

Relative or Absolute Standards in Assessing Medical Knowledge Using Progress Tests

A.M.M. MUIJTJENS, R.J.I. HOOGENBOOM*, G.M. VERWIJNEN** and
C.P.M. VAN DER VLEUTEN*

*Dept. of Medical Informatics, *Dept. of Educational Development & Research, and*

***Skillslab, University of Maastricht, P.O. Box 616, 6200 MD Maastricht, The Netherlands.*

E-mail: arno.muijtjens@mi.unimaas.nl

Abstract. Norm-referenced pass/fail decisions are quite common in achievement testing in health sciences education. The use of relative standards has the advantage of correcting for variations in test-difficulty. However, relative standards also show some serious drawbacks, and the use of an absolute and fixed standard is regularly preferred. The current study investigates the consequences of the use of an absolute instead of a relative standard. The performance of the developed standard setting procedure was investigated by using actual progress test scores obtained at the Maastricht medical school in an episode of eight years. When the absolute instead of the relative standard was used 6% of the decisions changed: 2.6% of the outcomes changed from fail to pass, and 3.4% from pass to fail. The failure rate, which was approximately constant when using the relative standard, varied from 2% to 47% for different tests when an absolute standard was used. It is concluded that an absolute standard is precarious because of the variations in difficulty of tests.

Key words: knowledge assessment, medical education, problem based learning, progress test, standard setting

Introduction

The match between an educational program and the assessment method is vital, because tests and examinations drive student learning. Educational objectives should be reinforced by the assessment program to prevent students from using a 'hidden curriculum' of assessment objectives. To serve this purpose, the University of Maastricht developed the progress test in order to reinforce self-directed learning in the problem-based curriculum (Van der Vleuten et al., 1996). The progress test is a comprehensive examination that reflects the end-objectives of the curriculum. It consists of approximately 250 true/false questions and samples knowledge across all disciplines and content areas in medicine relevant for the medical degree. Growth of medical knowledge is assessed continuously by administering the progress test four times a year during the six year undergraduate medical program. The format of the test precludes the possibility for students to prepare themselves specifically, thus preventing the undesirable effects of objective tests such as memorization of facts and interference with tutorial group functioning. The test is taken by all students, so, in all there are 24 test occasions (6 times 4) during the time course in the program (the undergraduate medical training program requires 6 years of

training). The progress test is used in a formative way to inform students about their progress, possible gaps in their knowledge, and their relative position with regard to the end-objectives. The test is also used to reach pass/fail decisions, requiring the development of a standard (Hambleton and Rogers, 1991; Jaeger, 1989; Livingston and Ziecky, 1989; Meskauskas, 1986). In the distribution of the test score of a group of students, a cutoff score is defined in order to decide who passed and who failed the test. In the past 20 years a relative standard was used by defining the pass/fail cutoff score at the mean minus the standard deviation of the test scores (Wijnen, 1971). The distribution of the students scores is approximately gaussian, so the used relative standard results in a percentage of failing students which is fixed at approximately 15%. This is in sharp contrast with an absolute standard where the location of the cutoff score is fixed, and, as a consequence the failure rate per test may vary. However, a relative standard also shows some serious drawbacks: 1) a fixed fraction of examinees is bound to fail, regardless of their ability, 2) the examinees can deliberately influence the standard, 3) the standard is not known in advance, 4) heterogeneity of the student population reduces the validity of the standard. An absolute standard does not have these drawbacks, but it raises some other questions: 1) how should the cutoff score of the absolute standard in a progress test be obtained, and 2) what are the effects of the absolute standard when it is applied to progress test data.

Several standard setting procedures are proposed in the literature for absolute standards (Hambleton and Rogers, 1991; Jaeger, 1989; Livingston and Ziecky, 1989; Meskauskas, 1986). The best procedures are criterion-referenced, such as Angoff's and Ebel's method, and require panel judgments of individual items. In an annual production cycle of approximately 1000 items these procedures surmount the resources of the medical school. Moreover, the item-judgment procedure in a progress-testing situation is further complicated because a cutoff score is required in all years of teaching (year groups). In the case of the Maastricht school this would entail an item-judgment of 1000 items on six levels of minimal competence. Therefore, only a simple and more arbitrary fixed cutoff score determination procedure is preferred. The consequences of this approach were investigated in the current study. The major research questions were: when an absolute instead of a relative standard is used to obtain pass/fail decisions in a progress test i) do the test outcomes substantially change, and ii) is there a substantial increase of the variation in the failure rate per test.

Method

The question was investigated by applying conditions simulating absolute standards on scores drawn from a database of actual progress test scores. The database was developed at the Maastricht Medical School over a period of 8 years (1986–1993).

In general, no major differences between average progress test scores were found when the results of our students were compared to those of students from

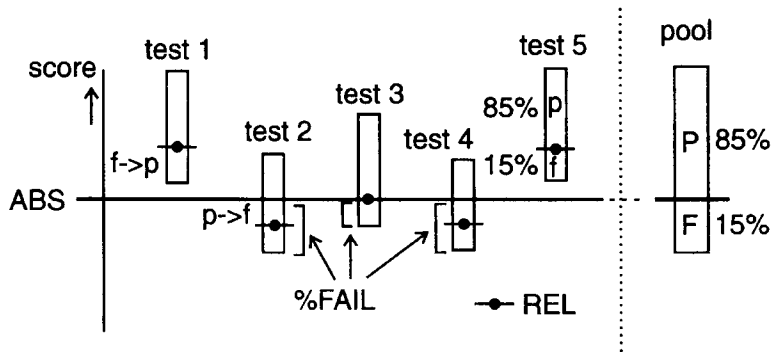


Figure 1. The level of the absolute standard is obtained by pooling the score distributions of several tests and calculating the relative standard for the pool. The example shows a pool of 5 test score distributions. In the current study a pool of 8 tests' results was used for each year of the 6 year undergraduate medical program.

other medical schools in the Netherlands, or to the results obtained by national reference groups of recently graduated doctors (Van der Vleuten et al., 1996). Therefore, the cutoff score of the absolute standard was chosen at the cutting score that retained the overall failure rate at the same level as obtained with the relative standard, i.e. 15%. The method is explained in Figure 1. The vertical axis corresponds to the test score. The boxes represent test score distributions for several consecutive progress tests. The level of the relative standard (REL) for each test is indicated by a dot and line and the pass and fail parts by 'p' and 'f'. The scores of the consecutive tests are pooled and the absolute standard is defined at the mean minus the standard deviation of the pool, as shown at the right. The horizontal line represents the corresponding level of the absolute standard (ABS). When using the absolute instead of the relative standard the test outcome for some students may change as indicated (f→p or p→f). Also the failure rate may vary from test to test. In the current study the old relative standard was compared to the new absolute standard applied to old data. Investigated aspects of interest were: 1) the change in pass/fail decisions, and 2) the variation of the failure rate per test. In the current study, the relative as well as the absolute standards were calculated for each occasion separately. The 8 year data set contains 8 tests for each of the 24 occasions.

Results

In Table I the pass/fail decisions obtained in the study are shown as percentages of the total number of test outcomes ($n \approx 27000$). The rows correspond to the outcomes fail and pass for the relative standard, and the columns refer to the outcomes, had an absolute standard been applied. When using the absolute instead of the relative standard, 2.6% of the outcomes change from fail to pass, and 3.4% from pass to fail. So, in all, 6% of the decisions changed. The histogram (Figure

Table I. Contingency table of test outcomes (pass or fail) obtained with the relative and the absolute standard in progress test data of 8 years, presented in percentages of the total number of outcomes

Relative standard	Absolute standard		Total
	Fail	Pass	
Fail	11.0	2.6	13.6
Pass	3.4	83.0	86.4
Total	14.4	85.6	100.0

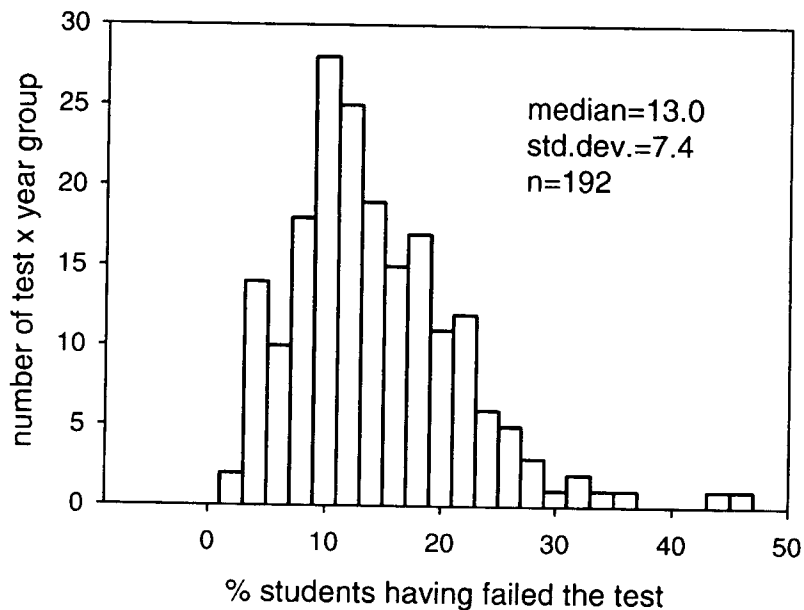


Figure 2. Histogram of the percentage students having failed a test, had an absolute standard been applied. Using the database of 8 years of progress test scores, the total number of units shown in the histogram equals 192 (4 tests per year in 6 year groups over a period of 8 years).

2) shows the distribution of the percentage students having failed a test, had an absolute standard been applied. The number of units in the histogram is equal to 192 (4 tests per year in 6 year groups over a period of 8 years). With approximately 150 students per year group, the total number of test outcomes in the database accounts to approximately 27000. Figure 2 shows that the failure rate varies substantially across tests ranging from almost no failures at all to nearly half of the group that fails.

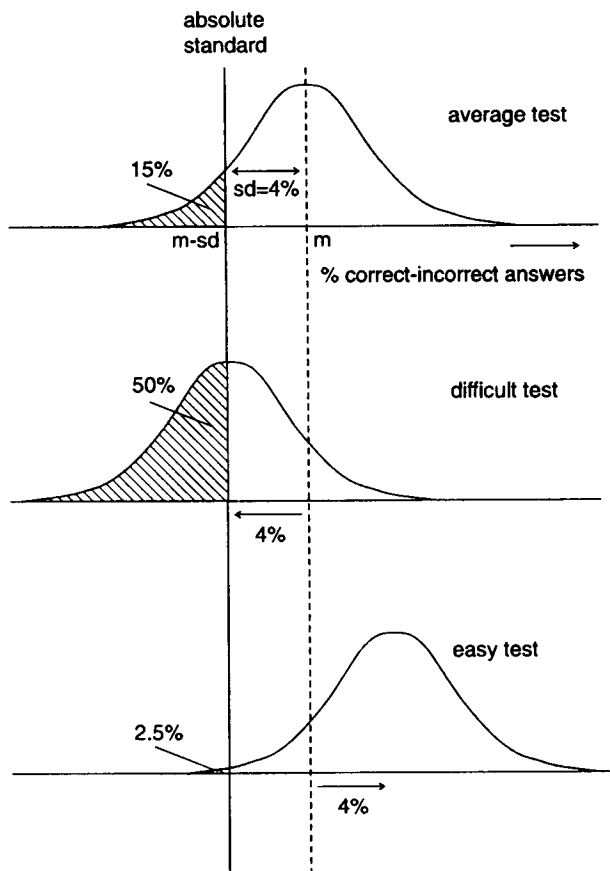


Figure 3. Variation of the failure rate per test when an absolute standard is used. Assuming a) the percentage correct minus incorrect answers per test is gaussian distributed with mean m and standard deviation (sd) equal to 4%, and b) the variation of the test difficulty can be represented by a maximum change of $\pm 4\%$ in the mean value (m) of the score distribution (upper panel), the failure rate per test varies between 2.5% and 50% (lower and middle panel, respectively).

Discussion

The effect of using a fixed absolute standard instead of a relative standard in progress tests was investigated by applying conditions simulating absolute standards to scores drawn from a database of actual progress test scores. The database was developed at the Maastricht Medical School in a period of 8 years (1986–1993). The study does not incorporate the effects on student behavior that a real introduction of an absolute standard might have. So, the obtained results should be interpreted with some reserve.

When the absolute standard was used instead of the relative one a substantial part of the pass/fail decisions changed. Whether these changes were valid or not

can not be answered in the current study. The conclusion is that the pass/failure rate changes considerably as a result of the standard setting method being used.

The large variation of the failure rate per test however is quite disturbing. Three factors come into consideration as a possible cause of this variation: differences between groups of students taking the tests, differences in education, and differences between tests. When comparing the results of several cohorts of students taking the same sequence of progress tests, the average percentage correct score as a function of time showed similar patterns of variation for different cohorts (Muijtjens et al., 1988; Van der Vleuten et al., 1996). The same pattern of variation was found for the average correct score of a national reference group of doctors which recently graduated at one of the dutch medical schools (Van der Vleuten et al., 1996). These observations indicate a test-effect which is of the order of $\pm 4\%$ on the percentage correct minus incorrect scale. In our 8 year data set the standard deviation (sd) of this score was found to range from 3% to 9%. For a gaussian score distribution with an sd of 4% and a variation of the location of $\pm 4\%$ a fixed standard would result in a failure rate ranging from 2.5% to 50% (Figure 3). The observed failure rate found in our study is approximately similar. In conclusion, it is indicated that test-difficulty is a major source of variation while cohort and education effects probably are minor.

Conclusion

The use of fixed absolute standards in progress tests developed in a norm-referenced setting, is precarious because of the variations in difficulty of different tests. Possible remedies for this problem are: 1) constructing a test by selecting items from a bank of items of known difficulty, which enables measurement and control of the test-difficulty, or 2) a more expensive standard setting procedure which is based on item judgment by a panel of experts (Angoff, 1971; Norcini et al., 1987; Swanson et al., 1990). In the absence of these provisions to criterion-referenced testing, the use of a simple fixed cutoff score in progress tests is difficult to maintain and defend.

References

- Angoff, W.H. (1971). Scales, Norms, and Equivalent Scores. In Thorndike, R.L. (ed.) *Educational Measurement*, 508–600. American Council on Education: Washington D.C.
- Hambleton, R.K. & Rogers, H.J. (1991). Advances in Criterion-Referenced Measurement. In Hambleton, R.K. & Zaal, J.N. (eds.) *Advances in Educational and Psychological Testing*, 3–43. Kluwer Academic: Boston.
- Jaeger, R.M. (1989). Certification of Student Competence. In Linn, R.L. (ed.) *Educational Measurement*, third edition, 485–514. McMillan: New York.
- Livingston, S.A. & Zieky, M.J. (1989). A Comparative Study of Standard-Setting Method. *Applied Measurement in Education* 2(2): 121–141.
- Meskauskas, J.A. (1986). Setting Standards for Credentialing Examinations. *Evaluation and the Health Professions* 9(2): 187–203.

- Muijtjens, A.M.M., Imbos, T., Theunissen, M.J.A., & Roos, J.M.A. (1988). *Exploratory Analysis of Progress Test Data*. Onderzoek van Onderwijs nr. 38, Dept. of Educational Development and Research, Maastricht University, The Netherlands.
- Norcini, J.J., Lipner, R.S., Langdon, L.O., & Strecker, C.A. (1987). A Comparison of Three Variations on a Standard-Setting Method. *Journal of Educational Measurement* **24**: 56–64.
- Swanson, D.B., Dillon, G.F., & Postell Ross, L.E. (1990). Setting Content-Based Standards for National Board Exams: Initial Research for the Comprehensive Part I Examination. *Academic Medicine* **65**: S17–S18.
- Van der Vleuten, C.P.M., Verwijnen, G.M. & Wijnen, W.H.F.W. (1996). Fifteen Years of Experience with Progress Testing in a Problem-Based Learning Curriculum. *Medical Teacher* **18**: 103–109.
- Wijnen, W.H.F.W. (1971). *Onder of boven de maat*. Ph.D. diss. University of Groningen, Groningen, The Netherlands.

Address for correspondence: A.M.M. Muijtjens, Ph.D, Dept. of Medical Informatics, University of Maastricht, P.O. Box 616, 6200 MD Maastricht, The Netherlands. Tel.: +31 43 3882235/40; Fax: +31 43 3671052