Notes and Comments—Notes et commentaires

Adjusting the Nominal Maximum for Tests of Unknown Difficulty

GEOFFREY P. MASON*

ABSTRACT

The percentage system of grading is still in common use in post-elementary education, and is likely to remain so. It is suggested that some of the difficulties experienced by instructors in attempting to conform to the arbitrary percentage standards of their institutions while setting tests of unknown difficulty may be overcome by using an adjusted maximum based on the performance of the top students in a class.

RÉSUMÉ

L'échelle des cotes exprimées en pourcentage s'emploie encore fréquemment dans l'enseignement post-élémentaire. D'ailleurs, il est probable que cette pratique continue. L'enseignant éprouve des situations difficiles lorsqu'il essaile de soumettre la correction de ses examens, qui sont d'une difficulté indéterminée, aux normes arbitraires des cotes exprimées en pourcentage utilisées par son institution. La solution suivante se propose pour un certain nombre de situations: la détermination, dans chaque cas, d'un pourcentage maximal ajusté en fonction des meilleurs étudiants dans la classe.

The method of determining standards by the use of percentage scores is still widely used in high schools and universities (e.g., Thorndike, 1969; Terwilliger, 1966). Although the trend is towards *reporting* by letter grade, nevertheless the assignment of the grade is frequently based on a percentage scale of some kind. This percentage scale, more often than not, is prescribed by an institution with explicit statements on transcripts or reports such as, A = 93% – 100%, B = 85% – 92%, and so on.

A continual problem facing any instructor working within such a system is the development of tests of precisely the right difficulty level, in order that the grading policy of the institution will be consistent across instructors and disciplines and great discrepancies in

* Faculty of Education, University of Victoria

the grades allotted not occur from one instructor to another. Another way of stating this is that for an institutional grading system with specified percentages to work, tests from one instructor to another must be of comparable difficulty. If tests vary in difficulty, there can be no consistent definition of the referent for "one hundred percent," and any grades based on standards set in percentages become relatively meaningless.

There have been several attempts to overcome this particular problem (Ebel, 1962; Hively, 1968; Osburn, 1968; Bormuth, 1970). In essence, all attempt to define the population of test items (and hence the knowledges and skills) of a given domain which is to measured. By selecting an appropriate sample from the universe of test items, generalization from performance on the test sample to performance on the universe of items is possible. By using sampling theory (and ignoring a minor problem of sampling error) the referent for "one hundred percent" on the test becomes complete mastery of the universe of items of the domain, and in a similar way any lesser percentage mark on the test sample generalizes to a similar lesser percentage of mastery of the universe of knowledges and skills.

This general approach appears profitable with relatively straight-forward tasks in highly structured disciplines such as mathematics, although even here, considerable work remains to be done. But the approach currently offers cold comfort to the many instructors in disciplines where the development of performance competencies on many tasks can be almost infinitely great so that the concept of learning for mastery is inappropriate. Consider the problem inherent in a definition of the universe of knowledges and skills involved in the writing of an extended literary essay, an understanding of Marxist-Leninism, or the development of an historical perspective, however any of these might be operationally defined. At present it appears extremely difficult, if not impossible, to define the population of learning outcomes which may be subsumed under any one of the above educational objectives without trivializing the objective. If it is not possible to specify the population of behaviours from which to sample, use of sampling theory is unjustified, and any assumption that the percent mark obtained on a test reflects the percentage of the course content known or the percentage of the learning outcomes mastered is on shaky ground indeed.

Another approach to the meaning of "one hundred percent" on a test is in terms of the "reasonableness" of the test items. The maximum possible mark is deemed here to indicate the highest quality of performance on the test which could reasonably be expected of a student who has been exposed to the course content and method. While, in this approach, there should still be an assumption that the test reflects a fair and logical sampling of course content or outcomes, the meaning of a perfect score is not dependent upon this assumption. Very high or very low competence, it is assumed, is indicated by very high or very low marks on the test.

The weakness of this approach is apparent. The level of performance on a test, as indicated by the mark obtained, depends to a great extent on the difficulties of the test items which, unless the items have been used previously and extensively, are unknown. Consequently, the one hundred percent simply refers, in this case, to a maximum performance on test items which the examiner has estimated to be of appropriate difficulty. Any complaints from students about examinations being too difficult challenges the validity of this estimation procedure. It is suggested in this paper that an empirical approach to a

fair and reasonable standard is preferable to one dependent upon prior judgment.

There are several approaches to grading which utilize a knowledge of test results. In essence, all are normative and based either on simple ranking procedures such as A = top 7%, B = next 24%, or on standard score methods such as A = above 1.5z, B = .5z to 1.5z, and so on. Ranking procedures are often inappropriate for homogeneous groups of students, while the use of standard scores as the basis for grade intervals, almost by definition forces half of a group into categories with minus values and frequent negative connotations. Further, comparability over the years with standard scores is based on the assumption of comparable means, which is often untenable for small classes and dubious with large ones.

From the writer's point of view, a better approach to the idea of a very high performance on a given test of unknown difficulty is one which utilizes the top few scores obtained. While this does not eliminate the necessity for subjective judgment, it does provide an empirical base from which the judgment can be made. The proposed procedure, which employs the concept of the "adjusted maximum" is as follows:

1. Test results are examined and the top two, three or four students identified. It is suggested that more than one student's score be considered even though the extreme scores of a distribution tend to have relatively high reliability.
2. The quality of the work of these top few students is assessed using existing information obtained from assignments, previous tests, classroom contributions and, if necessary, interviews. On the basis of this assessment and the marks obtained by the students an "adjusted maximum" is established for the test.
3. Marks for all students are now calculated as percentages of the adjusted maximum rather than of the nominal maximum for the test.
4. Previously established standards in terms of percentages are applied.

An example will clarify this procedure. Let us suppose that a grading policy has been established as follows: A = 91% – 100%, B = 81% – 90%, C = 71% – 80%, etc., and that an examination with a nominal maximum of 165 marks has been given. Marks for the four top students, A, B, C, and D are 134, 132, 130 and 128 respectively. Student A's performance is considered, on the basis of all his work in the course and an interview in which he was questioned on his examination answers, to be at the 99 percent level. Student B, C and D are also considered to be very good students on the basis of the same criteria, so that their performances confirm the judgment with respect to A. In the light of this and A's score of 134, the adjusted maximum is set at 135 and all other raw scores on the test calculated as a percentage of it. Should Student A's work have been considered to be at the 90 percent level instead of the 99 percent level then the "adjusted maximum" would have been set at $134 \times \frac{100}{90} = 149$.

It will be noted that the scaling which occurs is not of the test items but of the students themselves. In justification of this procedure it is believed that many instructors develop a fairly accurate calibration of the performance of their top students from one year to the next which is probably more accurate than their calibration of the mean of each of their classes from year to year. In a small survey conducted by the author, 24 of 30 randomly selected members of an Arts Faculty stated that they used a knowledge of the test achievement of their best students in estimating the difficulty level of a given test.

There are several advantages to the system outlined above:

1. The meaning of one hundred percent becomes clear – it is the test performance expected of a truly excellent student. (It might be noted parenthetically that if a standardization of the meaning of "one hundred percent" could be effected across institutions and disciplines the number of problems pertaining to admissions and scholarships would be substantially reduced.)
2. Students see the process as essentially reasonable. It is difficult to argue that too high a standard is being demanded on too difficult a test in the light of an adjusted maximum to the work of other students.
3. When setting an examination consideration of the difficulty of certain items need not be too inhibiting.
4. It is possible for all members of a highly homogeneous group to receive a high grade. The system per se does not force students into categories designated as "below average", as does, for example, the stanine system.
5. The application of the system is straightforward and easily understood.

Problems

1. There is the difficulty of determining the competence of the top students. However, with a very large group one can reasonably set the top score at or near 100 percent or whatever maximum percentage is customarily given. On one occasion in the last five years the writer found it necessary with an exceptionally able and unusual student whose mark far exceeded that of two or three very good students, to set the adjusted maximum at 95 percent of the student's mark.

With smaller classes the "adjusted maximum" has to be set on the basis of a good knowledge of the competencies of the top two, three or four students. In the writer's experience, this is far less difficult than arbitrarily assigning a difficulty level to a test in the absence of student performance, trying to rate the class average, or justifying the forcing of the grades of students into a predetermined distribution.

2. The "adjusted maximum" procedure has with an objective test such as multiple choice, the effect of tossing out the difficult items for the top scorer if his performance is adjusted to something near 100 percent. If a large adjustment has been made and if the item intercorrelations are low, comparisons of other scores with the "adjusted maximum" may be based on slightly different accumulations of items. There may be no way out of this problem which, in any case, is probably not a serious one as the overlap of the sets would inevitably be very considerable. However, with a subjective type test such as an essay test it could be argued that the "adjustment" is spread throughout each item and over all the items, so that the problem disappears.

References

Ebel, R.L., "Content standard test scores," *Educational and Psychological Measurement*, 1962, *22*, 15-25.

Hively, W., Patterson H.L., & Page, S.H., "A universe-defined system of arithmetic achievement tests," *Journal of Educational Measurement*, 1968, *5*, 275-290.

Osburn, H.G., "Item sampling for achievement testing," *Educational and Psychological Measurement,* 1968, *28*, 95-104.

Terwilliger, J., "Self-reported marking practices and policies in public secondary schools," *Bulletin of the National Association of Secondary-School Principals,* March 1966, 5-37.

Thorndike, R.L., "Marks and marking systems," in R.L. Ebel (Ed.) *Encyclopedia of Educational Research.* (4th. Ed.). New York: The Macmillan Company, 1969.

University Planning: Functional or Futile?

ANDREW GRINDLAY*

Profit-oriented corporations have been in the business of planning for longer than have the universities, mainly because for many of them, it was a matter of survival. It is only in recent years, with declining student enrollments and shrinking research funding, that universities have felt the need for a more careful look at where they are going. When times are good and students are clamoring to get in, people at universities are too busy launching new courses, building new buildings and hiring new faculty to devote much attention to the longer term. When adversity appears imminent, people in both business and universities move toward more formalized methods of deciding their destiny.

Planning can be described on two dimensions. First, it can be said to be either "top-down" or "bottom-up." The former is found in highly authoritarian types of organizations in which the senior administrator and perhaps a few of his close associates map out the future of the organization. They have the authority to implement whatever plans are developed and they do it. "Bottom up" planning is more appropriate to a democratic type of management style. With it, the senior administrators set out the general overall objectives and people who are lower in the organization plan for their own units, these plans being consolidated as they are moved up the organization.

The other dimension by which planning can be described is formality, meaning the degree of formality of the planning process. It can be measured on a continuum, at one end of which is informal planning, with formal planning at the other end. Planning is informal if it is done without documentation at the various steps, if the goals and target dates are kept in someone's head and communicated to nobody else except perhaps a few close associates. At the formal end of the continuum, the planning process is clearly spelled out, goals and objectives are written down and there are documents which are known as constituting "the plan."

Organizations with an authoritarian management style tend to use top-down planning whereas those with a democratic management style plan from the bottom up. Informal planning is usually used when the organization is growing, with the degree of formality

*Professor of Business Administration, The University of Western Ontario.