

# Generalizability of a study sample assessment procedure for entrance selection for medical school

PAUL OOSTERVELD & OLLE TEN CATE

*University Medical Center Utrecht School of Medical Sciences, Utrecht, The Netherlands*

**SUMMARY** *To optimize the validity of instruments for the selection of students for admission to medical school a close resemblance between selection criteria and activities in medical school and patient care is proposed. A study sample assessment procedure (SSAP), focusing on independent studying, collaboration with peers and providing information to standardized patients, was designed and has been applied as a selection tool at UMC Utrecht Medical School since 2001. The interviews with standardized patients are observed and rated on the quality of information provided and the quality of communication. This study investigates the psychometric properties of this observational procedure. Generalizability theory was applied to estimate the reliability of the SSAP and to compare it with other procedures carried out in the same populations, such as the rating of application forms and a structured interview procedure. Data from three years were analysed. The G-coefficients for the SSAP (0.84 to 0.90) were higher than those for the interview and the application form (0.74 to 0.83 and 0.53 to 0.61 respectively). In conclusion, the SSAP appears to be a feasible and reliable procedure. The number of raters could, if necessary, be reduced from three to two.*

## Introduction

Selection procedures for students in the health professions generally consist of assessments of cognitive abilities and personal qualities. Common criteria are pre-admission grades (e.g. Grade Point Average), aptitude tests (e.g. Medical College Admission Test), interviews, written admission forms and letters of reference. In an overview of procedures, Salvatori (2001) concludes that the GPA is the best predictor of academic performance. The MCAT also shows predictive validity for academic success, especially when used in combination with GPA. The reliability of interviews is still open to debate, but the use of a structured format and training of interviewers improves reliability and validity. Little support is found for application forms and other written measures, such as essays and letters of recommendation.

The predictive validity of selection decisions should improve when selection criteria are used that resemble outcome measures, although much of the variance still remains unexplained (Salvatori, 2001). To establish predictive validity, reliable instruments are a prerequisite. Therefore continued search for reliable ways of assessing these criteria is warranted and it is important to incorporate measures that closely resemble the domain of medical training, i.e. actual academic or professional activities. This is the line of reasoning underlying the use of so-called assessment centres in job selection measurement in which samples of actual job activities are used in the assessment

procedure. In-basket or in-tray exercises, role-playing assignments and other activities are recommended for personnel selection and have established construct validity (Lievens, 1998). Job sampling techniques have also been used successfully for the assessment of clinical competence (Turnbull *et al.*, 2000). Therefore, it seems logical to explore the measurement characteristics of training samples with regard to selection for medical training.

## *The study sample assessment procedure*

In 2001 the University Medical Center Utrecht decided to include a study sample assessment procedure (SSAP) in the entrance selection procedures (Ten Cate & Smal, 2002). The aim was to introduce a task that could be objectively judged, would reflect activities that are typical of the medical curriculum, and would be feasible to include in a selection procedure. We selected the following activities: studying a text, information exchange with a peer student, explanation of learned materials to a third party and communication with standardized patients (SP). All these elements are representative of student activities in the junior years of the Utrecht medical programme.

The full selection procedure included the following elements: an application form with several focused questions to be answered in 75-word statements, a 20 minute interview, a study sample assessment procedure (SSAP), in 2002 a questionnaire and in 2003 a knowledge test as a pre-selection procedure. The SSAP and the interview are carried out during one half-day session at the medical school location.

In the SSAP procedure two applicants are invited at a time and are interviewed by two parallel selection committees. Then both candidates are asked to perform the following activities:

- (1) individually study a three-to-five page text about diagnostic and therapeutic procedures of a particular disease (A) during one hour;
- (2) explain the studied material to the other candidate and receive reciprocal information from her about a different disease (B) which she has studied. This exchange lasts one hour;
- (3) answer the questions of a standardized patient about disease A in 15 minutes;
- (4) answer the questions of another standardized patient about disease B in 15 minutes.

*Correspondence:* Prof Dr Th.J. ten Cate, Onderwijsinstituut UMC Utrecht, PO Box 85060, 3508 AB Utrecht, The Netherlands. Tel: +31-30-2538349/2338; fax: +31-30-2538200; email: t.j.tencate@med.uu.nl

|            | Candidate 1  | Candidate 2                      |
|------------|--|----------------------------------|
| 1 hour     | Studying text of disease A                         | Studying text of disease B       |
| 15 minutes | Break  |                                  |
| 1 hour     | Mutual information exchange about diseases A and B |                                  |
| 15 minutes | Break  |                                  |
| 15 minutes | Explanation of disease A to SP-A                   | Explanation of disease B to SP-B |
| 15 minutes | Explanation of disease B to SP-B                   | Explanation of disease A to SP-A |

**Figure 1.** The SSAP steps.

A trained selection committee of three people behind a one-way screen observes the two 15 minute interviews with the standardized patients and rates these using standardized forms. The various steps in the SSAP are shown in Figure 1.

The entire procedure requires that all candidates are observed by a selection committee twice: once during a 20 minute interview and once for 30 minutes during the dialogues with the two SPs. A committee of six (two clinical staff members, two students, two student counsellors) is split in two, ensuring that each candidate is interviewed by one half of the committee and observed by the other half. Figure 2 shows the schedule of a typical selection day with 12 candidates.

#### *Assessing reliability with generalizability theory*

Generalizability theory (GT) is a widely recommended general framework for the study of the reliability of multi-faceted measurement procedures in healthcare (Cronbach *et al.*, 1972; Cardinet *et al.*, 1981; Swanson *et al.*, 1999). In generalizability theory it is possible to quantify effects of several sources of measurement error in such a way that the magnitude of the error can be estimated. Sources of error can be quantified by incorporating replications of each factor (called a *facet*). For example, the interviewer error can be studied by using more than one interviewer for each candidate, enabling an estimate of the interviewer variance to be made. The error introduced by specific tasks can be studied by incorporating several tasks.

In GT reliability is expressed as a generalizability coefficient, a measure similar to the intraclass correlation, and the standard error of measurement. Generalizability coefficients and standard errors are functions of the variance. The generalizability coefficient is a measure of homogeneity of the combination of facets, for example items and interviewers. GT can be used not only to estimate the accuracy of the current measurement procedure, but also to estimate reliabilities of hypothetical alternative designs, such as varying the number of interviewers, or of more complex designs, such as nesting the facets.

#### *Aim of the study*

This study explores the generalizability of the study sample assessment procedure in 2001, 2002 and 2003, in comparison with other methods. The SSAP was incorporated in a selection procedure for candidates for University of Utrecht Medical School. Other elements of the procedure that we investigated during the years 2001 to 2003 were a written application form and an interview.

#### **Method**

##### *Subjects*

The number of candidates in 2001, 2002 and 2003 was 61, 59 and 52 respectively. The prerequisite for entering the selection procedure was an undergraduate diploma in a biomedical field or equivalent level of education after leaving school.

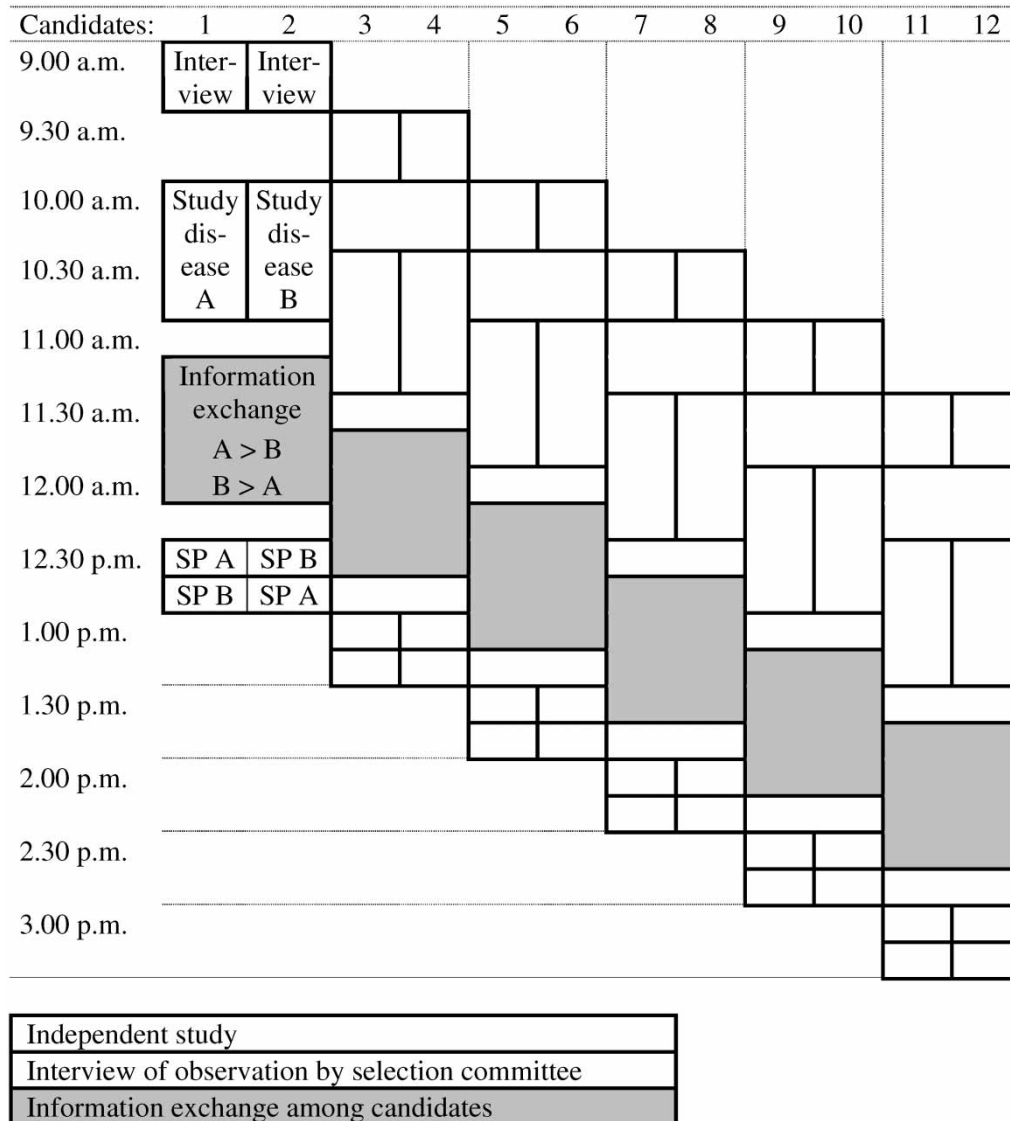
##### *Procedures*

The application form was mailed to the candidates; the interview and the SSAP took place during selection days at the UMC Utrecht.

The application form contained questions on the desired medical profession, the surplus value of the prior education, the expected ease and tempo of studying and (in 2003) academic interests. Each question required a written motivation with a maximum of 75 words. Each answer was independently rated on a four-point scale by a staff member, a student counsellor and a medical student with regard to specificity and sense of reality.

The interview was conducted by a committee of three over 20 minutes during which the same topics and rating scales were used as in the application form. A general impression of the student was included, but no distinction was made in the interview between specificity and sense of reality.

For the SSAP, the two 15 minute interviews with the standardized patients were observed by a three-person selection committee from behind a one-way screen.



**Figure 2.** Schedule of a typical selection day with interview and SSAP.

For each patient interview (regarding disease A and B respectively) the committee members independently rated (a) correctness of the given information (no major flaws?), (b) the comprehensiveness of the information given (no important information left out?) and (c) the quality of communication and the attitude towards the patient. The average scores for these criteria were weighted equally to arrive at a total score. In addition, each candidate received a score (d) for correctness and comprehensiveness of the information given in the patient interview by the *other* candidate. This was added to stimulate the information exchange between the two candidates concerning the disease information. In this way a candidate would be punished for withholding information to the colleague candidate regarding the disease studied. The sum of all ratings was used for the final evaluation of the candidates. The procedure required that candidates be assessed in pairs, being interviewed and observed by two parallel selection committees of three persons each. For each pair of students there were two committees. Committee A interviewed student A and rated

the SSAP of student B. Committee B interviewed student B and rated the SSAP of student A.

#### Data analysis

The reliability of the entire procedure was estimated with Generalizability Theory. Variance components were estimated through restricted maximum likelihood (SPSS, 1999). These analyses were performed for the total set of ratings as well as for each of the three parts of the procedure separately (the application form, the interview with the committee and the SSAP).

Generalizability coefficients and standard error were computed for relative measurement, as we were only interested in the rank order of the candidates, not in the absolute value of the scores. Generalizability coefficients for relative judgements express how well raters and/or criteria agree in their ranking of candidates, whereas generalizability coefficients for absolute judgements express how well they agree on the absolute value of the candidate's score.

**Table 1.** Generalizability coefficients and standard errors of measurements for the selection procedure.

|                    | No. of raters | 2001        |             | 2002        |             | 2003        |             |
|--------------------|---------------|-------------|-------------|-------------|-------------|-------------|-------------|
|                    |               | G           | sem         | G           | sem         | G           | sem         |
| Combined procedure | 3             | <b>0.72</b> | <b>0.17</b> | <b>0.71</b> | <b>0.15</b> | <b>0.74</b> | <b>0.15</b> |
|                    | 2             | 0.65        | 0.20        | 0.64        | 0.17        | 0.68        | 0.17        |
|                    | 1             | 0.49        | 0.28        | 0.51        | 0.23        | 0.53        | 0.23        |
| Application form   | 3             | <b>0.61</b> | <b>0.17</b> | <b>0.53</b> | <b>0.16</b> | <b>0.61</b> | <b>0.15</b> |
|                    | 2             | 0.52        | 0.20        | 0.43        | 0.20        | 0.53        | 0.18        |
|                    | 1             | 0.36        | 0.28        | 0.28        | 0.27        | 0.37        | 0.25        |
| Interview          | 3             | <b>0.83</b> | <b>0.20</b> | <b>0.74</b> | <b>0.23</b> | <b>0.75</b> | <b>0.19</b> |
|                    | 2             | 0.80        | 0.22        | 0.72        | 0.24        | 0.70        | 0.22        |
|                    | 1             | 0.72        | 0.28        | 0.68        | 0.26        | 0.60        | 0.27        |
| SSAP               | 3             | <b>0.86</b> | <b>0.17</b> | <b>0.84</b> | <b>0.18</b> | <b>0.90</b> | <b>0.17</b> |
|                    | 2             | 0.84        | 0.19        | 0.80        | 0.20        | 0.88        | 0.19        |
|                    | 1             | 0.78        | 0.23        | 0.72        | 0.26        | 0.83        | 0.23        |

Note: Actual coefficients in bold type, estimated coefficients in roman type.

**Table 2.** G-coefficients of separate SSAP criteria.

|  | 2001 |      | 2002 |      | 2003 |      |
|--|------|------|------|------|------|------|
|  | G    | sem  | G    | sem  | G    | sem  |
| Correctness of information given on own case (A)               | 0.56 | 0.43 | 0.83 | 0.45 | 0.81 | 0.44 |
| Comprehensiveness of information given on own case (A)         | 0.77 | 0.43 | 0.66 | 0.46 | 0.77 | 0.38 |
| Communication with SP A  | 0.85 | 0.38 | 0.83 | 0.43 | 0.91 | 0.38 |
| Correctness of information given on transmitted case (B)       | 0.85 | 0.44 | 0.80 | 0.49 | 0.84 | 0.44 |
| Comprehensiveness of information given on transmitted case (B) | 0.84 | 0.42 | 0.80 | 0.39 | 0.84 | 0.43 |
| Communication with SP B  | 0.91 | 0.38 | 0.80 | 0.46 | 0.93 | 0.32 |

## Results

### Reliability

Generalizability coefficients and standard errors for the entire assessment procedure as well as the separate parts are given in Table 1. The coefficients are given for the actual number of raters (3) and estimated for 2 and 1 raters.

For the entire procedure with 3 raters the G-coefficients exceed the 0.70 mark, indicating sufficient reliability (Nunnally, 1978). The standard errors are quite low, with mean 95% confidence intervals around the mean item scores of  $\pm 0.286$  for 2003,  $\pm 0.293$  for 2002 and  $\pm 0.332$  for 2001, each with regard to four-point scales. The G-coefficients for a smaller number of raters are less than 0.70.

The separate parts of the assessment procedure show mixed results. The application form is much less reliable than the entire procedure or its other two parts. The interview and the SSAP show adequate reliability coefficients and the G-coefficients for these two components suggest that fewer raters would be acceptable.

Table 2 shows G-coefficients for the SSAP criteria separately. Most components show satisfactory reliabilities, with the highest G-coefficients being for communication with both patients. There was no significant difference in the reliability scores between the interviews conducted with the two different standardized patients.

## Discussion

The aim of this study was to assess the psychometric properties of an assessment centre procedure (SSAP) for use in selection of candidates for medical school. The results suggest that the SSAP is more reliable than a written application form with free-response questions concerning reasons and motivation to attend medical school and also more reliable than our interview procedure. The interview procedure did produce satisfactory reliabilities, however, probably due to the use of a structured format and training of interviewers.

Other studies have previously shown that written application letters have a low reliability (Salvatori, 2001) and we have confirmed this in this study. This does not necessarily imply that raters are not able to assess application letters adequately. Forms filled in at home may contain other sources of uncontrolled variance, such as length of time taken in thinking about and completing the form, and the influence of others. Written testimonies might therefore not reflect the genuine nature of the candidates as well as an interview or the observation of activities.

The combined G-coefficients are lower than the separate G-scores of the SSAP and the interview. Using only the SSAP would, however, reduce the content validity of the selection as a whole. Correlations between the three procedures show that only the application form and the interview correlate (Table 3). This is not surprising,

**Table 3.** Pearson correlations between selection procedures.

|                              | 2001 | 2002 | 2003 |
|------------------------------|------|------|------|
| SSAP ★ application form      | 0.07 | 0.25 | 0.28 |
| SSAP ★ interview             | 0.19 | 0.30 | 0.30 |
| Application form ★ interview | 0.61 | 0.61 | 0.58 |

since both address the same topics. Reduction of effort could therefore best be achieved by leaving out the application form.

Generalizability theory may be applied to optimize the design of a measurement procedure. Estimated G-coefficients and standard errors for alternative designs with one or two raters per element show that using less than three raters would compromise the reliability of most measurements. The interview and certainly the SSAP, however, show sufficient reliability with only two instead of three raters so a reduction in the number of raters in these components could be considered. Increasing the number of raters is not necessary as the reliability of the overall measurement is adequate. We would not want to reduce the number of participants at the interview, because each was chosen because of his/her specific background (a clinical staff member, a student counsellor and a student) and added perspectives during the interview that could be rated by others.

The G-coefficients of the separate SSAP criteria are generally satisfactory, but the standard errors are quite large. Since the selection decision is made on the basis of the total score, where standard errors are considerably lower, this is not a great problem.

A new element in a selection procedure cannot only be judged on its formal psychometric aspect. The feasibility of the implementation is also highly relevant. From the perspective of the school the burden is large, but not excessive if organized well, as shown in Figure 2. It requires a two- to three-person committee that judges the performance of a candidate in a 30 minute observation, which is comparable to the investment required for an interview with the same committee. For the candidates it requires two hours' preparation and instruction time. Given that the candidates were already present for the interview this did not pose a problem. On the contrary, for the candidates it was an extra opportunity to show their abilities in a relevant exercise and, judging from evaluation forms, the SSAP is considered by candidates to be an acceptable selection procedure with high face validity.

It has been advocated that cognitive and non-cognitive criteria should be combined in selection procedures—cognitive criteria as a threshold to define a selected group among whom other competences are measured (Albanese *et al.*, 2003). The procedure described in this study does just this. Candidates are first tested with a long written examination. The highest scoring candidates are admitted to a set of non-cognitive assessments including the

SSAP, which appears to offer favourable psychometric characteristics.

### Practice points

- The reliability of the SSAP is at least as high or higher than other selection procedures.
- SSAP is a feasible selection instrument that takes little more time than an interview.
- The face validity of study sampling is high.

### Notes on contributors

PAUL OOSTERVELD, PhD, is methodologist at University Medical Center Utrecht School of Medical Sciences, The Netherlands.

OLLE TH.J. TEN CATE, PhD, is Professor and Associate Dean of Education, School of Medical Sciences, University Medical Center Utrecht, The Netherlands.

### Acknowledgements

The authors wish to thank the Dutch State Department of Education, Culture and Science for financial support with this study.

### References

- ALBANESE, M.A., SNOW, M.H., SKOCHELAK, S.E., HUGGETT, K.N. & FARRELL, P.M. (2003) Assessing personal qualities in medical school admissions, *Academic Medicine*, 78, pp. 313–321.
- CARDINET, J., TOURNEUR, Y. & ALLAL L (1981) Extension of generalisability theory and its applications in educational measurement, *Journal of Educational Measurement*, pp. 183–204.
- CRONBACH, L.J., GLESER, G.C., NANDA, H. & RAJARATNAM, N. (1972) *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores of Profiles* (New York, Wiley).
- LIEVENS, F. (1998) Factors which improve the construct validity of assessment centers: a review, *International Journal of Selection and Assessment*, 6, pp. 141–152.
- NORCINI, J.J., BLANK, L.L., DUFFY, F.D. & FORTNA, S. (2003) The Mini-CEX: a method for assessing clinical skills, *Annals of Internal Medicine*, 138, pp. 476–481.
- NUNNALLY, J.C. (1978) *Psychometric Theory*, 2nd edn (New York, McGraw-Hill).
- SALVATORI, P. (2001) Reliability and validity of admissions tools used to select students for the health professions, *Advances in Health Sciences Education*, 6, pp. 159–175.
- SPSS (1999) *Advances Models 9.0* (Chicago, IL, SPSS Inc.).
- SWANSON, D.B., CLAUSER, B.E. & CASE, S.M. (1999) Clinical skills assessment with standardised patients in high stakes tests: a framework for thinking about score precision, equating and security, *Advances in Health Sciences Education*, 4, pp. 67–106.
- TEN CATE, O. & SMAL, K. (2002) Educational assessment center techniques for entrance selection in medical school, *Academic Medicine*, 77, p. 737.
- TURNBULL, J., MACFADYEN, J., VAN BARNEVELD, C. & NORMAN, G. (2000) Clinical work sampling: a new approach to the problem of in-training evaluation, *Journal of General Internal Medicine*, 15, pp. 556–561.