Educational Research and Evaluation
1995, Vol. 1, No. 3, pp. 276-286

1380-3611/95/0103-0276$6.00


# A Written Knowledge Test for Postgraduate Medical Students: Reliability in Relation to Different Educational Goals

Yvonne D. van Leeuwen[1], Marjan C. Pollemans[2], Herman Düsman[2], Cees P.M. van der Vleuten[1] and Richard P.T.M. Grol[1]
[1]University of Limburg, Department of General Practice
[2]University of Utrecht, Department of General Practice

## ABSTRACT

The use of written knowledge tests in (medical) education is widespread. Only few of them are thoroughly validated. Usually, validity studies are restricted to establishing 'face-validity', the apparent similarity between test-material and real life problems. Reliability studies are usually restricted to estimation of the coefficient alpha, representing the reproducibility of rank-ordering of students at repeated test administration. This study addresses reliability from a broader perspective, using generalizability theory. The approach enables faculty to gain insight into the suitability of the test to serve different educational goals.

A written knowledge test was examined, applied in postgraduate training for general practice in the Netherlands. Test-reliability was approached from different perspectives: the norm-oriented perspective, aiming at rank-ordering (groups of) students, the domain-oriented perspective, aiming at determining the absolute score level of (groups of) students and the decision-oriented perspective, aiming at taking pass-fail decisions.

Reliability estimates differed for the different perspectives. The implication of the results and feasible options to increase reliability are discussed.

Correspondence: Yvonne D. van Leeuwen, University of Limburg, Department of General Practice, P.O. Box 616, 6200 MD Maastricht, the Netherlands.

Manuscript submitted: March 17, 1995
Accepted for publication: September 19, 1995

# RELIABILITY OF A CASE-BASED KNOWLEDGE TEST FOR GENERAL-PRACTICE TRAINEES

## Postgraduate training for general practice

During the last two decades, written knowledge tests, usually multiple choice tests, have conquered the educational world. They were welcomed because of their objectivity and feasibility. In the course of their existence, however, they have often been reviled because of their assumed low validity. In view of their wide-spread use, remarkably few thorough validity studies have been performed. Such studies should include scrutiny of the test-content and test-results in relation to the competence that is to be assessed as well as scrutiny of test-reliability.

In this study, the quality of a written knowledge test for postgraduate students in general practice in the Netherlands is examined. The validity of the test is described elsewhere (Van Leeuwen et al., 1995). Here, the issue of test-reliability is addressed.

Classical test theory provides a means to estimate the reproducibility of the rank order of test scores, reflected in the coefficient alpha. The aim in education, however, is not merely to rank-order students, but also to establish their absolute level of competence and/or to determine whether students have a sufficient level of knowledge for qualification. Generalizability theory (Cronbach et al., 1972; Brennan 1983) addresses reliability from all these perspectives.

The objective of this study is to explore the reliability of the test, administered to (groups of) postgraduate medical students in general practice, related to these different educational goals.

Postgraduate training for general practice in the Netherlands is a compulsory training that consists of two (since September 1994 three) years which are mainly spent in general practice under the supervision of a GP trainer (Dubois et al., 1987). The assessment mainly refers to their performance 'on the job', which is evaluated by their trainers. Tests on knowledge, technical skills and communication skills have been developed for nationwide use (Pollemans et al., 1988). Up to now, these tests have had a merely educational function. Also, their use as instruments for programme evaluation has been limited. The reason is that their qualities were not fully known.

## The Knowledge Test

The knowledge test consists of about 80 patient-problems (cases) with a total of 160 case-related items (Kramer & Pollemans, 1990; Van Leeuwen & Van Hessen, 1990). The cases are derived from general practice.

The response format is of the true/false/question mark type. One mark is given for a correct answer, a negative mark for an incorrect answer, whereas no credit is given for the question mark option. The question mark option is introduced to discourage guessing as well as the habit of doctors to pretend omniscience. Consequently, the test score is composed of the total percentage correct minus incorrect answers.

Tests are administered at regular intervals during postgraduate training, to all (about 500) postgraduate students in general practice in the Netherlands. Sequential tests have a similar format but differ in content. At each test administration all students take the same test, regardless of their training level. Each test is set at the level of general practitioners at the moment of certification. The test is thus designed to provide longitudinal information, and to record progress during training (Kramer & Pollemans, 1990; Van der Vleuten & Verwijnen, 1990).

## METHODS

### Materials

For purposes of analysis the correct minus incorrect scores of three knowledge tests were used, those of June 1991, October 1991 and February 1992. All students in postgraduate training for general practice in the Netherlands participating in these three tests were included in the study. This implies that students may appear up to three times in the study (at 4, 8 and 12 months of training, respectively).

Table 1 contains the number of students per group for the three tests included in the analysis. For statistical analysis it was necessary to balance the number of students per test per group of training level. Mean scores of the selected groups did not deviate from the original groups.

### Reliability Estimates

Reliability is not a characteristic of the test as such, but varies with the object of measurement and with the interpretation of test scores (Thorndike, 1988). Generalizability theory was used as a framework to estimate reliability according to three perspectives of test score interpretation (Cronbach et al., 1972; Brennan, 1983). Reliability of test scores of individuals may imply the following (Brennan, 1983):

*Reliability of the ranking of individuals:* the norm-oriented perspective. Thus defined, reliability implies that the ranking order of individual scores is reproducible from test to test. Variation in test (or item) difficulty is not relevant, because the position of an individual score in the rank-

Table 1. The Selected Number of Respondents (N) per Training Level, their Corresponding Mean Scores (Percentage Correct Minus Incorrect (% C-IC)) and Standard Deviation (SD) for the Tests of June 1991 (151 items), October 1991 (150 items) and February 1992 (146 items).

| | June 1991 | | | October 1991 | | | February 1992 | | |
|---|---|---|---|---|---|---|---|---|---|
| Training level (in months) | N | Mean | SD | N | Mean | SD | N | Mean | SD |
| 4 | 52 | 41.7 | 10.5 | 45 | 29.3 | 8.2 | 46 | 33.3 | 10.2 |
| 8 | 52 | 51.1 | 9.8 | 45 | 39.8 | 11.5 | 46 | 41.4 | 10.3 |
| 12 | 52 | 50.4 | 9.3 | 45 | 36.3 | 8.8 | 46 | 40.4 | 10.1 |
| 16 | 52 | 55.4 | 11.7 | 45 | 42.0 | 9.0 | 46 | 39.6 | 9.6 |
| 20 | 52 | 57.3 | 9.8 | 45 | 45.7 | 9.9 | 46 | 45.7 | 9.2 |
| 24 | 52 | 57.9 | 9.5 | 45 | 47.5 | 14.4 | 46 | 48.8 | 10.1 |
| Total | 312 | 52.3 | 11.5 | 270 | 40.1 | 12.9 | 277 | 41.5 | 11.0 |

ing order is not related to a specific level of competence. An example is presented in Figure 1.

The students A, B and C score 65, 59 and 53 respectively on test one, and 75, 49 and 43 respectively on test two. Reliability is maximal here,

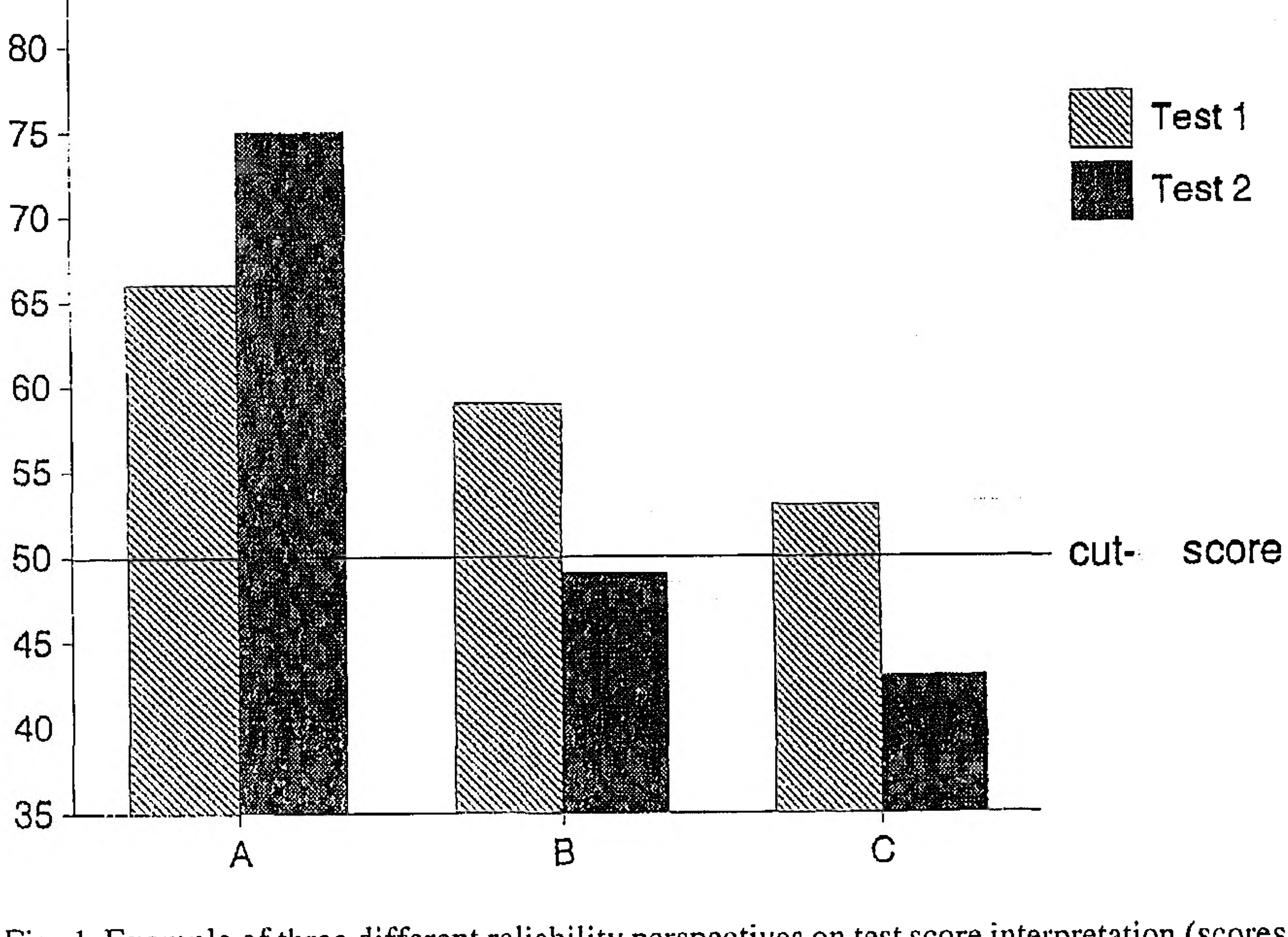


Fig. 1. Example of three different reliability perspectives on test score interpretation (scores on test 1 and test 2 for students A, B and C).

because the ranking remains the same despite the fact that the scores have changed.

*Reliability of the estimation of the absolute level of competence:* the domain-oriented perspective. Here, individual scores represent an absolute level of competence. A score of 65 implies that the student masters 65% of the knowledge domain of general practice. Reliability implies that the estimated level of competence is reproducible. In the example of Figure 1, reliability is less than for the norm-oriented perspective, because the absolute score for all the students is not identical for test one and test two.

*Reliability of pass/fail decisions:* the decision-oriented perspective. In this case, reliability refers only to the accuracy of the decision related to the score being above or below an established cut-off score. The absolute level of competence, its 'distance' to the cut-off score, is irrelevant: if the cut-off score in the example is 50 all three students pass on test one, but student B and C fail on test two, which affects reliability.

In addition, reliability of group mean scores were estimated from the three perspectives mentioned above.

## Statistical Analysis

For estimation of the reliability of individual scores for each of the three tests, an all random Person-by-Item analysis of variance (ANOVA) was carried out (p x i design) followed by variance component estimation. Earlier analyses indicated that case and item scores yielded equal reliability estimates. This implies that it makes no difference whether clusters of items related to a specific case or isolated items were taken as 'entity'. Therefore, simple item scores were used for analysis here. The variance components were pooled across the three tests by averaging the components. Generalizability coefficients (norm-referenced) and dependability coefficients (domain-referenced and decision-referenced), as well as Standard Errors of Measurement (SEMs) were subsequently estimated following regular procedures (Brennan, 1983). Reliability from the decision-oriented perspective was estimated using several cut-off scores. The SEM reflects the size of the measurement error and may be used to estimate a confidence interval for individual scores. Adding and subtracting the SEM to/from a single examinee's score provides the 67% confidence interval. Multiplying the SEM by 1.96 (the appropriate z-value under the normal curve) provides the 95% confidence interval.

For estimation of the reliability of group mean scores, (groups of students of the same training level) an all random Person-nested-within-Group-by-Items ANOVA was conducted per test (i x (p:g)) followed by

variance component estimation, which were similarly pooled across tests. Subsequently, generalizability coefficients, dependability coefficients and SEMs were estimated. Generalizability, dependability and SEMs are presented as a function of item sample and number of students within groups.

## RESULTS

Table 1 presents the mean scores percentage correct minus incorrect- and standard deviation for the three tests of the participating groups that were selected for analyses. The scores increase with training level for all three tests. The total mean scores of the three tests differ, indicating that there is a difference in item difficulty between the tests.

### Reliability of Individual Scores

Table 2 shows the reliability of individual scores estimated from the norm-oriented and from the domain-oriented perspective.

The reliability estimates from a norm-oriented perspective appear to be low. The degree of imprecision can be derived from the corresponding Standard Error of Measurement (SEM). The confidence interval for a test of 160 items is 12 (2 x SEM) and 24 (4 x SEM) for the 67% and 95% confidence level. This implies that a given score of for example 55 (percentage correct minus incorrect items) refers to a true score that lies between 49 and 61 with 67% certainty, and between 43 and 67 with 95% certainty. About the same applies for the reliability estimates from a domain-oriented perspective, with consequences for the absolute interpretation of test scores.

Table 3 shows reliability estimates from the decision-oriented perspective using different cut-off scores.

Table 2. Reliability Coefficients and Standard Errors of Measurement (SEM) for Individual Scores from a Norm and Domain-oriented Perspective, as a Function of Numbers of Items and Testing Time (Hours).

| Items | Hours | Norm-oriented | | Domain-oriented | |
|---|---|---|---|---|---|
| | | generalizability coefficient | SEM | dependability coefficient | SEM |
| 80 | 1 | .55 | 8 | .50 | 9 |
| 160 | 2 | .71 | 6 | .67 | 7 |
| 240 | 3 | .79 | 5 | .75 | 5 |
| 320 | 4 | .83 | 4 | .80 | 5 |

Table 3. Dependability Coefficients for Individual Scores from a Decision-oriented Perspective Using Different Cut-off Scores (Mean over Three Tests) as a Function of Number of Items; Mean score = 44% (Percentage Correct Minus Incorrect Items).

| Items | Dependability coefficients | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 80 | .89 | .77 | .67 | .56 | .52 | .60 | .71 | .80 | .90 |
| 160 | .94 | .87 | .80 | .72 | .67 | .75 | .83 | .89 | .95 |
| 240 | .96 | .91 | .86 | .79 | .77 | .82 | .88 | .92 | .96 |
| 360 | .97 | .93 | .89 | .89 | .81 | .86 | .91 | .94 | .97 |
| cut-off scores | 20 | 30 | 35 | 40 | 45 | 50 | 55 | 60 | 70 |

Table 4. Reliability Coefficients (SEM in Brackets) of Group Mean Scores from a Norm-oriented Perspective as a Function of Number of Individuals per Group and Test Length (Number of Items).

| Items | Reliability coefficients | | | | |
|---|---|---|---|---|---|
| 80 | .67 (4) | .75 (3) | .79 (3) | .82 (3) | .85 (2) |
| 100 | .69 (4) | .77 (3) | .81 (3) | .84 (2) | .86 (2) |
| 120 | .71 (3) | .78 (3) | .82 (3) | .85 (2) | .87 (2) |
| 140 | .72 (3) | .79 (3) | .83 (2) | .86 (2) | .88 (2) |
| 160 | .73 (3) | .80 (3) | .84 (2) | .87 (2) | .89 (2) |
| 240 | .76 (3) | .82 (3) | .86 (2) | .88 (2) | .90 (2) |
| 360 | .77 (3) | .83 (2) | .87 (2) | .89 (2) | .91 (2) |
| Individuals | 10 | 15 | 20 | 25 | 30 |

Table 5. Dependability Coefficients (SEM in Brackets) of Group Mean Scores from a Domain-oriented Perspective as a Function of Individuals per Group and Test Length (Number of Items).

| Items | Dependability coefficients | | | | |
|---|---|---|---|---|---|
| 80 | .50 (5) | .54 (5) | .56 (5) | .58 (5) | .59 (5) |
| 100 | .54 (5) | .58 (5) | .61 (4) | .63 (4) | .64 (4) |
| 120 | .57 (5) | .62 (4) | .64 (4) | .66 (4) | .67 (4) |
| 140 | .60 (4) | .64 (4) | .67 (4) | .69 (4) | .70 (4) |
| 160 | .62 (4) | .67 (4) | .69 (4) | .71 (4) | .72 (3) |
| 240 | .67 (4) | .72 (3) | .75 (3) | .77 (3) | .78 (3) |
| 360 | .71 (3) | .76 (3) | .79 (3) | .81 (3) | .82 (3) |
| Indiv. | 10 | 15 | 20 | 25 | 30 |

Reliability varies depending on the position of the cut-off score: the more distant the cut-off score is from the mean (here 44%), the more reliable pass-fail-decisions are. With the given test length of 160 items a cut-off score of 35 or 55 yields reliability estimates of about .80.

**Reliability of Group Mean Scores**

Tables 4, 5, 6 and 7 present reliability coefficients and SEMs for group mean scores as a function of the number of items and the number of students within groups, from the norm-oriented, domain-oriented and decision-oriented perspective.

Reliability estimates of group mean scores from the norm-oriented perspective (Table 4) attain .80 with groups of 15 individuals at a test length of 160 items. A test length of 80 items may be sufficient to compare groups of 25 individuals each. It is noteworthy that the differences in

Table 6. Dependability Coefficients for Group Main Scores from a Decision-oriented Perspective Using Different Cut-off Scores, as a Function of Number of Items and Individuals per Group. Individuals = 15; Mean Score = 44% (Mean % Correct Minus Incorrect over Three Tests).

| Items | Dependability coefficients | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 80 | .96 | .90 | .82 | .65 | .55 | .72 | .86 | .92 | .97 |
| 160 | .98 | .94 | .88 | .75 | .67 | .82 | .91 | .95 | .98 |
| 240 | .98 | .95 | .91 | .80 | .73 | .85 | .93 | .96 | .98 |
| 360 | .99 | .96 | .92 | .83 | .77 | .88 | .94 | .97 | .99 |
| Cut-off scores | 20 | 30 | 35 | 40 | 45 | 50 | 55 | 60 | 70 |

Table 7. Dependability Coefficients for Group Main Scores from a Decision-oriented Perspective Using Different Cut-off Scores, as a Function of Number of Items and Individuals per Group. Individuals = 30; Mean Score = 44% (Mean % Correct Minus Incorrect over Three Tests).

| Items | Dependability coefficients | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 80 | .97 | .92 | .89 | .69 | .60 | .76 | .88 | .93 | .97 |
| 160 | .98 | .95 | .91 | .80 | .73 | .85 | .93 | .96 | .98 |
| 240 | .99 | .96 | .93 | .85 | .79 | .89 | .95 | .97 | .99 |
| 360 | .99 | .97 | .95 | .88 | .83 | .91 | .96 | .98 | .99 |
| Cut-off scores | 20 | 30 | 35 | 40 | 45 | 50 | 55 | 60 | 70 |

SEM are small: between 80 items for 10 individuals and 360 items for 25 individuals the difference is only 2.

Reliability estimates from the domain-oriented perspective are lower, as was to be expected.

Tables 6 and 7 demonstrate that the reliability of group mean scores, even of relatively small groups, from a decision-oriented perspective, is sufficient for cut-off scores 10% from the mean.

## DISCUSSION

The reliability estimates vary considerably with the object of measurement and the perspective taken (norm, domain or decision-oriented). Starting from a given test length of 150-160 items, corresponding with two hours of testing time, reliability of individual scores from a norm-oriented and domain-oriented perspective seems problematic. Reliability from a decision-oriented perspective with a fair range of cut-off scores seems satisfactory. It might be concluded that the overall reliability for individual scores is insufficient. However, relatively low reliability sometimes yields confidence intervals that are not substantially more extended than those related to a (generally accepted) reliability of .80. It shows that this often used benchmark reliability coefficient is rather arbitrary. It should be noted that reliability is also a function of the population. It is usually lower in a homogeneous population than in a heterogeneous population. Using the test to compare postgraduate students and certified general practitioners may yield better reliability estimates.

Several solutions to increase reliability may be suggested. The most obvious solution is lengthening of the test. This might, however, prolong the testing time unduly and is not feasible in the present situation. An alternative is to aggregate the longitudinal information of two or three successive knowledge tests, across the four months intervals, although during this period the level of knowledge changes under the influence of training. A composite score of two or more tests, therefore, represents a combination of the actual level of knowledge and the growth in knowledge during the last four to eight months. For this purpose, tests should be equated to correct for differences in item-difficulty. A third alternative is using composite scores of different tests from a battery, e.g. knowledge and skills tests. A last alternative is to adopt the decision-oriented perspective. High reliability is achievable, depending on which cut-off score is selected. Here, however, it is validity that poses the essential dilemma: which cut-off score is realistic and tenable? In other words, can it be

made plausible that a score below the cut-off score represents a level of knowledge that is insufficient for good performance? This issue needs further investigation.

The reliability of group mean scores is satisfactory. If fewer items are included more students should participate and vice versa. Given a test length of 150-160 items, present group mean scores may be used in the context of programme evaluation.

The conclusion seems warranted that care should be taken in basing judgements about the student's level of knowledge on their individual test scores. Feedback may consist of the student's actual score and the total group mean score whereas the accompanying confidence interval might be reported to give the students insight into the significance of their results. The feasibility of composite scores should be examined, as well as the validity of different cut-off scores. The use of group mean scores to evaluate the training programme should be encouraged. Either different programmes may be compared or the effect of a programme may be evaluated in a pretest-posttest design. It depends on the educational goal which perspective is used; one could rank-order groups, compare their absolute score level or decide wether they have acquired sufficient versus insufficient knowledge (pass/fail) on the basis of their programme.

Assessment of test-reliability, using generalizability theory, seems a fruitful procedure, which enables faculty to establish the potentials of the test in relation to different educational goals.

## REFERENCES

Brennan, R.L. (1983). *Elements of Generalizability Theory.* Iowa City, Iowa: American College Testing Publications.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements. Theory of Generalizability for Scores and Profiles.* New York: Wiley.

Dubois, V., Everwijn, S., Van Geldorp, G., Groeneveld, Y., Grol, R., Pieters, R., Pollemans, M., Verheij, Th., & Van der Werve, T. (1987). *The Construction of a new Curriculum of Vocational Training for General Practice in the Netherlands.* Utrecht: Royal Dutch Medical Association.

Kramer, A.W.M., & Pollemans, M.C. (1990). Nation-wide progress tests assessing knowledge in vocational training for general practice. In W. Bender, R.J. Hiemstra, A.J.J.A. Scherpbier & R.P. Zwierstra (Eds), *Teaching and Assessing Clinical Competence* (280-282). Groningen: BoekWerk Publications.

Pollemans, M.C., Van Geldorp, G., & Tan, L.H.C. (1988). Naar kwaliteitsbewaking van de beroepsopleiding tot huisarts. [Towards quality assessment of vocational training of general practitioners]. *Medisch Contact, 43*(46), 1429-1430.

Thorndike, R.L. (1988). Reliability. In J.P. Keeves (Ed.), *Educational Research, Methodology, and Measurement. An International Handbook* (330-343). Oxford: Pergamon Press.

Van der Vleuten, C., & Verwijnen, M. (1990). A system for student assessment. In C. van der Vleuten, & W. Wijnen (Eds.), *Problem-based learning: Perspectives from the Maastricht experience* (27-49). Amsterdam: Thesis.

Van Leeuwen, Y.D., Pollemans, M.C., Mol, S.S.L., Eekhof, J.A.H., Grol, R., & Drop, M.J. (1995). The Dutch knowledge test for general practice: issues of validity. *European Journal of General Practice, 1*, 113-117.

Van Leeuwen, Y.D., & Van Hessen, P.A.W. (1990). Clinical competence and objective questions: tac-tics to realize a true/false format assessing competence. In W. Bender, R.J. Hiemstra, A.J.J.A. Scherpbier, & R.P. Zwierstra (Eds.), *Teaching and Assessing Clinical Competence* (233-236). Groningen: BoekWerk Publications.