STANDARDIZING TEACHERS' EXAMINATIONS *and the* DISTRIBUTION OF CLASS MARKS

BY
ROBERT S. ELLIS

PUBLIC SCHOOL PUBLISHING COMPANY
BLOOMINGTON, ILLINOIS

STANDARDIZING TEACHERS' EXAMINATIONS
and the
DISTRIBUTION OF CLASS MARKS

BY

ROBERT S. ELLIS, PH. D.
Associate Professor of Psychology,
Colorado College

Edited by GUY M. WHIPPLE

PUBLIC SCHOOL PUBLISHING COMPANY
BLOOMINGTON, ILLINOIS

PUBLIC SCHOOL PUBLISHING COMPANY
Bloomington, Illinois

PREFACE

The past decade has seen rapid strides in the direction of standardizing intelligence tests and educational tests, especially for use in the elementary schools. Thus far, however, this movement has had but little effect on improving the reliability of the tests made by the average teacher. Most of the examining, from the elementary school to the university, is still done by methods which the recent studies in mental measurement have shown to be very fallible. I am a strong advocate of the use of the standardized tests now on the market, but I do not believe they are likely ever to be satisfactory for more than a small part of the total amount of testing that needs to be and will be done. It is, therefore, very important that every teacher understand the elementary principles of constructing and scoring tests, so that the present errors in classroom testing may to a considerable extent be overcome. This little manual is an attempt to supply in convenient form and in language as simple as possible a guide for the construction and scoring of tests, so that the average teacher without special training in psychology, statistics, and mental measurement may be enabled to give very much better tests and with considerable economy of time.

A second problem of almost equal importance is that of converting test scores into marks, and of interpreting the meaning of these marks. At the present time there is even more confusion here than exists with reference to the methods of constructing tests. This manual attempts to present a solution to the problem with a minimal use of statistics and even of arithmetical calculation.

I am indebted to various studies in this general field, but in view of the character of the manual I have thought it best to give references chiefly to a few standard works on statistics and mental measurement. However, a number of references to recent articles in the periodical literature are given at the end of Chapter VII. The student interested in pursuing the matter further will find these very helpful, and he will find in both the texts and the periodical articles further references to the literature.

The manual has been read in manuscript by Dr. M. J. Zigler and by Dr. G. M. Whipple, and both have made constructive suggestions. The writer takes this opportunity to express his obligation for that service.

Colorado Springs, Colorado ROBERT S. ELLIS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Importance of Measurement in Education

Education is an attempt to produce definite changes in individuals. The success of a teacher or of an educational system is proportional to the extent to which these changes are brought about, and intelligent teaching requires that the teacher know the extent to which such changes have been produced. This requires some kind of measurement. In practice, teachers give examinations to determine the extent of the progress of pupils in the course of study. Promotions are based largely on the results of these examinations: a failure causes the pupil either to repeat the work or to quit school; an undeserved promotion places the pupil in a class for which he is not prepared. Under such conditions it is clearly of great importance that these measurements be accurate. For the best teaching it is also desirable that tests be given both as a stimulus to the pupils and as an aid to the teachers in determining how well the subjects taught have been assimilated.

Present Methods of Testing Are Unsatisfactory

The methods of examining now in common use are not satisfactory for two reasons: they are not reliable, and they require too much of the teacher's time. The grading of papers is considered by many to be the worst single phase of a teacher's work. If done in the conventional way, it requires an enormous amount of time and energy, and very often neither of these is available when the teacher has done his regular and more immediately necessary work. When classes are very large, it becomes impossible for the teacher to do very much of this work in the usual way. Even, however, if the teacher has a strong sense of duty and a stronger constitution and grades the papers of numerous tests, the results are of uncertain value as measures of the achievements of students. The grade attached to a paper depends very largely on the teacher grading the paper and on the time when it is graded. Experiments have shown that if a number of teachers grade the same paper, they will assign very different grades; and if the same teacher grades the same papers at different times, the grades will again vary considerably.

Experiments on the Unreliability of Grades

A Sample Experiment

As an indication of the unscientific and unreliable character of ordinary grades, the reader is referred to Table I, which shows the grades assigned to some papers by the members of a class in mental tests. In this case a test was given to the class on the work they had just covered, and at the end of the examination they were told to come prepared

Table I.—Variations in Grades Assigned to the Same Papers by Different Graders

Papers	Grades						
A	83	32	70	40	46	53	63
B	92	70	89	78	81	83	45
C	25	63	81	49	79	75	92
D	63	72	90	87	80	74	77
E	74	47	46	58	57	80	78
F	77	93	88	85	80	80	100
G	97	90	93	93	77	100	87
H	64	75	49	68	95	82	59
I	66	98	69	71	82	80	70
J	42	35	27	60	30	21	41

the next day to grade the papers. The class was composed mostly of experienced teachers of an average age of about thirty years. Nearly all had had a great deal of experience in grading papers. The papers did not have the names of the writers on them, but they were lettered A, B, C, etc. Each person graded seven papers in a fifty-five minute period, but the papers were so arranged that no one graded his own paper. The material on which the examination was based was relatively simple, and the average intelligence of the class as shown by a mental test was above that of the average college class. Yet, in spite of this, they showed the most remarkable variation in the grades assigned to the same papers. Paper A, for example, was marked from 32 to 83 by different graders. Paper B was marked from 45 to 92. In every case there is considerable variation.

The objection will be raised to this experiment that it is not fair since the subject matter of the test was new to those grading the papers. Unfortunately, however, the same outcome appears when experienced teachers grade papers on the subject with which they are most familiar.

Further Experiments

Starch and Elliott established in a convincing manner the unreliability of teachers' marks. A paper in plane geometry was photographed so that facsimile reproductions could be sent to the geometry teachers in each of the high schools of the North Central Association. They were asked to grade the paper on a scale of 100 and according to the standards they were in the habit of using. The paper was graded and returned by 114 teachers. With what result? Mathematics papers had generally been supposed to be of such character that they would show uniformity in grading if uniformity in grading could be found in any subject. But such uniformity in grading was not found. The grades were distributed all the way from 28 to 92. Thirty teachers (26%) assigned grades of 62 and below. Twenty-nine teachers (25%) assigned grades of 78 and above. Assuming a passing mark of 75, then 41 percent of the teachers vote in favor of a pass, while 59 percent of them vote in favor of a failure. If a passing mark of 70 is assumed, they are about equally divided. And these are qualified teachers regularly engaged in teaching geometry—geometry, the subject often supposed to train one in exact thinking!

Starch and Elliott extended their investigations to other subjects with results that were essentially the same as in the case of the geometry paper, except that the range of variation in the other subjects was less. Starch also made an investigation of the grading of freshman English papers in the University of Wisconsin by having ten final examination papers graded by ten instructors of freshman English. One paper was marked by one instructor as low as 20 and by another instructor as high as 68; another paper was marked from 44 to 81. The least variation occurred in the case of a paper that was marked from 79 to 96. The average of the grades assigned by different instructors to the ten papers also varied greatly; for one instructor the average of the ten grades was 65.5, and for another instructor it was 85.1.

The findings of these investigators have been confirmed by numerous other studies; the same general result is found from the elementary school to the university—wherever tests and examinations of the ordinary kind are given.

Examinations as Lotteries

These studies have left no room for doubt that when a student takes an examination, he is investing more or less heavily in a lottery. If the goddess of chance is with him, he needs little merit in order to win; if the goddess is against him, he can be sure of a passing grade only by having an exceptionally good paper. In practice, the success of a student depends to an undue extent on what particular teacher does the grading. Some teachers are low graders; some are high graders; some fluctuate between the two extremes.

The writer once had occasion to compare the term grades in two large freshman science classes. In one case the highest grade in the class was 73; in the other case about one-third of the class received a grade of 100, and the lowest grade in the class was 85. Fifty percent of the grades were above 92. In this case there was no question but that the first class had done more work than the second, but the grades these students received carried with them nothing to indicate whether they were high or low, and the 73 in the first class would ordinarily be rated as inferior to the 85 in the second class in spite of the fact that it undoubtedly represented both vastly more and better work. Such is the condition of much of our grading.

The Causes of Unreliability

To overcome such errors we must find the causes and remove them. There are numerous detailed factors that enter into the production of this condition, but the cause, stated in general terms, is found in the fact that no definite objective standard is used as a basis of assigning marks. Grades depend too much on subjective factors, *i.e.*, on the personal judgment of the teacher. Teachers follow no systematic plan of selecting questions; they grade the papers according to no uniform plan. Under such conditions variability in results is inevitable. The remedy is to be found in the adoption of standardized methods—in the elimination so far as possible of the personal equation. Tests must be made objective. And the distribution of grades must be governed by some definite principle.

The Development of Standard Tests

In recent years rapid progress has been made in the process of standardizing school