

Reliability of the Fellowship Examination of the Royal Australian College of General Practitioners

R. B. Hays

*University of Queensland
Townsville, Australia*

W. E. Fabb

*Royal Australian College of General Practitioners
Melbourne, Australia*

C. P. M. van der Vleuten

*University of Limburg
Maastricht, The Netherlands*

The examination of the Royal Australian College of General Practitioners is a battery of eight subtests using different methods of assessment to assess different components of competence (domains) regarded as appropriate for Australian general practice. A pass/fail decision is made by combining subtest scores. This article reports an analysis of the reliability of the 1991 examination using multivariate generalizability theory. Results demonstrate that although reliability scores of individual tests vary, the combination of several tests, each providing information about a different component of competence, may produce reliable information about performance of candidates. Replication of this study with data from other candidate cohorts is required to enhance the generalizability of its findings.

The examination for Fellowship of the Royal Australian College of General Practitioners (FRACGP), introduced in 1967, is the accepted means of certifying competence to practice as an independent general practitioner in Australia. Attainment of this certification has been voluntary, although from January 1995, possession of the FRACGP will become a mandatory requirement for entry to the vocational register of general practitioners.

This examination has purposely been directed along what could be called a "high validity" route to assessing competence. Realizing that assessment of competence is a complex task, developers chose to define carefully the components of competence appropriate to Australian general practice and then sought methods that would assess those components. The focus of assessment was not on measuring simple recall of factual knowledge but rather on application of knowledge and communication skills and how candidates deal with patients and their problems. Where possible, patient simulations (either written or role-played) are used.

A conceptual framework for the assessment, depicted in Figure 1, was developed. Competence was thought to consist of five "domains": knowledge, prob-

lem solving, clinical interpretation, psychomotor skills, and affective behavior.

Recognizing that no single assessment method was able to assess competence defined as broadly as in the conceptual framework adopted, a range of test methods was employed as appropriate to the task. The result was a complex battery of tests requiring approximately 15 hr of formal assessment in addition to assessment of the practices of individual candidates. This examination is therefore longer than any other test of competence for general practice documented in the literature.

Our study investigated the reliability of the 1991 examination. The purpose of the analysis was threefold; first, to determine the reliability of the segments or subtests; second, to estimate the reliability of the examination as a whole; and third, to assess the relative contribution to reliability of the individual segments.

The Structure of the Examination

Detailed discussion of the FRACGP examination, including its development, administration, and scoring, has been published elsewhere and should be consulted

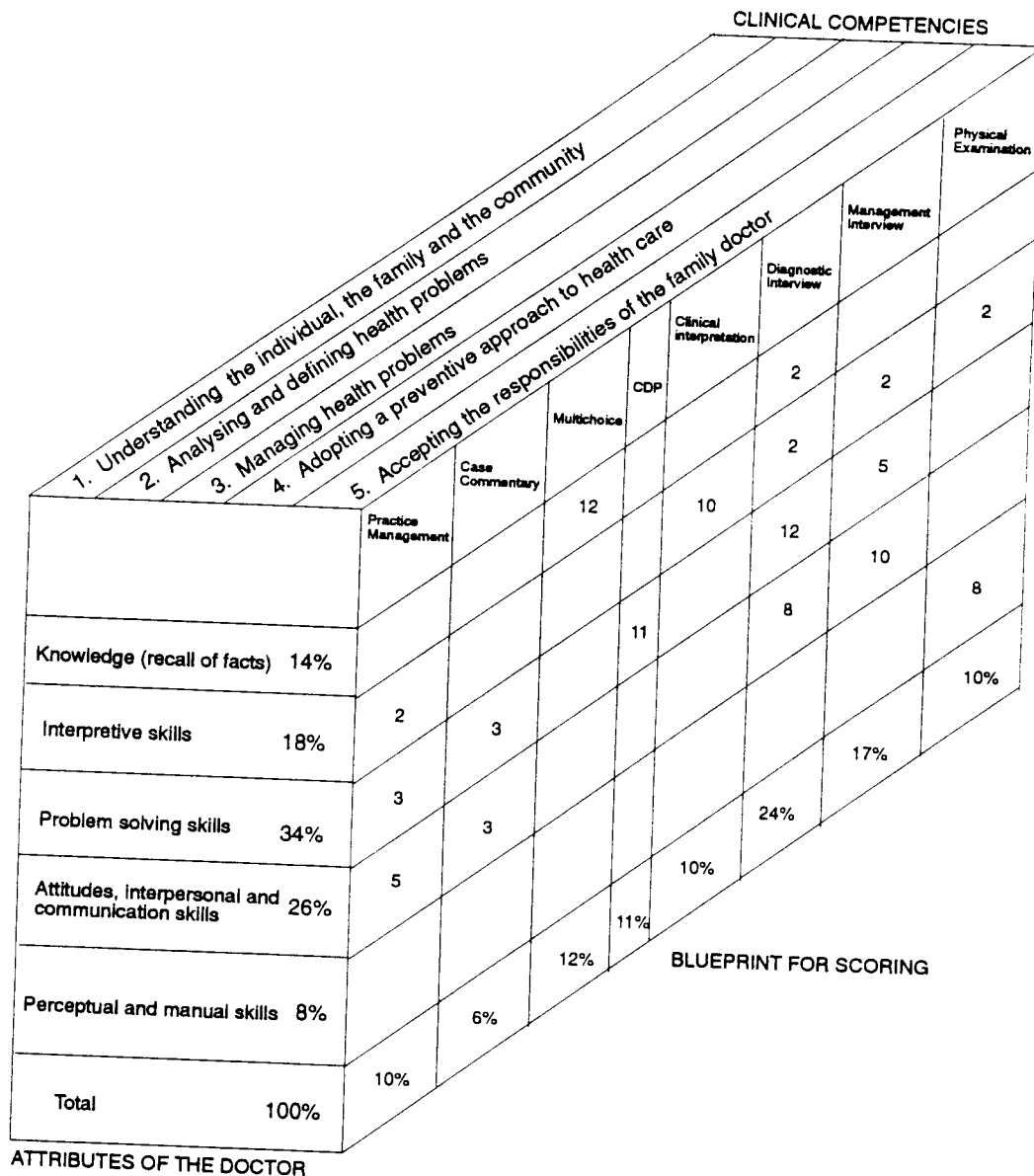


Figure 1. The conceptual framework for assessment.

for complete information.^{1,2} In brief, the 1991 examination comprised eight segments, details of which are summarized in Table 1. The multiple choice questions (MCQ) are selected from a computer bank of questions designed for general practice according to a formula designed to give a balanced sample of the universe of general practice. The Clinical Interpretation (CI) examination consists of 80 cases of highly visual material (slides of rashes, obvious clinical signs, radiographs, ECGs, laboratory data, etc.), which must be interpreted within either 1 or 2 min, according to the brief clinical context provided for each case. Candidates select their answers from a long list of possible answers. In both MCQ and CI, all examinees are given the same items.

Computerized diagnostic problems (CDPs) are interactive computerized simulations based on genuine clinical cases and marked by computer. Candidates are randomly allocated cases from the bank held on computer file. In each case, the candidate is presented with a clinical scenario that includes the presenting symptoms and the context in which the doctor is working. The candidate then selects from a 10-item menu what he or she wishes to do first. Except in emergency cases, the candidates will usually elect to take a history followed by a physical examination, office tests, and investigations, after which treatment may be given. In each instance, the candidate seeks information through the computer keyboard by typing in the first three letters

Table 1. Subtests of the FRACGP Examination

| Subtest | Number of Items | Duration in Hours | Mean | Standard Deviation |
|------------------------------------|-----------------|-------------------|--------------------|--------------------|
| Multiple Choice (MCQ) | 200 | 3.0 | 73.54 | 7.58 |
| Clinical Interpretation (CI) | 80 | 2.0 | 71.87 | 6.35 |
| Computer Diagnostic Problems (CDP) | 2 | 3.0 | 69.05 | 9.07 |
| Case Commentaries (CC) | 2 | 3.0 ^a | 78.12 | 7.96 |
| Diagnostic Interview (DI) | 3 | 1.5 | 73.32 | 8.27 |
| Management Interview (MI) | 2 | 0.5 | 72.93 | 13.72 |
| Physical Examination (PE) | 5 | 1.5 | 75.78 | 9.31 |
| Practice Assessment (PA) | 1 ^b | 0.5 | — | — |
| Composite test | — | 15 | 73.51 ^c | 5.40 |

^aNo formal time limit set; global estimate of average testing time. ^bAlthough multiple cases are assessed, only a single overall mark is given; subtest excluded from further analysis. ^cBlueprint weights used in composite test score calculation (see D-study 1 in Table 4).

of the information required. Candidates are encouraged to think clinically and ask the questions that will confirm or refute their hypotheses generated by the presenting symptoms. They are able to work their way through the case as in clinical practice, arrive at diagnostic conclusions, and institute initial management.

Case commentaries (CC) require candidates to submit two 1,500- to 2,000-word case presentations demonstrating continuing care of a patient or family within the candidate's own practice. Two markers independently score candidate performance using a 5-item marking form. The five role-played consultations, diagnostic interviews (DIs), and management interviews (MIs) assess primarily communication skills but also applied knowledge and problem solving. Performance is marked on rating scales by both the role-playing patient (another general practitioner) and an observing examiner (different examiners for each case). Physical examination (PE) assesses a candidate's ability to elicit clinical signs (process and findings) in four genuine patients with stable signs, ability to perform one of a range of practical procedures on simulation mannikins, and ability to perform cardiopulmonary resuscitation on a mannikin. Two examiners score examinee performance.

Practice assessment (PA) is a structured oral examination based on a logbook recording of 100 consecutive patients from the candidate's own practice. Although multiple markers are used in DI, MI, PE, and PA, marks are given by consensus, and only a single score per case is available. In addition, not all examinees are given the same cases in these subtests.

Candidate scores for each of the eight subtests are apportioned, according to a predetermined key or blueprint, to scores for each of the five domains. To pass the examination, candidates must achieve a minimum of 66% in each domain score and in the whole examination. Borderline performance is judged with the aid of written feedback from scorers in each subtest, and it is possible to pass with a slightly lower mark in either two

domains (63%–65.9% in each) or one subtest (61%–62.9%), if performance is otherwise unblemished. Hence there is an absolute cut-off of 61% in all domain scores, below which a candidate is regarded as having failed the domain and the whole examination. Candidates may be asked to retake only those subtests that compose the majority of the relevant domain score; for example, if the knowledge domain score is below 61%, the candidate will be asked to retake the MCQ subtest.

Methods

The 1991 examination was used for analysis. Of the 140 candidates, only 79 took all subtests at that examination. This was the first attempt for all 79 candidates; there were no repeaters. The item scores within each subtest of these 79 examinees were used. An item score was defined as a score on the smallest available independent score within a subtest. For some subtests, these were individual item scores (e.g., MCQ); in others, these were case scores (in which cases could consist of multiple questions or ratings; e.g., MI). All item scores were expressed in a percentage score. The subtest scores were calculated by averaging item scores, and the composite score was an average of the subtest scores weighted by the blueprint weights (see Table 2).

The reliability of the individual subtests was estimated using generalizability theory.^{3,4} Where multiple marks from multiple examiners were available (only for CC; the others were marked by consensus), scores were averaged to a single score. In the original design, raters were nested within cases and within candidates: r:i:p). The PA subtest was dropped from analysis, because only one overall score and no separate item scores were available. This resulted in a person-by-item score matrix for the seven remaining subtests. Although there were design differences across subtests—in some, examinees were given the same items; in others, different items—all subtests were submitted to a random items-

Table 2. Composite Reliability Results for Several Decision Studies, Using Different Numbers of Items and Subtest Weights

| | MCQ | CI | CDP | CC | DI | MI | PE |
|---|------|-------|------|------|------|------|------|
| D-Study 1: Actual Test | | | | | | | |
| Number of Items | 200 | 80 | 2 | 2 | 3 | 2 | 6 |
| Test Duration in Hours | 3 | 2 | 3 | 3 | 1.5 | 0.5 | 1.5 |
| Weights | 0.13 | 0.11 | 0.12 | 0.07 | 0.27 | 0.19 | 0.11 |
| Universe Score Contribution | 2.33 | 2.07 | 3.09 | 1.48 | 7.30 | 8.08 | 2.33 |
| Relative Contribution | 0.09 | 0.08 | 0.12 | 0.06 | 0.27 | 0.30 | 0.09 |
| Optimal Number of Items | 88 | 76 | 15 | 10 | 43 | 38 | 26 |
| Composite Universe Score Variance | | 26.69 | | | | | |
| Composite Error Score Variance | | | 7.08 | | | | |
| Domain-referenced Reproducibility Coefficient | | | 0.79 | | | | |
| Mastery-referenced Reproducibility Coefficient | | | 0.92 | | | | |
| Standard Error of Measurement | | | 2.66 | | | | |
| D-Study 2: Equal Subtest Weights | | | | | | | |
| Number of Items | 200 | 80 | 2 | 2 | 3 | 2 | 6 |
| Test Duration in Hours | 3 | 2 | 3 | 3 | 1.5 | 0.5 | 1.5 |
| Test Duration in Hours | 3 | 2 | 3 | 3 | 1.5 | 0.5 | 1.5 |
| Weights | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 | 0.14 |
| Universe Score Contribution | 2.72 | 2.73 | 3.73 | 3.01 | 3.43 | 5.18 | 3.14 |
| Relative Contribution | 0.11 | 0.11 | 0.16 | 0.13 | 0.14 | 0.22 | 0.13 |
| Optimal Number of Items | 89 | 92 | 16 | 18 | 21 | 26 | 32 |
| Composite Universe Score Variance | | 23.29 | | | | | |
| Composite Error Score Variance | | | 5.33 | | | | |
| Domain-referenced Reproducibility Coefficient | | | 0.82 | | | | |
| Mastery-referenced Reproducibility Coefficient | | | 0.94 | | | | |
| Standard Error of Measurement | | | 2.31 | | | | |
| D-Study 3: Reduced MCQ, double DI and MI | | | | | | | |
| Number of Items | 100 | 80 | 2 | 2 | 6 | 4 | 6 |
| Test Duration in Hours | 3 | 2 | 3 | 3 | 3 | 1 | 1.5 |
| Weights | 0.13 | 0.11 | 0.12 | 0.07 | 0.27 | 0.19 | 0.11 |
| Universe Score Contribution | 2.33 | 2.07 | 3.09 | 1.48 | 7.30 | 8.08 | 2.33 |
| Relative Contribution | 0.09 | 0.08 | 0.12 | 0.06 | 0.27 | 0.30 | 0.09 |
| Optimal Number of Items | 60 | 52 | 10 | 7 | 29 | 26 | 18 |
| Composite Universe Score Variance | | 26.69 | | | | | |
| Composite Error Score Variance | | | 4.49 | | | | |
| Domain-referenced Reproducibility Coefficient | | | 0.86 | | | | |
| Mastery-referenced Reproducibility Coefficient | | | 0.94 | | | | |
| Standard Error of Measurement | | | 2.12 | | | | |

nested-within-persons analysis of variance (ANOVA) (i:p design). Subsequently, variance components were estimated. A comparable design across all subtests was required for estimation of the composite reliability (see the following). However, because a domain-referenced score interpretation was adopted for estimating reproducibility coefficients, design differences resulting from crossing or nesting items had no consequences for error variance estimation and hence for the reliability of the subtests. As opposed to a norm-referenced perspective (scores relevant to each other), a domain-referenced score interpretation gives absolute meaning to the test results. Scores are interpreted as absolute indicators of mastery in a domain (e.g., a candidate has 80% mastery in history taking). Absolute meaning is alternatively given to test results when pass/fail decisions are considered. Pass/fail decisions indicate whether a

score has or has not exceeded a particular cut-off score (in other words, whether a domain is mastered or not). This is referred to as a mastery-referenced score interpretation. Here, for each subtest a reproducibility coefficient was calculated for both a domain-referenced score interpretation and a mastery-referenced score interpretation,^{4,5} using 66% as a cut-off for the latter. In addition, the standard error of measurement (SEM) was calculated. For each score interpretation, these reliability indices were estimated for the actual items used in the subtest. To allow meaningful comparison across subtests, they were also standardized for unit of testing time, in this case for 1 hr of testing time.

The candidates in this examination were tested at different sites. There was, however, no random assignment of candidates to the different sites. Similarly, in those subtests in which different markers were involved, candidates were not randomly as-

signed to examiners. This may have caused some variation between the scores of the candidates above that attributable to variation in the ability of candidates. To some extent, this confounding may therefore have inflated the variance component for persons.

The reliability of the overall examination was estimated by using multivariate generalizability theory.^{3,4,6} In multivariate generalizability theory, an object of measurement may have multiple universe scores (or true scores) each connected to a specific level of a fixed facet. In the FRACGP examination, subtests can be considered a fixed facet, with each subtest associated with a separate universe score. Multivariate generalizability theory not only provides convenient estimation of composite reliability but also allows for investigating the relative contribution of each of the subtests to the overall universe score, therefore providing suggestions for improving the reliability. The design of the overall examination is an items-nested-within-persons-within-subtests design, in which subtests is considered a fixed facet ($(i:p)xs, s$ fixed). A matrix of person variance components and covariance components is estimated (S_p) representing the multiple universe score (co-)variances across subtests. Similarly, a matrix of error score variance (S_{ip}) is estimated. The off-diagonal values of this matrix are zero because of independent sampling of items and assumptions about uncorrelated residual effects.⁴ A composite universe score variance and error variance may be estimated by summation of respective matrix elements after weighing the entries by their appropriate number of items and subtest or blueprint weights. The composite universe score variance can be broken down into relative proportions of subtest variance in relation to the total composite universe score variance. By comparing the relative universe score contribution with the actual number of items, it is possible to infer an optimal number of items, suggesting alterations for improvement of reliability by differential weighing of subtests (either by lengthening or shortening sub-

tests by changing the number of items or by changing the subtest weights directly in total score summation across subtests). Using the composite universe score variance and error score variance, reproducibility coefficients can be calculated in a regular way. A domain-referenced and a mastery-referenced reproducibility coefficient will be reported here, using a cut-off score of 66%. It should, however, be realized that the latter does not directly reflect the actual decision-making process of the FRACGP because no full compensation across subtests is actually allowed, and decisions are based on a combination of domain or subtest scores (or both). For that purpose, the decision reliabilities of individual subtests are more appropriate.

Results

Table 1 provides descriptive statistics for individual subtests and the test as a whole. There is some variation in subtest difficulty, but they uniformly exceed the 66% cut-off score. The variation of subtest scores is highest for MI. Although there are more subtests with few cases, the wide variation could be caused by the limited number of cases that are used in this subtest (only two). For the same reason, but in the opposite direction, the variation of composite scores from the test as a whole is smaller than its components.

Subtest Reliability

In Table 3, reproducibility coefficients and SEMs are reported for each subtest, separate for the actual testing time used per subtest and for the standardized 1 hr of testing time. These reliability indices are derived from the diagonal entries of the matrices S_p and S_{ip} in Table 4, representing the regular variance components for true and error variance respectively associated with the individual subtests.

The domain-referenced reproducibility coefficients for the actual testing time used vary considerably from

Table 3. *Reproducibility Coefficients and Standard Error of Measurements for Individual Subtests from a Domain-referenced and a Mastery-Referenced Score Interpretation*

| Subtest | Domain-Referenced | | Mastery-Referenced | | Standard Error of Measurement | |
|------------------------------|---------------------|-------------|---------------------|-------------|-------------------------------|-------------|
| | Actual Testing Time | 1-Hour Test | Actual Testing Time | 1-Hour Test | Actual Testing Time | 1-Hour Test |
| Multiple Choice | 0.84 | 0.64 | 0.94 | 0.82 | 3.08 | 5.32 |
| Clinical Interpretation | 0.48 | 0.32 | 0.70 | 0.63 | 5.01 | 7.08 |
| Computer-Diagnostic Problems | 0.63 | 0.36 | 0.71 | 0.45 | 5.52 | 9.32 |
| Case Commentaries | 0.38 | 0.17 | 0.85 | 0.65 | 6.27 | 10.84 |
| Diagnostic Interview | 0.50 | 0.40 | 0.77 | 0.69 | 5.92 | 7.25 |
| Management Interview | 0.56 | 0.72 | 0.69 | 0.82 | 9.05 | 6.40 |
| Physical Examination | 0.47 | 0.37 | 0.79 | 0.71 | 6.85 | 8.39 |

Table 4. *Estimated Variance and Covariance Components (\pm Standard Errors) of p (S_p) and $i:p$ ($S_{i:p}$)*

| Subtest | MCQ | CI | CDP | CC | DI | MI | PE |
|-----------|------------------------|-------------------------|----------------------|----------------------|-----------------------|-----------------------|-----------------------|
| S_p | | | | | | | |
| MCQ | 50.67 ± 9.07 | | | | | | |
| CI | 25.07 ± 4.77 | 23.45 ± 6.38 | | | | | |
| CDP | 6.80 ± 5.78 | 22.49 ± 4.64 | 51.85 ± 13.87 | | | | |
| CC | 12.35 ± 4.14 | 13.29 ± 3.04 | 32.07 ± 5.33 | 24.08 ± 11.76 | | | |
| DI | 13.08 ± 4.91 | 18.06 ± 3.78 | 24.69 ± 5.49 | 20.56 ± 3.97 | 34.73 ± 11.47 | | |
| MI | 7.67 ± 8.25 | 13.70 ± 5.79 | 29.89 ± 8.95 | 25.80 ± 6.34 | 42.00 ± 8.26 | 106.32 ± 32.47 | |
| PE | 17.14 ± 5.52 | 17.21 ± 4.01 | 14.58 ± 5.49 | 19.27 ± 4.17 | 14.61 ± 4.59 | 28.18 ± 8.14 | 42.34 ± 14.15 |
| $S_{i:p}$ | | | | | | | |
| MCQ | 1898.79 ± 59.04 | | | | | | |
| CI | | 2007.13 ± 108.90 | | | | | |
| CDP | | | 60.83 ± 9.50 | | | | |
| CC | | | | 78.71 ± 12.29 | | | |
| DI | | | | | 105.24 ± 11.85 | | |
| MI | | | | | | 163.95 ± 25.61 | |
| PE | | | | | | | 234.82 ± 19.77 |

subtest to subtest. Most reliable is the MC subtest, least reliable are the CC. Standardized for time, the MI subtest turns out to be most reliable, with MC second, and CC again least reliable.

Substantial differences in reliability emerge when shifting from a domain-referenced score interpretation to a mastery-referenced perspective. Unreliable subtests from a domain-referenced perspective may still yield reliable decisions. This is a reflection of the measurement error in relation to the difference between mean performance and cut-off score. For pass/fail decisions to be reliable for examinees scoring in the vicinity of the cut-off score, the measurement error of the test should be quite small (i.e., the test should be quite reliable). When examinees perform well above (or below) the critical score, the measurement error may be larger and yet allow reliable decision making. Even with sizable measurement error, the subtest decision reliability may therefore still be high, depending on the distance of the average examinee performance in relation to the cut-off score. Although accepted guidelines for judging decision reliabilities do not exist, the actual decision reliabilities here are all moderate to high.

The SEM reflects the magnitude of measurement error on the original score scale. It provides a very useful index to interpret the reliability. The SEM may

be used to estimate a confidence interval for (individual) test scores. Adding and subtracting the SEM gives an estimate of the range in which an examinee's true score will lie with 68% of certainty. By multiplying the SEM by 1.96 or 2.58 (the respective z values in the normal curve), the 95% or 99% confidence intervals are obtained. The SEMs allow a direct interpretation of the reliability. The minimum reliability value of 0.80 commonly reported for educational tests becomes a less dichotomous benchmark if one realizes the amount of error associated with this value. For instance, the MC test has an "adequate" reliability of 0.84, but any score derived from this subtest has an approximate 6% (95% confidence level) of error associated with it. Table 2 shows that for the actual testing time used in the FRACGP, the SEMs vary across subtests, ranging from approximately 3% (MC test) to 9% (MI).

Composite Reliability

The estimated variance and covariance components are reported in Table 4. The matrix S_p is the multivariate counterpart of variance associated with persons. The diagonal elements are the variance components obtained from the individual subtest analyses. The MI

component is relatively large compared to the other values (explaining the high reproducibility coefficients with only a few cases), whereas CI and CCs are relatively low. The off-diagonal elements are the covariance components indicating associations between subtests. Although there is substantial variability in the covariance components, many values are only slightly lower than their diagonal values. This suggests that there are (high) correlations between the universe scores between the subtests. The pattern of covariances indicates a cognitive cluster (MC with CI), a performance-based cluster (DI, MI, and to some extent, PE), and a cluster of written and computerized clinical measures (CCs and CDPs).

The S_{ip} matrix represents the multivariate error variance. The error variance of MCQ and CI is relatively large. However, they are difficult to compare because the unit of an item is quite different across formats (e.g., an MCQ requires substantially less time than a computer problem). Standardizing the error variances of Table 4 by expressing them at a standard length of 1 hr of testing time yields more comparable values. These values are 27.92 (MCQ), 50.18 (CI), 90.79 (CDP), 117.48 (CC), 52.62 (DI), 40.99 (MI), and 70.49 (PR). In this comparison, particularly the MCQ, because of its efficiency of sampling, provides the smallest error variance.

For many entries in Table 4, the standard errors in both matrices are sizable. This is a reflection of the sample sizes on which the estimates are based. Some caution is therefore in order.

The variance and covariance components from Table 4 were used to estimate the composite reliability. In Table 2, results are reported for several testing situations (decision studies), including the actual test situation and a few alternative test situations in which the number of items within subtests or blueprint weights (or both) are varied.

Decision study 1 reflects the actual test situation of the FRACGP, using actual numbers of items and (proportional) subtest weights defined by the blueprint. A satisfactory reliability is achieved with an overall domain-referenced reliability of 0.79, a decision reliability of 0.92, and an SEM of 2.66. The breakdown of universe score contribution shows considerable variation across subtests, suggesting variability in reliability contribution of subtests. DI and MI contribute heavily to the universe score. In part, this is a natural consequence of the heavy weights these subtests are given in the blueprint. The proportional or relative universe score contribution provides suggestions for improving the composite reliability. A perfect balance in reliability would be obtained if both actual weights and relative universe score contribution were equal. If the relative contribution to the universe score variance is larger than the weight given, then the particular subtest apparently contributes heavily to the reliability, and by adjusting

the weight or the number of items, the reliability could be improved. The last row in decision study 1 contains suggestions for this adjustment by minimizing the error variance through estimating the optimal number of items. The optimal item numbers suggest that the MC and CI subtest could be shortened, whereas all others should be lengthened, especially the performance-based components DI, MI, and PE.

To inspect the impact of the differential blueprint weights, decision study 2 reflects an equal weight application. The composite reliability results improve slightly. However, DI and especially MI remain best in their universe score contribution. The optimal item numbers invariantly indicate shortening the MC subtest, whereas the CI subtest is approximately balanced at the current length. For all other components, especially the performance-based components, an increase of the sample is still in order.

An infinite number of item combinations and weights could be investigated. Each option will affect the validity of the examination and resources required. Changes will also affect the subtest reliability. To investigate a practically attainable testing situation not affecting the validity of the test, decision study 3 reports a situation using current blueprint weights, but with the MC test reduced to half its size to 100 items and DI and MI doubled in length to six and four cases, respectively. A reasonable improvement is the result of this alternative: the domain-referenced reproducibility improves from 0.79 to 0.86, the decision reliability improves slightly from 0.92 to 0.94, and the SEM decreases from 2.66 to 2.12.

Discussion

As is reported consistently in the literature of clinical competence assessment, adequate reliability of test scores for many testing methods is difficult to attain and requires substantial testing time.^{7,8} The (domain-referenced) subtest reliabilities reported here are no exception to that finding. Clinical competence in one particular situation (e.g., test item, station, problem) is not very predictive of competence in another situation, regardless of the method being used. Therefore sufficient sampling of different situations is the problem. Although there is no linear relationship, generally those methods that are more efficient—defined by the time they require to test a single situation (item)—do better in terms of their reliability. The efficient MC subtest here was found to be most reliable despite having the largest error variance component, and the inefficient CCs subtest was found least reliable, with other subtests in between. The score reliability of the MI subtest appeared an exception, with a fairly high reliability per unit of testing time. It remains to be seen whether this can be attributed to chance or to true measurement

characteristics of this subtest. Previously, other studies showed that communication skills are less variable across different situations or patient cases.⁹

Also in accordance with findings in the literature¹⁰ is that reliability conclusions alter drastically when a mastery-referenced perspective is adopted, focusing on the reliability of decisions rather than on scores. For all subtests, the decision reliability improves substantially. Even a subtest with a very unreliable score reliability may have an adequate decision reliability. As was explained earlier, this is dependent on the achieved score distribution of the group of examinees in relation to the position of the cut-off scores. The average score on all subtests here exceeded the cut-off score to a sizable extent, implying that for most examinees, reliable pass/fail decisions are possible even with fairly imprecise test scores.

The composite reliability of the battery of tests as a whole appeared reasonably satisfactory for both the domain-referenced score reliability and the decision reliability. Adopting a 95% confidence level, composite scores on the FRACGP can be interpreted with an approximate 5% of precision.

It should be noted that all these reliability findings are to some extent an underestimate of the actual reliabilities, because one subtest (PA) was discarded from analysis.

The multivariate generalizability analyses allowed estimation of the differential contribution to reliability of the various subtests. Particularly the performance-based component contributed heavily to the overall reliability (especially DI and MI and, to a lesser extent, PE). This is a reflection of their relatively large individual true score variances, their positive relationships with other subtests, and their heavy blueprint weight. Compared with the other subtests, the performance-based subtests have small numbers of items. To improve the reliability of the FRACGP, lengthening these components is in order. As was demonstrated, the overall reliability was improved significantly by doubling the number of items in the DI and MI subtests, even with shortening the MC subtest.

Recognizing this, the College examiners are considering a change to the conceptual model, which would involve the replacement of the present domains with four sets of skills relating to clinical practice: cognitive skills, communication skills, physical examination skills, and practice skills. There is vigorous debate about whether consulting and physical examination

skills should be joined to form a larger group, clinical skills. These changes would increase the number of MI and DI clinical cases used to derive a score for clinical skills. It is expected that this change in perspective will be adopted in late 1993.

In general, it can be concluded that the approach of using a multitude of different methods for a licensing examination aiming at assessing a range of complex professional abilities may produce reliable information. Replication of this study using other data sets (from other cohorts) is in order to enhance the generalizability of these findings.

References

1. RACGP, Fabb WE (Ed.). *The examination and assessment system of the RACGP. A manual for examiners*. Melbourne: RACGP, 1991.
2. RACGP, Fabb WE (Ed.). *Fellowship of the RACGP by examination and assessment. A handbook for candidates*. Melbourne: RACGP, 1992.
3. Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. *The dependability of behavioral measurements: Generalizability for scores and profiles*. New York: Wiley, 1972.
4. Brennan RL. *Elements of generalizability theory*. Iowa: American College Testing Program, 1983.
5. Brennan RL, Kane MT. An index of dependability for mastery tests. *Journal of Educational Measurement* 1977;14:277-8.
6. Jarjoura D, Brennan RL. A variance components model for measurement procedures associated with a table of specifications. *Applied Psychological Measurement* 1982;6:161-71.
7. Swanson DB. A measurement framework for performance-based tests. In I Hart, R Harden (Eds.), *Further developments in assessing clinical competence*. Montreal: Can-Heal, 1987.
8. Van der Vleuten CPM, Newble D, Case S, Holsgrove G, MacCrae C, McCann B, Saunders N. Methods of assessment in certification. In D Newble, B Jolly, R Wakeford (Eds.), *The certification and recertification of doctors: Issues in the assessment of competence* (pp. 105-25). Cambridge: Cambridge University Press, 1994.
9. Van Thiel J, Kraan HF, Van der Vleuten CPM. Reliability and feasibility of measuring interviewing skills using the revised Maastricht History Taking and Advice Checklist. *Medical Education* 1991;25:224-9.
10. Colliver JA, Verhulst SJ, Williams RG, Norcini JJ. Reliability of performance on standardized patient cases: A comparison of consistency measures based on generalizability theory. *Teaching and Learning in Medicine* 1989;1:31-7.

Received 22 July 1993

Final revision received 28 March 1994