Selection for postgraduate training for general practice: the role of knowledge tests

VAN LEEUWEN Y D

MOL S S L

POLLEMANS M C

VAN DER VLEUTEN C P M

GROL R

DROP M J

SUMMARY

Background. *Postgraduate training for general practice is a legal requirement in most countries of the European Community, and includes posts in general practice as well as in hospitals. The effectiveness of the training has not been fully evaluated, and it is largely unknown whether the results are satisfactory or what the impact of the separate training components is — nor is it known which characteristics or prior achievements of the trainee influence the end-of-training performance.*

Aim. *To determine the value of knowledge tests in the context of entry selection for postgraduate training in general practice.*

Methods. *Three (equated) knowledge tests were administered during the two years' postgraduate training of 85 Dutch trainees. The first test was taken at entrance, the second eight months later, and the third shortly before the end of the entire training period. Complete data for 67 trainees were available for analysis. A multiple regression analysis was performed to estimate the predictive values of test 1 and test 2 scores, separately and in combination, for test 3 scores. Since the knowledge test may be used for selection purposes, the analysis was repeated using logistic regression with two pass/fail criteria: a 'minimum criterion' and an 'exellence criterion'.*

Results. *Neither of the two analyses yielded a predictive value of test 1 that was high enough to warrant the use of knowledge tests in the context of entry selection. A 'below minimum' score on test 2 correlated 100% with a 'below minimum' score on test 3. However, the positive predictive value of an above minimum score on test 2 was only 86%.*

Conclusions. *The knowledge tests used in this study are not suitable in the context of entry selection. However, trainees that score 'below minimum' after eight months of training may be regarded as 'at risk' in that they will probably score 'below minimum' at the end of training.*

Y D van Leeuwen, MD, general practitioner, Department of General Practice; S S L Mol, MD, general practitioner, Department of General Practice; M C Pollemans, MD, medical educationalist, Department of General Practice; C P M van der Vleuten, PhD, educationalist, Department of Education and Educational Research; M J Drop, PhD, sociologist, Department of Medical Sociology, University of Limberg, The Netherlands. R Grol, PhD, psychologist, Centre for Research on Quality in General Practice, University of Nijmegen and Limberg, The Netherlands.

Submitted: 20 February 1996; accepted: 14 November 1996.

Keywords: vocational training; knowledge test; selection.

Introduction

POSTGRADUATE training for general practice is a legal requirement in most countries of the European Community.[1] The structure and content of the training vary from country to country, but nearly all training schemes include posts in general practice as well as in hospitals. Although its value for general practice is unquestioned, the effectiveness of postgraduate training has only marginally been evaluated. It is largely unknown whether the result of the training is satisfactory or what the impact of the separate training components is — nor is it known which characteristics or prior achievements of the trainee influence the end-of-training performance.

The relationship between 'process' and 'outcome' has been investigated in a study recently performed in the Netherlands.[2,] Growth in knowledge during the first eight months of training was chosen as an outcome measure. The process variables included in the study were the number of patient encounters made by the trainee; the trainer and training practice; and the trainee's education. No convincing relationship was found between process and outcome. However, the trainee's level of knowledge at the start of training appeared to explain a substantial part of the variance in the level of knowledge after eight months (25%). A similar result has been found by Leigh *et al*[3] for the relationship between in-training examination results for family residents and their certification results, and again for the relationship between the level of knowledge of GPs at certification and their level of knowledge as demonstrated at recertification examinations six or more years later.

These findings give rise to the question whether the trainees' level of knowledge at certification may be predicted by their knowledge test scores at entry to, or in the course of, their training. A positive answer might warrant the use of knowledge tests in the context of selection. In most European countries, selection at entry to postgraduate training for general practice is based on the assessment of a written application and a structured interview, which focuses on motivation and prior professional experience. The agreement between interviewers is known to vary considerably, while the predictive validity of this selection procedure for performance during the training is notoriously poor.[5,6] Selection during training is mainly based on assessment of the trainee's performance by the GP trainer.[7,8] Although the trainer seems the most appropriate judge of the trainee's performance, the trainer's simultaneous role of colleague, supervisor and assessor, hinders an objective assessment. Complementary, preferably objective, assessment methods would be desirable. The question is whether knowledge testing would be a valuable addition to the selection methods already used. In other words, can the use of knowledge tests improve the accuracy of the decision to allow trainees either to enter into or to continue their training? The answer depends, in part, on the predictive validity of these tests.

This study investigates the predictive value of two knowledge tests, taken at entry and after eight months of training, for the knowledge possessed by a trainee shortly before certification.

Methods

Context of the study

Since September 1994, postgraduate training in general practice in the Netherlands has consisted of three training blocks of 12 months each. The first and third are spent in general practice, the second in hospitals and other non-primary care settings. At the time of this study (1992), the entire training period was two years, with three blocks of eight months. The structure and content were largely the same throughout. During the entire training period, one day a week was reserved for academic education through day-release courses at the training institute (a the department of general practice at one of the eight faculties of medicine in the Netherlands). At the end of the first block, the training institute formally decided whether or not to allow the trainee to continue training. This decision was based on the trainer's judgement of the trainee's performance and professional growth during the first block. There was no summative assessment at the end of the entire training period.

Subjects

A longitudinal study was conducted with the participation of all 85 trainees who started their training in January 1992 at one of the training institutes in the Netherlands.

Instruments

Knowledge was assessed by the National Knowledge Test for trainees in general practice; this is a paper and pencil test that is routinely administered to all GP trainees in the Netherlands at fixed intervals during their training.[9] The test is used for feedback purposes, not for pass/fail decisions.

The test consists of about 80 patient cases as they are usually presented to the general practitioner (GP), followed by one or more items. The items (160 in total) focus on the key features of the problem.[10] Test content is selected on the basis of a multidimensional blueprint, established by consensus among GPs.[11] The test is designed to assess progress,[12] and is set at the level of the qualified GP at the moment of certification. Successive tests are similar in format but vary in content, and all trainees take the same test regardless of their training level. The questions are of the true or false type with an additional 'don't know' option. The final score is the number of correct answers minus the number of incorrect answers expressed as a percentage of the maximum score. The validity of the test is reported in detail elsewhere.[13] Until now, the knowledge test has only had an educative function, enabling the trainees to detect their strengths, weaknesses and progress.

Procedure

Three knowledge tests were included in the study: the first taken shortly after the start of training (test 1), the second eight months later (test 2), and the third at the end of the entire two-year training period (test 3).

Analyses

To allow correction for item difficulty, the tests were equated through a horizontal linear equation procedure with an anchor test consisting of 20% of the items of each test.[14] The equation was performed on pairs of tests, each pair containing the same anchor test. The group mean scores per test (expressed as described above) were computed as well as the standard deviation and the 95% confidence interval. The correlations (Pearson's *r*) between the three tests were computed, including a correction for attenuation (unreliability).

A multiple regression analysis was conducted, with test 3 as a dependent variable and tests 1 and 2 as independent variables. In addition, a logistic regression analysis was performed with the same independent variables. In this analysis, the dependent variable (test 3) was dichotomous: 'pass' or 'fail'. As cut-off scores, two criteria were chosen. The first was the trainees' group mean score minus one standard deviation: the 'minimum criterion'. Scoring below this implied that the trainee belonged to the 16% poorest-scoring examinees. The second was the trainees' group mean score plus one standard deviation: the 'excellence criterion'. Scoring above this implied that the trainee belonged to the 16% best-scoring examinees.

Results

All three tests were taken by 67 of the 85 trainees. At one institute, test 3 could not be administered for practical reasons, resulting in missing data for 12 trainees. The remaining missing data were due to trainees being absent because of illness, pregnancy, etc. The group mean scores on test 1 and test 2 for the response and non-response group did not differ significantly (test 1: $t = 0.05$; test 2: $t = 0.64$; df = 13.66; $P>0.05$).

Table 1 shows the mean test scores, the standard deviations and 95% confidence intervals for the three tests. There is a significant increase in score throughout the training period; this increase is most pronounced in the first eight months.

Table 2 shows the reliability (Cronbachs alpha) of each of the three tests, and the Pearson's correlations between the three tests, uncorrected and corrected for attenuation (unreliability). The correlations are all significant. The highest correlation is found between tests 2 and 3.

Table 3 shows the results of the stepwise multiple regression analysis. The explained variance (R^2) of test 3 in terms of tests 1 and 2 is 20%; this is largely due to the contribution of test 2 (18%). Again, the best fitting model in the logistic regression analysis includes test 2 only. Adding test 1 does not enhance the prediction (minimum criterion: $\chi^2 = 0.8$, df = 1, $P = 0.38$; excellence criterion: $\chi^2 = 0.12$, df = 1, $P = 0.73$).

Table 4 shows the results of the logistic regression analyses using the 'minimum criterion' and the 'excellence criterion' in

Table 1. A comparison of month of test administration, number of test items, test mean score (M), standard deviation (SD) and 95% confidence interval (95% CI) on the three (equated) tests (n = 67 trainees).

Month	Items	M	SD	95% CI
1	146	46.7	9.4	44.3–49.1
9	153	58.1	8.4	56.1–60.1*
21	148	62.5	9.0	60.3–64.7*

*Significant difference compared with preceding test(s) (P<0.05).

Table 2. Reliability (α, bold diagonal entries), correlations (Pearson r, in upper corner) and correlations after correction for attenuation (italics, in lower corner) between the three tests (n = 67).

	Test 1	Test 2	Test 3
Test 1	**0.68**	0.36*	0.28*
Test 2	0.51*	**0.71**	0.43*
Test 3	0.39*	0.60*	**0.73**

*Significant (P<0.05).

terms of odds ratios. An odds ratio of more than 1 indicates that the test has a predictive value that is more than a 'random' prediction. As is shown, the odds ratios barely exceed 1; test 1 even yields confidence intervals of less than 1.

An alternative way to present predictive power is to show the parameters used for diagnostic tests; for example, sensitivity and specificity. A cross-tabulation with test 3 as 'gold standard' and test 2 as predictor is presented in Table 5. The scores on test 2 were assigned to 'pass' and 'fail' on the basis of the best cut-off score produced by the logistic regression analysis: number of correct – incorrect answers equals 42.6% of the total for the minimum criterion and 69.6 % for the excellent criterion.

For the minimum criterion, the sensitivity for 'pass' is 55/55 x 100% = 100%, implying that all those who finally passed also passed on test 2. The negative predictive value (fail) of test 2 is also 100%; there are no false negatives. The specificity, however, is only 3/12 x 100% = 25%, implying that of those who finally failed only 25% also failed test 2. Conversely, the positive predictive value is 55/64 x 100% = 86%, meaning that of those who passed test 2 only 86% also passed on the final test. The sensitivity and specificity for the excellence criterion are 29% and 98% respectively, with corresponding percentages for false negative and false positive predictions of 16 (10/62 x 100%) and 20 (1/5 x 100%).

Discussion

The results presented here warrant the following conclusions concerning the predictive validity of the 'entry test' and the 'eight month test'. The entry test has little predictive value for the test at certification. The eight month test does contribute significantly to the prediction of 'pass' or 'fail' at certification. Both regression analyses, however, show that the use of this test as predictor is not substantially better than random prediction. Using the minimum criterion, the predictive power of test 2 is, at first sight, not unfavourable if the first concern is to minimize the number of trainees who are unjustly failed after test 2 (0% false negatives). If, however, the major concern is to minimize the number of trainees who fail on test 3 having passed on test 2, the outcome is less favourable (14% false positives). Reducing the latter (for example to less than 5%) implies increasing the former (fewer than 5% false positives would mean more than 75% false negatives!).

The acceptability of the figures depends on the interests at stake. Boards responsible for the outcome of the training should decide what price to pay in terms of percentages of 'incorrect' decisions for selection or early detection of either poor or excellent performers. Alternative criteria for 'pass' and 'fail' may have yielded quite a different outcome. The scores of experienced GPs, for example, might provide such a criterion. In a cross-sectional study performed earlier, however, these GPs scored lower than trainees at certification.[15] Nearly all trainees would have met this criterion, which yields redundant predictions.

Moderate but significant correlations between tests, as found in this study as well as in others,[2,3,4] do not imply that the predictive power of the tests at stake are high enough to justify using the test scores for selection of individual trainees. Parameters like the positive and negative predictive values do give a better insight into the consequences of pass/fail decisions. Regrettably, the problem of selection at entry remains unsolved. No instrument is yet known to have an acceptable predictive value for such an important decision. The results of this study are certainly no reason to alter this statement. For selection during training, assessment of performance by the trainer is generally accepted and should be maintained.

However disappointing the suitability of the tests as selection instruments, their value as instruments of feedback remains unaffected. Trainees have the opportunity to monitor their own performance and to detect their strengths and weaknesses. This may contribute to the efficiency of their self-directed learning. Moreover, monitoring the growth in knowledge of groups of trainees indicates the contributions made by the different training phases. The relatively large increase in knowledge during the first training phase, as demonstrated here, suggests that the knowledge required for general practice is decidedly different from knowledge acquired during undergraduate training.

Table 3. Stepwise multiple regression analysis with test 3 as dependent variable, and tests 1 and 2 as independent variables.

Independent variables	R^2	b	SE	ß	P
Test 1	0.02	0.13	0.14	0.14	0.250
Test 2	0.18	0.41	0.12	0.38	0.002

R^2 = explained version, b = estimate, SE = standard error of estimate, ß = standardized estimate, P = level of significance.

Table 4. Logistic regression analysis for the best model, using the minimum and the excellence criterion for test 3: odds ratios, with their confidence intervals and levels of significance (P).

	Odds ratio	CI	P
Minimum criterion			0.15
Model with test 1 as predictor	1.05	0.98–1.14	
Model with test 2 as predictor	1.13	1.02–1.25	
Excellence criterion			0.09
Model with test 1 as predictor	1.06	0.99–1.13	
Model with test 2 as predictor	1.14	1.05–1.24	

Table 5. Cross-tabulation of the predictor (test 2) versus the (gold) standard (test 3), using the minimum and excellence criterion.

Predictor (test 2)	'(gold) standard' (test 3)					
	Minimum			Excellence		
	below	above	total	below	above	total
Negative	3	0	3	52	10	62
Positive	9	55	64	1	4	5
Total	12	55	67	53	14	67

References

1. Boerma WGW, De Jong FAJM, Mulder PH. *Health care and general practice across Europe.* Utrecht: NIVEL/NHG, 1993.
2. Van Leeuwen YD. *Growth in knowledge of trainees in general practice: figures on facts.* [Thesis.] Maastricht: Datawyse, 1995.
3. Leigh TM, Johnson TP, Pisacano NJ. Predictive validity of the American board of family practice in-training examination. *Acad Med* 1990; **65:** 454-457.
4. Leigh TM, Young PR, Haley JV. Performance of family practice diplomates on successive mandatory recertification examinations. *Acad Med* 1993; **68:** 912-919.
5. Wright PM, Lichtenfels PA, Pursell ED. The structured interview: additional studies and a meta-analysis. *Journal of Occupational Psychology* 1989; **62:** 191-199.
6. Landsbergen EP. Nailing red jelly to the wall. Selectie voor de medische vervolgopleiding. (Selection for medical postgraduate training.) In: Van der Vleurten CPM, Scherpbier AJJA, Pollemans MC (eds). *Gezond onderwijs.* Houten: Bohn, Stafleu, Van Loghum, 1992.
7. Pereira Gray DJ. *Training for general practice.* London: Butler & Tanner Ltd, 1982.

8. Van Geldorp G (ed). *Opleiden en leren in de huisartspraktijk. (Training and learning in general practice.)* Utrecht: Bunge, 1985.
9. Kramer AWM, Pollemans MC. Nationwide progress tests assessing knowledge in vocational training for general practice. In: Bender W, Hiemstra RJ, Scherpbier AJJA, Zwierstra RP (eds). *Teaching and assessing clinical competence*. Groningen: BoekWerk Publ, 1990.
10. Bordage G, Page G. An alternative approach to PMPs: the 'key features' concept. In: Hart IR, Harden RM (eds). *Further developments in assessing clinical competence.* Montreal: Heal Publications, 1987.
11. Pollemans MC. *Kennistoetsing bij huisartsen [Dissertation]. (Testing of knowledge of general practitioners.)* Maastricht: Datawyse/Universitaire Pers Maastricht, 1994.
12. Van der Vleuten C, Verwijnen M. A system for student assessment. In: Van der Vleuten C, Wijnen W (eds). *Problem-based learning: perspectives from the Maastricht experience.* Amsterdam: Thesis, 1990.
13. Van Leeuwen YD, Pollemans MC, Mol SSL, *et al.* The Dutch knowledge test for general practice: issues of validity. *European Journal of General Practice* 1995; **1:** 113-117.
14. Crocker L, Algina J. *Introduction to classical and modern test theory*. Orlando: Holt, Rinehart and Winston, 1986.
15. Van Leeuwen YD, Mol SSL, Pollemans MC, *et al.* Change in knowledge of general practitioners during their professional career. *Fam Pract* 1995; **12:** 313-317.

Acknowledgements

Many thanks to H Düsman and P Portegijs for their help with the statistical analyses.

Address for correspondence

Dr Y D van Leeuwen, Ransdalerstraat 10, 6312 AH Ransdaal, The Netherlands.