DOCUMENT RESUME

ED 470 592 TM 034 611

AUTHOR	Winters, R. Scott
TITLE	Score Normalization as a Fair Grading Practice. ERIC Digest.
INSTITUTION	ERIC Clearinghouse on Assessment and Evaluation, College Park, MD.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO	EDO-TM-02-10
PUB DATE	2002-12-00
NOTE	4p.
CONTRACT	ED-99-CO-0032
AVAILABLE FROM	ERIC Clearinghouse on Assessment and Evaluation, 1129 Shriver Laboratory, University of Maryland, College Park, MD 20742. Tel: 800-464-3742 (Toll Free); Web site: http://ericae.net.
PUB TYPE	ERIC Publications (071) -- ERIC Digests in Full Text (073)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	Grades (Scholastic); *Performance Based Assessment; *Scores
IDENTIFIERS	ERIC Digests; Mean (Statistics); *Normalizing Transformation

ABSTRACT

This Digest outlines an appropriate way to handle score normalization in a fair and equitable manner. Using raw scores to calculate final grades may not entirely capture a student's true performance within a class. As variation in performance evaluation increases, so does the impact on the student's final ranking. Ideally, the distribution of individual student performance for all examinations should be equal, and fortunately the methodology for placing diverse assignments on an equitable scale is straightforward. Appropriate normalization requires nothing more than adjusting the examinations' means to be equal as well as their variances. A template for the normalization process is included, and an example derived from real data from a college biology course is given. (SLD)

Score Normalization as a Fair Grading Practice

R. Scott Winters
University of Pennsylvania

ERIC Digest EDO-TM-02-10 December 2002

Score Normalization as a Fair Grading Practice

R. Scott Winters
Department of Biology
University of Pennsylvania

Course instructors want to evaluate students in a manner that is fair and based upon the student's representative performance. Discussions of fair grading practice tend to focus on: grading methodology and individual assignments (i.e., Glenn, 1998), the determination of an appropriate metric and clearly articulating expectations to students (i.e., Davis, 1993). Few guidelines address practical considerations for integrating multiple assignments (e.g., determining final grades based upon multiple exams written by different instructors) and the prerequisite statistical methodologies (but see Cross, 1995). This Digest outlines an appropriate means to handle these situations in a fair and equitable manner. Included is a detailed example, based upon real class data, which illustrates the disparity in grade assignment with and without proper normalization.

All Scores Are Not Equal

While fair grading is easily understood when discussing a single assignment (such as an exam or paper) it becomes a more difficult issue when multiple assignments are considered. For instance, if a student gets a 50 on an exam that is very hard (hence the 50 is the highest grade among all students), and a 60 on a second exam that is very easy (hence the lowest grade among all students), are these exams equitable? If a student is given the option of dropping the "lowest grade" of the two, does it make sense to drop the exam that, a) reflects the lowest numerical score (the 50), or b) reflects poorer performance (the 60)?

If we set our evaluation criterion as a performance measure, then the score reflecting poor performance should be dropped. However, in order to make such an evaluation, the exams need to be converted into a common currency; specifically, they need to be placed upon a standard scale for comparison. Therefore, using raw scores to calculate final grades may not accurately capture a student's true performance within a class. As variation in performance evaluation increases, so does the impact on the student's final ranking.

Ideally, we would like the distribution of individual student performance for all exams to be equal, despite differences in time, instructor, teaching assistant, and other factors. Only then can evaluations be considered comparable. Without this common currency or scale, errors in grade assignment will result. Fortunately, the methodology for placing diverse assignments on an equitable scale is straightforward. Appropriate normalization requires nothing more than adjusting the exams' means to be equal as well as their variances. If different teaching assistants instruct different subsets of the class, then these subsets also need to be standardized for equal means and variances across teaching assistants.

The need for normalization is intuitive to most: an exam with a mean of 40 is not equitable to an exam with a mean of 70. The obvious correction is to readjust the scores such that the means are equal; this is a good first step, but alone, it is insufficient. Equally important is the need to correct for differences in the variances. A template for making such calculations is introduced below.

The Normalization Process

We begin by converting an individual score into a context-free evaluation of relative performance. Next, we will transpose this context-free evaluation into a performance measure (a normalized score) based upon a distribution that we define (that is, we will dictate what the mean and variance are to be). In this manner, scores from different evaluations (exams, instructors, laboratory sections, etc.) can be transposed onto a common scale. When all of the course's evaluations are based upon the same distribution, they can reasonably be compared.

The context-free evaluation we will work with is a z-score. A z-score captures an individual performance relative to the population's mean and variance.

$$z=(X-M)/S$$

where: z refers to the z-score, M is the estimate of the population's mean, S is the estimate of the population's standard deviation, and X is an individual score within the distribution having mean M and variance S.

Since z-scores give us a relative performance measure, then the same z-score can be derived from significantly different distributions. Thus, any score from one distribution can be converted into a score for a second distribution, while maintaining that same relative performance (the same z-score).

For any assignment in a class, we know the absolute score for every student and can estimate the mean and the standard deviation for that assignment based upon all students' scores. Therefore, we can convert each student's absolute score into a z-score. With z-score in hand, we can calculate a new absolute score for any distribution we define. That is, we can declare a mean and standard deviation we wish the new distribution to have and then solve for the absolute numerical value that the z-score would take. This is called the T-score or transformation score.

$$T=m + (s)(z)$$

where: T refers to the transformed score on the new distribution, m is the target mean, s is the target standard deviation, and z is the z-score.

Working through an example--one student

Let us take a specific example of one student's performance on three separate exams where we intend to drop the "lowest" exam score. The vernacular of "lowest exam score" is misleading since our true intention is to drop the grade representing the student's worst performance on any of the three exams. Table 1 gives the student's grades along with the average and standard deviation for the performance of all students on each exam.

Table 1			
	Exam 1	Exam 2	Exam 3
Student's Performance	69	75	72
Class Average	58	66	62
Class Standard Deviation	22	19	9

Normalization begins by choosing an arbitrary average and standard deviation for the distribution we wish to set as our baseline. In this example, an average of 70 and a standard deviation of 15 are selected. In order to normalize the student's performance on exam 1, we simply fill in those values that we have. Thus, for Exam 1, the student's z-score is

$$z = (69-58)/22 = .5$$

and

$$T = 70+(15)(.5) = 77.5$$

While the numerical value may have changed, the student's relative performance (the z-score) has not. A grade of 77.5 within a distribution having an average of 70 and standard deviation of 15 represents the same relative performance as a grade of 69 within a distribution having an average of 58 and a standard deviation of 22.

If we were normalizing the grades of an entire class, then we would use the same equation and change the values for the original grades for each student in order to obtain each student's normalized grade (T-score). Performing similar calculations for Exam 2 and Exam 3 generates normalized scores of 77.1 and 86.67, respectively. Therefore, Exam 2 should be dropped since the student's performance is the lowest.

Working through an example--an entire class

This example illustrates how final scores for individual students can change dramatically depending on whether normalization procedures are adopted.

The example is derived from real data for an introductory biology course taught at a large university and is based upon scores for 205 students. For each student, there are five grades: three exams, a final, and a laboratory score. It is the policy of the department that grades be calculated according to the following criteria:

A. the "lowest" of the three exam scores is to be dropped,

B. each of the two remaining exams is worth the same as the final, and

C. the laboratory score is worth one and one half times any exam (which represents one third of the course evaluation). Complicating the matter is the fact that students are pseudo-randomly assigned to one of seven laboratory instructors. Laboratory instructors vary tremendously in their knowledge, experience, and difficulty. Finally, two instructors co-lectured the course and exams were written independently (with the exception of the final).

For simplicity, let us assume that grades are based upon the following schema: the top 5% will receive an A+, the next 5% an A, the next 15% a B, the next 50% a C, the next 15% a D, and the last 10% an F. In reality, a far more complicated method is – and should be – used that bases an individual's grade on an absolute score rather than a relative measure such as intra-class competition.

Differences in grade assignment between pre-normalization (raw) and post-normalization are profound. Approximately 27% of the class (56 out of 205 students) would have been assigned the wrong grade had the instructors not normalized the scores. In fact, the grades for 52 students changed by one letter grade, and 4 students changed by two letter grades. Looking at one superficial aspect of these dynamics, we note that 37% of students have a different exam score dropped post-normalization. The effects of such changes influence the top, more competitive, tiers. Without normalization, 40% of A+ grades are incorrectly assigned and the ranking of the top three students is incorrect. In fact, the student who performed the best in class would have been wrongly assigned a B without normalization. More dramatically, prior to normalization, another student would have incorrectly been considered average, C, when in fact their work merited an A relative to his or her peers.

Acknowledgements

The author would like to thank Heather Trobert and Michael J. Balsai for helpful comments on an earlier draft of this manuscript.

References

Cross, L. 1995. Grading Students. *Practical Assessment, Research & Evaluation*, 4 (8). Available at: http://ericae.net/pare/getvn.asp?v=4&n=8

Davis, B.G. (1993). *Tools for Teaching*. San Francisco: Jossey-Bass Publishers.

Glenn, B.J. (1998). The golden rule of grading: Being fair. *The American Political Science Association Online*, 31 (4). Available at: http://www.apsanet.org/PS/dec98/glenn.cfm

This publication was prepared with funding from the Office of Educational Research and Improvement, U.S. Department of Education, under contract ED99CO0032. The opinions expressed in this report do not necessarily reflect the positions or policies of OERI or the U.S. Department of Education.