COMMENTS

The purpose of this department is to afford an opportunity for informal exchange of ideas on matters related to legal education. Typical comments will range from about 1200 to about 3000 words in length, and may either advocate innovations in curriculum or teaching method or respond critically to previously published material.

THE VALIDITY OF AN OBJECTIVE EXAMINATION IN CONSTITUTIONAL LAW

PAUL E. SPAYDE,* FRANK R. STRONG,† and W. R. FLESHER ‡

The empirical concept of validity is always present in the question, "Does this examination measure well what it is supposed to measure?" The validity of an examination in Constitutional Law, then, as in any other law subject, depends upon the degree to which it satisfactorily gauges individual student achievement in the field under investigation. Individual student achievement is itself difficult both of definition and of measurement. Capacity in Constitutional Law may be envisioned largely in terms of subject-matter knowledge, or it may be thought of in the entirely different sense of developed skill in synthesizing constitutional principles from the cases and in applying these principles to the resolution of constitutional problems put by newly emerging fact patterns. Admeasurement of achievement, however defined, is no easier. Success in practice, the criterion that so strongly appeals as being the ideal, is, upon closer analysis, seen to be a resultant of such diverse and often irrelevant factors as to raise doubt concerning its assumed adequacy, even aside from the question of feasibility of use. On the other hand, anything approaching a perfect criterion is rarely, if ever, to be found within the law-school experience itself. Among possible criteria of this type, the one most often used is that of the student's composite ratings as revealed in cumulative grade-point averages. Thus, an objective examination in Constitutional Law may be validated by ascertaining the correlation existing between the scores made on it and the cumulative grades of the examinees for all courses taken, or for all courses taken concurrently. The data needed to compute a correlation coefficient are merely the quantitative scores on the two variables, *i.e.*, the cumulative point-hour ratio and the examination score for each student. The experimenter should consult a statistician or a standard statistics textbook for details of computation.

If, as has been traditional, the law essay examination is accepted as a satisfactory measure of student achievement, an objective examination in law may be validated by correlating the results secured on it with criterion

* Research Assistant, Bureau of Educational Research, The Ohio State University.

† Professor, College of Law, The Ohio State University.

‡ Research Associate and Professor, Bureau of Educational Research, The Ohio State University.

scores obtained on an essay examination in the same subject matter. Comparability may be achieved by using on each of the two examinations (1) fact-patterns based upon the same principles of law or involving the same legal skills; (2) fact-patterns involving the same specific content, but with differences in the degree of specificity in the responses; or (3) fact-patterns identical save for the *form* of the response required. A small class in Constitutional Law, first taking four essay problems and then immediately reacting to the same four problems, each translated into *one* multiple-choice objective question, received, on the two types of examination, grades producing a correlation of .49.

With only nine students involved, no conclusion as to the validity of the objective-type examination can of course be drawn from this correlation. A partial explanation of the correlation of .49 may well lie in the fact that the part-credit grading technique, so universal with essay examinations, was used in the scoring of only one of the four items in the objective examination. In as much as the basic postulate—that the law essay examination is a satisfactory criterion—requires acceptance of the seemingly inevitable counterpart of part-credit grading, regardless of its pedagogic soundness, objective examinations will, under this method of validation, consequently require "adjustment" for this factor. Such adjustment can be made either by giving part credit for a second-best election under the first or the third type of comparable examination, or by developing several objective items for each essay item, thus in effect employing the second type described in the preceding paragraph. Even the correlation of .49, existing without adequate adjustment for the part-credit factor and based upon far too few instances to warrant definitive conclusions, was not only positive but high enough to be statistically significant, indicating that the two examinations measured, to a large degree, the same things.

It has been a long-known and commonly accepted theory in test construction that validity is a function not only of the correlation of each item with the accepted criterion, but also of intercorrelation of the various test items. Consequently, when a satisfactory external criterion is unavailable for the validation of an examination, individual test items are often correlated with the total scores of the examination. Of the various methods of item analysis the one most commonly used and recognized is that of internal consistency. Internal consistency should not, however, be confused with what is commonly known as validity. As above indicated, the generally accepted theory of validity is that of concomitance of examination scores with criterion scores; internal consistency is, as the term implies, simply a measure of how well each part or each item of an examination predicts the total score of that examination. It follows from this distinction that the relationship between the total examination scores and the criterion scores must be perfect in order for measures of internal consistency and measures of validity to be equivalent. Such a situation exists only in theory; it is seldom even approached in practice. Nevertheless, determination of the internal consistency of an examination is valuable, not only for providing a quick, easy method of investigating the satisfactoriness of individual test items, but also as an easily understood procedure which should tend, through elimination of "poor" items, to increase the validity of the examination as a whole.

An all-objective examination in Constitutional Law, consisting of twenty items (40 per cent of which were further subdivided) variously weighted for scoring, was subjected to analysis for internal consistency. Because of the unique type of this examination, *i.e.*, the method of scoring the different items making up the total score, item analysis based on criterion groups was believed to be the most valuable. The 200 examination papers were, therefore, divided into five criterion groups of forty each, based upon total marks in the examination.

A method of computing item difficulty from item scores is shown in the accompanying table.

A METHOD OF COMPUTING ITEM DIFFICULTY FROM ITEM SCORES OF CRITERION GROUP ONE.

Code Number of Examination Paper	Item Score								Total Score
	1	2	3	4	.	.	.	20	
1	2	1	2	1	.	.	.	0	17
2	1	1	3	2	.	.	.	0	22
3	1	1	3	1	.	.	.	0	23
4	1	2	1	2	.	.	.	1	23
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
40	1	2	2	1	.	.	.	3	33
Total actual score	65	67	115	56	.	.	.	50	1242[a]
Total possible score	160	120	200	200	.	.	.	280	4000
Mean % of possible score	39	56	58	28	.	.	.	18	32[b]

[a] This figure is a check on the accuracy of column and row totals.

[b] In Criterion Group One 32 per cent of the maximum possible score was attained. This is a measure of difficulty of the examination for this particular criterion group.

Steps in computing item difficulty, based upon data in the table, are as follows:

1. Give each examination paper a code number.
2. Use large sheets of graph paper or other paper appropriate for tabulating data as outlined in the table.
3. List credit given to each item and sub-item in each examination paper in the appropriate space in the table.
4. Compute column totals.
5. Compute the mean per cent of possible score by dividing column totals by maximum possible totals (assuming full credit for all items).
6. Repeat (1) through (5) for each of the other four criterion groups.
7. Graph each item and sub-item, using criterion group on the ordināte and mean per cent of possible scores on the abscissa.

In cases where items are scored as either completely right or completely wrong, the mean per cent of possible score is identical with the per cent of the criterion group passing the item. Where part credit is given for an item, a somewhat different interpretation must be made, *i.e.*, mean per cent of possible score (item difficulty).

Ideally, an item of 40 per cent difficulty (an item failed by 40 per cent of those taking the examination) should be passed by all of those in the up-

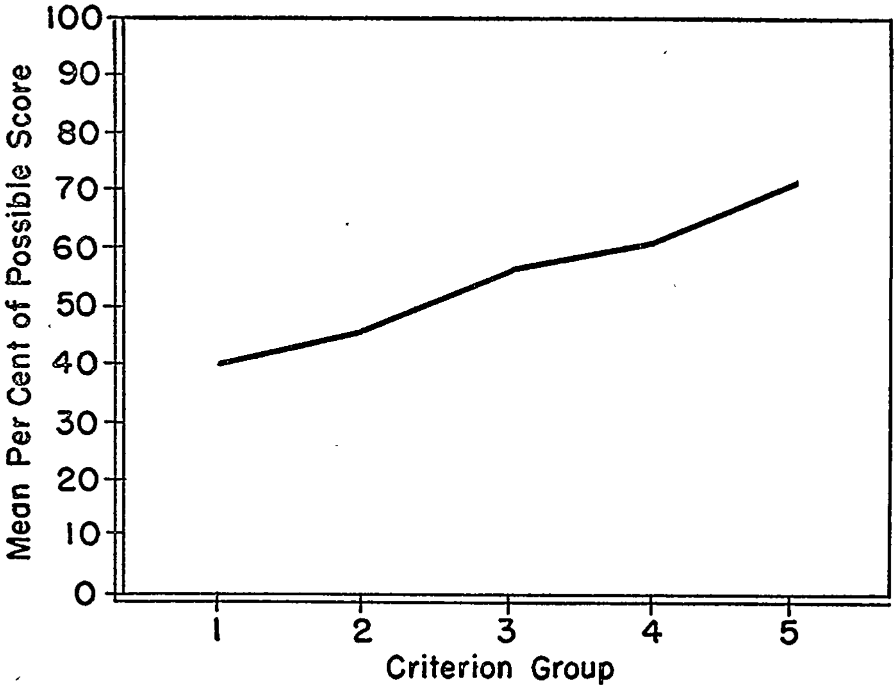

Fig. 1 A SATISFACTORY ITEM

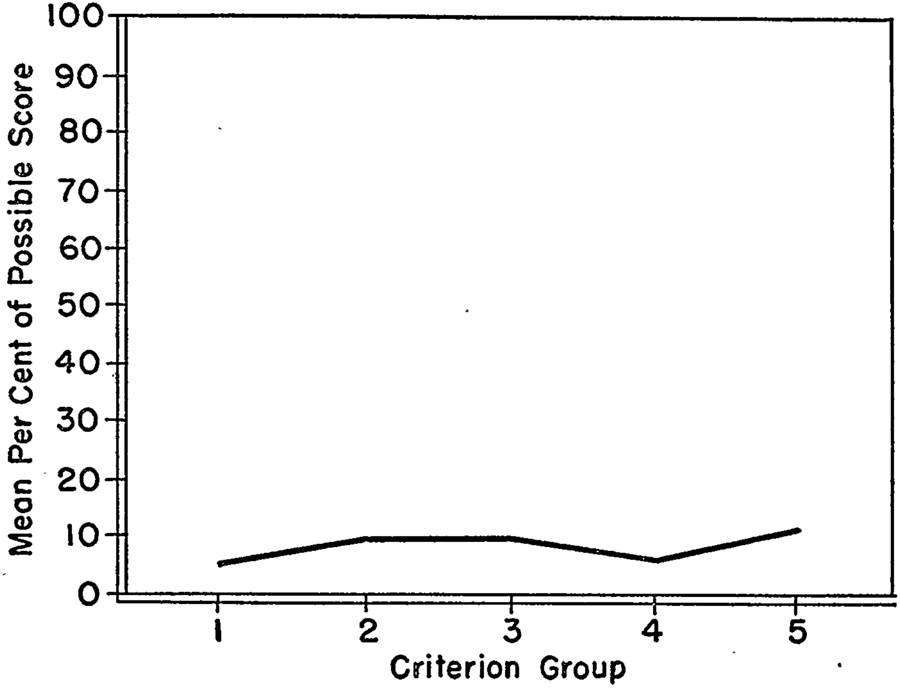

Fig. 2 AN UNSATISFACTORY ITEM

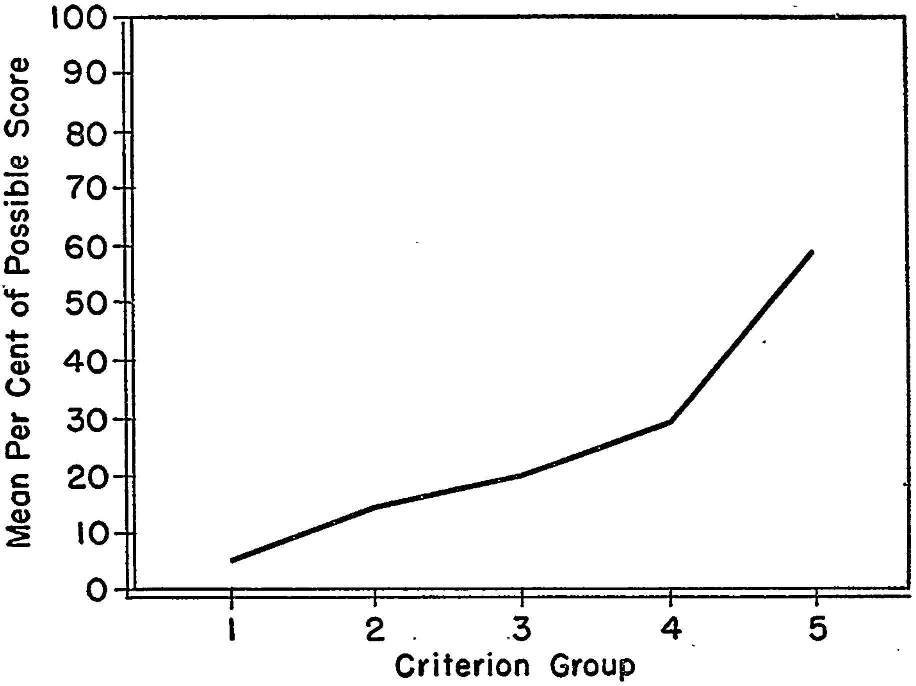

Fig. 3 A VERY GOOD ITEM

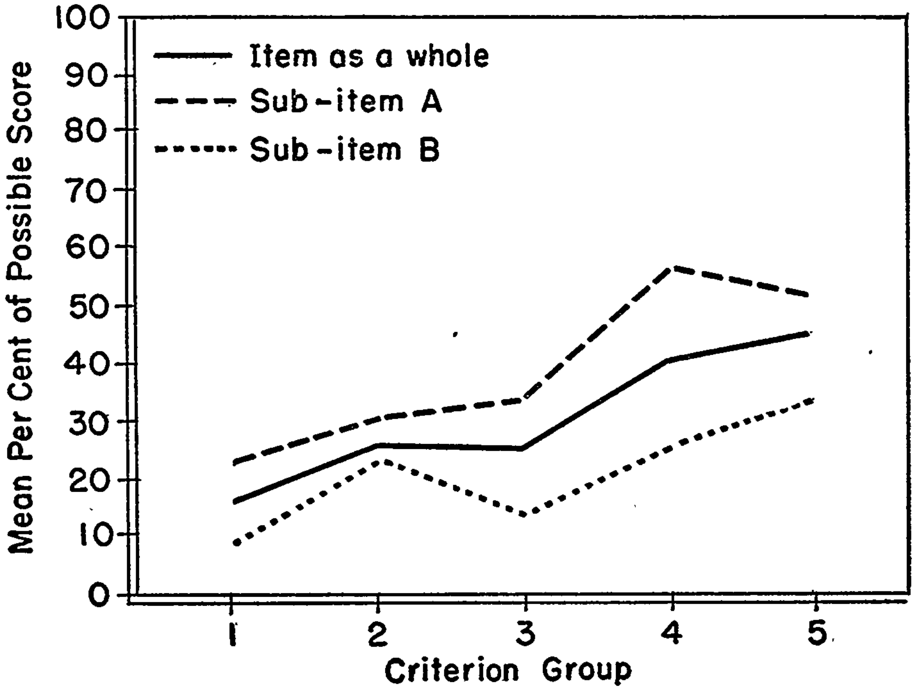

Fig. 4 AN ERRATIC ITEM

per three criterion groups (upper 60 per cent) and failed by Criterion Groups One and Two (lower 40 per cent). Where the number of cases is very large and the amount of credit given to an item can vary greatly, it would be ideally expected, with a perfect criterion, to have a direct relationship between criterion score and item credit. For example, the lowest E student would score zero on the item and the highest A student would get the maximum possible item score.

The difficulty and discriminating ability of a test item are the factors with which one is generally concerned in any method of item analysis. In a multiple-choice item it is important to know whether or not the percentage passing the item is greater than chance factors would allow. It is important also to know whether or not the better students tend to pass an item more frequently than the poorer students. The method of item analysis described in the previous sections and shown graphically in the figures on preceding pages will answer these questions.

The results of a satisfactory item are shown in Figure 1. The average difficulty of the item is slightly over 50 per cent. The item provides increasing discrimination among criterion groups; the groups of students scoring high on the total examination tended also to score high on this item.[1]

The results of an unsatisfactory item are shown in Figure 2. The item is very difficult and fails to discriminate adequately among criterion groups. This item consisted of four alternative answers plus the opportunity to write in a fifth alternative (the correct answer). It is hazardous to speculate on the effect of expecting students to write in a correct alternative when other alternatives are given.

Analysis of a very good item is shown in Figure 3. The graphic analysis shows that the item provides excellent discrimination between the upper two criterion groups as well as good discrimination among the other criterion groups.

Analysis of an erratic item is shown in Figure 4. The item contained two multiple-choice sub-items of five alternatives each. The item as a whole appears to be fairly satisfactory. However, on sub-item A Criterion Group Four did better than did Criterion Group Five, and in sub-item B Criterion Group Two did better than did Criterion Group Three.

It is important to remember that an item analysis such as has been described merely gives a basis for determining the relative merits of the different items of an examination; it does not reveal causal relationships, although some insights into causal factors may be revealed. The first tendency of the test-maker will probably be to eliminate those items which the item analysis indicates are unsatisfactory. To eliminate items, however, may result in discarding from the test certain basic concepts, the loss of which is likely to reduce markedly the over-all validity of the examination. Assuming that the test-maker desires to retain in his examination the concepts represented by the unsatisfactory or very poor test items, he should, in the revision of each such item, exercise care in direct relationship to the degree of the item's unsatisfactoriness as revealed by the item analysis.

[1] The mean slope of the curve in the accompanying figures cannot accurately be interpreted as a coefficient of correlation unless the scales have been "normalized," *i.e.*, scaled according to their respective standard deviations.