Validity of objective tests towards peer-rated competence by students

ROB DIJCKS, KATINKA J.A.H. PRINCE, Cees P.M. VAN DER VLEUTEN & ALBERT J.J.A. SCHERPBIER
Faculty of Medicine, Maastricht University, The Netherlands

SUMMARY The purpose of the study was to investigate the predictive value of written longitudinal tests (progress tests), block-related written tests and OSCEs in relation to peer-rated competence of students. Competence was rated for functioning as a student and for functioning as a future physician. Eight students were asked to rate their peers on both aspects of competence. This was related to past performance on four progress tests, eight block tests, and two OSCEs. Observed and true correlations were estimated and a discriminant analysis was performed. Both written tests fairly accurately predicted functioning as a student. Peer-rated competence as a future doctor was moderately predicted by the OSCE. Intercorrelations among the various test methods were theoretically interpretable suggesting construct validity of the objective tests used.

Introduction

Assessment procedures are continuously critically monitored by students, particularly in educational settings where students are actively organized and involved in salaried positions to represent students on educational committees such as in the Maastricht medical school (Visser et al., 1998). Periodically, the Maastricht assessment programme comes under attack. On one such occasion students argued that the results of the objective tests conflicted with their own perspective on the capabilities of their peers. These objective tests include the progress test (a longitudinal written test assessing general medical knowledge), written end-of-unit tests assessing course-related knowledge, and an annual OSCE assessing clinical skills. All these tests are used summatively to decide on student promotion. In particular the students felt that certain very bright or poor students were not identified by the assessment programme. They also thought that some of their fellow students were unlikely to make good doctors and that the assessment programme failed to detect this. This was reason for us to investigate their concern. We therefore related the scores of the objective tests to ratings of their colleagues’ competence by a group of students.

The concern of the students represented a practical reason to undertake this study. From a theoretical perspective, this study questions the predictive validity of our assessment procedures. Are our instruments able to predict who performs well and who does not? And do some instruments do better than others? Validity research always requires a criterion or gold standard and no single criterion will ever be perfect (Norman et al., 1996). In this case there was concern raised by the students based on their own perceptions of their peers’ functioning. It might be argued that this is a valid criterion, since students interact with each other frequently in an unobtrusive non-threatening way, particularly in a problem-based curriculum like that of Maastricht medical school, where students work mostly in small groups in which they need to collaborate and are dependent on each other. The composition of the small groups changes every six-week block period. Students also work in different small groups in the laboratory, for communication skills training and in clinical skills training sessions in the skills laboratory. Although clinical attachments do not start until the end of year 4, students do have patient contact both alone and in small groups, where they can observe their peers interacting with patients. The threat to the validity of the criterion naturally depends on the accuracy of how students perceive their peers’ competence (Dochy et al., 1999). This will also depend on the sample of peer assessments. We therefore used a group of students to judge the competence of their peers. The research question thus becomes: How well do our objective assessment instruments predict the combined peer ratings of student competence?

Method

Instruments

Progress test: This test is administered four times per year and covers all basic, clinical and behavioural disciplines relevant to undergraduate medicine. It consists of 250 questions of the multiple true/false type with an ‘I do not know’ option. Each test consists of strictly reviewed newly constructed items and the content is balanced through a test blueprint. Students cannot prepare for the test. The progress test is meant as a periodic unobtrusive general knowledge assessment. A student’s test score is the number of correct items (+1 score) minus incorrect items (−1 score) expressed as a percentage of the total score (formula scoring). More information about the progress test can be found in another publication (Van der Vleuten et al., 1996).

Block test: At the end of each unit or block in the problem-based curriculum a written test is administered covering the content of the block. In most cases it is of the same item format as the progress test (with approximately 150 to 180 items), but without the question-mark option. The percentage of correctly scored items gives the final student score. Some units use essay-type questions; test scores here consist

Correspondence: Prof. Dr. A.J.J.A. Scherpierz, Institute for Medical Education, PO Box 616, 6200 MD Maastricht, The Netherlands. Tel:+31433881815; fax: +31 433884165; email: a.scherpbier@oifdg.unimaas.nl
of the total number of marks expressed again as a percentage score. Test items are newly constructed and reviewed by a panel of teachers prior to test administration.

**OSCE**: At the end of the academic year an OSCE is administered assessing the clinical skills that are considered relevant up till that year of training. All stations require ‘hands-on’ performance. Depending on the content, station duration varies from 10 to 30 minutes and total testing time ranges from two to three hours. Examiners mark performance on checklists. A student score per station is the percentage correct across items of the checklist; the test score is the average across all stations.

**Procedure**

**Dependent variable**: Eight fourth-year medical students (from a six-year programme) were asked to independently judge their peers, excluding themselves \( n = 212 \). They were asked to rate peers on: (1) competence as a student, and (2) competence as a future doctor. The rating included four options: (1) ‘poor’, (2) ‘average’, (3) ‘good’, (4) ‘do not know’. An intraclass correlation treating differences between judges in rank ordering as error yielded a value of 0.68 for the first item and 0.66 for the second item. A final classification was computed by using only the data of students with the same classification given by five or more judges. Students’ final classifications into ‘do not know’ were dropped.

**Independent variables**: For each of the test formats past performance was aggregated for each student. Four progress test scores were used from the preceding third year. Scores were transformed into standard \( z \)-scores before being averaged to a single average progress test score. The reliability, using Cronbach’s alpha, was 0.88 taking the four test scores as ‘items’ for the analysis. For the block tests six previous tests from the third year were averaged using a similar procedure. The reliability was also 0.88. Finally, OSCE scores were obtained from the last two years of training and aggregated. The reliability of these scores was 0.61.

**Statistical analysis**

First, Pearson bivariate correlations were computed between all variables. To explore the theoretical associations between the variables, assuming perfect reliability of the measures, disattenuated correlations (true correlations) were computed using the above-mentioned reliability indices. The true correlation between the two ratings was assumed to be identical to the observed correlation, because the comparison was based on the same judges. Finally, a discriminant analysis was used with the two competence ratings as a dependent variable and the progress test, block test and OSCE scores as independent variables. The percentage of correctly classified students, i.e. the agreement between the classifications derived from the peer evaluations and the predicted classification derived from the test scores, is reported. Because the students argued that the tests were particularly insensitive to identifying good and weak students, the discriminant analyses were repeated for a subset of the data whereby the middle category ‘average’ was removed.

**Results**

After excluding all students with fewer than five of the eight judges agreeing, 128 students remained (60%). The test scores from this remaining sample did not differ significantly from those of the total year group. Table 1 reports the frequencies of classifications. It is interesting to note that within the subset used another quarter to one-third is indifferently classified as ‘do not know’. Apparently, on a large proportion of the students the student judges either disagree (84 or 40% of total class of 212) or are indifferent (32 and 33, 15 and 16% of total class respectively), in total constituting nearly half of the class. Only a small percentage is identified as being a poor student and future doctor (6 and 4% respectively) while a larger percentage is classified as being a good student and future doctor (13 and 17% respectively).

Table 2 reports the bivariate correlations. All correlations are statistically significant but more interesting is the pattern of associations, particularly of the true correlations. Both competence ratings appear highly intercorrelated but do not reach unity. Apparently, there are some differences between the rating of being a good student and being a good future doctor. The best univariate predictor for competence as a student is the block test, followed by a near equal correlation by the progress test. This is not surprising because the intercorrelation between the block and progress test is also high. For competence as a future doctor all correlations are lower and the best predictor now becomes the OSCE. Finally, it is interesting to note that the association between the block test and the OSCE is higher than the association between the progress test and the OSCE.

The discriminant analysis classified 63% of the students correctly for the competence rating as a student, with the progress test and skills test being identified as significant predictors. For functioning as a future doctor only the OSCE was identified as a significant predictor and the percentage correctly predicted was 45%. Omitting the middle classification ‘average’ and using the extremes ‘poor’ and ‘good’ only, these percentages were respectively 87% (this time progress test predictor significant only) and 87% (OSCE only).

**Discussion**

The peer rating yielded quite a few indifferent outcomes. Approximately 50% of the students were either classified as ‘do not know’ or there was substantial disagreement between the judges. Although this is sizeable, it is not too surprising. It is not to be expected that student judges should know all of

<table>
<thead>
<tr>
<th>Classification</th>
<th>Competence as a student</th>
<th>Competence as a future doctor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poor</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Average</td>
<td>72</td>
<td>56</td>
</tr>
<tr>
<td>Good</td>
<td>16</td>
<td>13</td>
</tr>
<tr>
<td>Do not know</td>
<td>32</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>128</td>
<td>100</td>
</tr>
</tbody>
</table>
are able to discriminate between students with poor and good performance. In that respect they demonstrate predictive validity. Moreover, the patterns of intercorrelations are theoretically meaningful.

One could justifiably wonder why both block tests and progress tests need to be maintained when they are so highly correlated. First, the data indicated that the unique information from the progress test is related to its ‘unobtrusive’ nature. Second, feedback is an important function of any form of assessment and both types of tests provide different sources of information: the progress test gives longitudinal broad-based feedback across all disciplines, whereas the block test gives specific course-related information. Finally, one could also agree with the argument and use a progress test as a broad periodic cognitive assessment and design block-related evaluations that are less knowledge oriented but which tap into other non-factual or non-cognitive skills.

This study has several drawbacks. First, the sample of student ratings was rather low. With a larger sample of approximately 20 students we would have obtained better reliability and more accurate identification of student functioning. Second, the total sample on which the correlations are based excluded both the students with poor inter-judge agreement and the students classified as ‘do not know’. Therefore the subsample was only part of the original total class of fourth-year students. Although non-significant differences were found between students within and outside the analysis, some sampling bias may have occurred and the estimation of the size of the associations may be less accurate than when a larger sample had been used.

Finally, validity studies have bi-directional implications. Peer assessment was used here as the gold standard. Reversing the relationship and taking the objective tests as criterion, the peer assessment of the students was valid. The restriction seems to be—as with all measures of competence—to obtain a large enough sample of ratings to allow for reliable inferences.

**Notes on contributors**

**Rob Dijcks** is a medical student and research assistant at the Skillslab, Faculty of Medicine, Maastricht University.

**Katinka J.A.H. Prince, MD** is a General Practitioner in the Skillslab, Faculty of Medicine, Maastricht University.

**Cees P.M. van der Vleuten, PhD** is Chair of the Department of Educational Development and Research, Faculty of Medicine, Maastricht University.
ALBERT J.J.A. SCHEPBIER, MD PhD is Scientific Director of the Institute for Medical Education, Faculty of Medicine, Maastricht University.

References


