed by the rheumatologist was scored on the predefined case-specific checklist. This number was expressed for each encounter as a percentage of the total number of actions listed on the checklist. This process was also carried out for essential key items. In this way, 2 percentage scores were calculated for every rheumatologist for each of 8 cases. The correlation between the overall score and the essential score was 0·97. As this would inevitably yield very similar reproducibility estimates, it was decided to report the overall score only. The average case score was the mean score for an individual rheumatologist across 8 cases.

We investigated the level of reproducibility using generalizability theory.[23] A simple, all-random person-by-case design was used for estimating variance components. The design may have been slightly biased due to the fact that each role was played by 2 SPs. Differences between SPs may have inflated the variance of the 'person' component. This confounding effect was not preventable. Reproducibility coefficients were estimated from the variance components using a norm-referenced perspective (generalizability coefficients) and a domain-referenced perspective (dependability coefficients) as functions of the number of cases. From a norm-referenced perspective, scores are valued relative to each other; from a domain-referenced perspective, scores are interpreted in an absolute manner. A

reproducibility coefficient can be interpreted as the expected correlation with a hypothetical other measurement involving a random sample of different patients and cases, assuming that the scores are valued relative to each other (generalizability coefficient) or relative to an absolute standard (dependability coefficient).

Results

Due to unforeseen logistical problems (long waiting lists, temporary stops on new patients, absence of referral letters, etc.), some consultations were unavailable for analysis. The results of all cases seen by 22 of the rheumatologists were available and these were included in the analysis. Therefore, 176 (8 cases for each of the 22 rheumatologists) of the 201 first visits were included in the analysis. As results of follow-up visits were supplementary to those of first visits, the first and related follow-up visit were analysed as one visit. Table 1 provides descriptive statistical data on the total scores per case.

Table 2 reports the estimated variance components. Most of the variance is associated with cases.

Table 1 shows that the mean number of items performed across cases varies considerably. The various clinical problems apparently present quite different challenges to individual rheumatologists. The second largest variance component is the general error term, which represents approximately one quarter of the total variance. Usually, this component is the largest in clinical competence measurements. The remaining one fifth of the variance is attributable to the variance between rheumatologists. This indicates the ability of the measurement to differentiate between doctors. As the purpose of the instrument is to differentiate between rheumatologists, this is considered a desirable variance. Although it is, as usual, the smallest component, it cannot be called small in an absolute sense. The general error term (PxC, error) represents undesirable variance, or error variance, both in the norm-referenced and the domain-referenced score interpretations. From the domain-referenced perspective, the case variance is also considered part of the error variance. The standard errors are quite large for all variance components. This is due to the relatively small sample sizes on which these estimations are based (22 doctors and 8 cases).

Table 3 reports the generalizability and dependability coefficients as a function of a (hypothetical) number of cases. There is a sizeable difference between the 2 types of reproducibility coefficients. This is due to the large case variance component, which is entered in the dependability coefficients. The sample size of 8 cases used in this study produced values of 0·86 and 0·69, respectively. If an arbitrary value of 0·80 is regarded as a minimum reproducibility coefficient, 6 cases may suffice for norm-referenced scores, whereas about 14 cases are needed for an absolute score interpretation.

Table 2 Estimated variance components, standard errors and percentage of total variance

Source of variance	Estimated variance component	Standard error	Percentage of total variance
Persons (P)	74·14	25·82	19·17
Cases (C)	206·57	99·65	53·42
P × C, error	105·96	12·28	27·40

Discussion

This study has 2 methodological weaknesses. The first concerns the use of 2 different SPs for each case. Differences between rheumatologists due to differences between these SPs are erroneously taken as either person variance or desirable variance. This could not be prevented for practical reasons. However, the inflation is likely to be small since the SPs were trained to portray their role and score the checklist very accurately and in the same way. It has been shown that differences between SPs do not statistically interfere with doctors' performances.[21]

The second problem is more serious and concerns the small sample size of the doctors. Naturally, this originates from the huge logistical demands of this study. As a consequence of the small sample size,

Table 3 Reproducibility coefficients using a norm-referenced (generalizability coefficients) and a domain-referenced score interpretation (dependability coefficients) as functions of the number of cases and estimated testing time

Number of cases	Estimated testing time (in hours)	Generalizability coefficient	Dependability coefficient
2	1	0·61	0·35
4	2	0·76	0·52
6	3	0·82	0·62
8	4	0·86	0·69
10	5	0·89	0·73
12	6	0·90	0·77
14	7	0·92	0·79
16	8	0·93	0·81

standard errors of the estimated variance components are sizeable and the data should be interpreted with some caution.

The most notable finding in relation to the variance components obtained was the large variance caused by the differences between the cases in terms of the extent of the challenge posed to doctors. This may be due to the sampling problem mentioned. Usually, the general error term is larger than the variance associated with cases. The person component was relatively large. As a result, the reproducibility coefficients were quite acceptable. With a relative score interpretation, a relatively small sample of approximately 6 cases suffices to achieve a reproducibility coefficient of 0·80. Many more cases are needed for an absolute interpretation of the scores. Naturally, when the cases present considerable differences in degree of difficulty, as in this study, a large sample of cases is required in order to achieve a generalization of the results of performance on random samples of other cases. Our data indicated that approximately 14 cases would be needed to obtain a reproducibility coefficient of 0·80.

A more salient finding than the difference between the 2 score interpretations and their effect concerns the overall level of reproducibility found in this study. The purpose of this study was to investigate whether this direct performance assessment method could yield reproducible findings. The answer to this question is affirmative, provided the sample of cases is large enough.

We also wanted to know how the reproducibility of this method compares to that of other methods. When we compare the reproducibility coefficients at comparable testing times, the values found with this method are by no means worse than the values reported in OSCE assessment.[11] Furthermore, the norm-referenced reproducibility coefficients appear to be considerably better. In order to compare our method with the practice video assessment format, we compared our dependability coefficients with the values reported by Ram *et al.*[13] With the video assessment instrument, real patients were randomly used, all cases were dissimilar and case difficulty was automatically included in the error term. Our reproducibility data are slightly worse at comparable testing times. However, we wish to emphasize that these reproducibility coefficients should not be interpreted too absolutely, because of the noise in the estimates. Finding of comparability between the results is much more compelling than finding of observed differences.

The information collected in our study related only to the medical content of the consultation and did not include attitudinal and communication skills. However, we think that data on attitudinal and communication skills would yield a much more consistent pattern among different rheumatologists and might therefore result in even better reproducibility results.

Overall, this study has demonstrated that the method of assessing authentic practice performance using incognito SPs can yield scores that are as reproducible as those of other authentic measures of competence and performance. With a sufficiently large sample of measurements, reproducible measurement can be achieved with direct assessment of performance in daily clinical practice.

Contributors

SG acted as main researcher for this project. J-JR assisted in practical aspects of the study and in writing this paper. DvdH participated in the development of the project and in training of SPs. AS actively participated in writing this paper. SvdL actively participated in the development of the project and in training of SPs. HH participated in the development of the study. CvdV participated in statistical analysis and in writing this paper.

Funding

This study was subsidized by the Dutch Arthritis Association (grant no: 97-2-201).

References

1 Van der Vleuten CPM. The assessment of professional *competence: Developments, research and practical implications. Adv Health Sci Educ* 1996;**1**:41–67.
2 Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). *Med Educ* 1979;**13**:41–54.
3 Rethans J-J, Westin S, Hays RH. Methods for quality assessment in general practice. *Family Prac* 1996;**13**:468–76.
4 Kane MT. The assessment of professional competence. *Eval Health Profess* 1992;**15**:163–82.
5 Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990;**65**:S63–7.
6 Rethans J-J, Sturmans F, Drop R, van der Vleuten C, Hobus P. Does competence of general practitioners predict performance? Comparison between examination setting and actual practice. *BMJ* 1991;**303**:1377–80.
7 Streiner C. Clinical ratings – ward rating. In: Shannon S, Norman G, eds. *Evaluation Methods: A Resource Handbook.* Hamilton; 1995;29–32.
8 Gray JD. Global rating scales in residency education. *Acad Med* 1996;**21** (Suppl.):S55–S63.

9 Swanson DB. A measurement framework for performance-based tests. In: Hart I, Harden R, eds. *Further Developments in Assessing Clinical Competence*. Montreal: Can-Heal Publications; 1987;13–45.
10 Van der Vleuten CPM, Norman GR, De Graaff E. Pitfalls in the pursuit of objectivity: Issues of reliability. *Med Educ* 1991;**25**:110–8.
11 Van der Vleuten CPMD, Swanson D. Assessment of clinical skills with standardized patients: State of the art. *Teaching Learning Med* 1990;**2**:58–76.
12 Vu NV, Barrows HS. Use of standardized patients in clinical assessments: Recent developments and measurement findings. *Educ Res* 1994;**23**:23–30.
13 Ram P, Grol R, Rethans J-J, Schouten B, van der Vleuten CPM, Kester A. 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**:447–54.
14 Rethans J-J, Sturmans F, Drop MJ, van der Vleuten CPM. Assessment of performance in actual practice of general practitioners. *Br J General Pract* 1991;**41**:97–9.
15 Tamblyn RM, Berkson L, Dauphinee D, Gayton D, Huang A, Isaac L *et al.* Unnecessary prescribing of NSAIDS and the management. *Ann Intern Med* 1997;**127**:429–38.
16 Norman GR, Tugwell P, Feightner JW. A comparison of resident performance on real and simulated patients. *J Med Educ* 1982;**57**:708–15.
17 Rethans J-J. Needs assessment in continuing medical education through standardized patients. *J Cont Educ Health Prof* 1998;**18**:172–8.
18 Beullens J, Rethans J-J, Goudhuys J, Buntinx F. The use of standardized patients in research in general practice. *Family Pract* 1997;**14**:58–62.
19 Gorter S, van der Linden S, Brauer J, van der Heijde D, Houben H, Rethans J-J *et al.* Rheumatologists' performance in daily practice. *Arthritis Care Res* 2001;**45**:16–27.
20 Gorter S, Rethans J-J, Scherpbier A, van der Linden S, van Santen-Hoeufft M, van der Heijde D *et al.* How to introduce incognito standardized patients into outpatient clinics of specialists in rheumatology. *Med Teach* 2001;**23**:138–44.
21 Tamblyn RM, Grad R, Gayton D, Petrella L, Reid T. Impact of inaccuracies in standardized patient portrayal and reporting on physician performance during blinded visits. *Teaching Learning Med* 1997;**9**:25–38.
22 Vu NV, Marcy MM, Colliver JA, Verhulst SJ, Travis TA, Barrows HS. Standardized (simulated) patients' accuracy in recording clinical performance checklist items. *Med Educ* 1992;**26**:99–104.
23 Brennan RL. *Elements of Generalizability Theory*. Iowa: ACT Publications; 1983.

Received 31 *January* 2001; *editorial comments to authors* 19 *April* 2001, 7 *February* 2002; *accepted for publication* 22 *March* 2002