DOCUMENT RESUME

ED 076 660 TM 002 667

AUTHOR Gillmore, Gerald M.
TITLE Evaluation by Students for University-Wide Comparative Purposes.
PUB DATE 27 Feb 73
NOTE 7p.; Paper presented at American Educational Research Association Convention (New Orleans, Louisiana, February 25-March 1, 1973)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Course Evaluation; Data Collection; *Evaluation Techniques; *Questionnaires; Reliability; Speeches; *Student Attitudes; *Teacher Evaluation; Validity

ABSTRACT

The use of a short, face valid, objectively scorable questionnaire to obtain students' evaluations of courses as a whole is discussed. The instrument is described from the standpoints of content domain, reliability, face validity, and university-wide applicability. This instrument would provide reliable normed data for use by campus-level administrators. (DB)

Evaluation by Students for University-Wide Comparative Purposes[1]

Gerald H. Gillmore
University of Illinois at Urbana-Champaign

I would like to mention at the outset that I believe the evaluation of instruction is best considered as a highly complex system with inputs from administrators, teachers and students and with outputs to administrators, teachers and students. The purpose of the system is to maximize the objectives of these three groups by means of evaluation, diagnostic feedback and information transmittal. However, I will limit myself to input from one group to one other group for one purpose. Specifically, I will address myself to input from students to administrators for evaluation. Unfortunately, because of time constraints, I will undoubtedly oversimplify even this one facet of the issue.

The Instrument

I would like to suggest to you that data can be collected for this purpose with use of a short, face valid, objectively scorable instrument, possibly containing in the neighborhood of one-half dozen or less broad, general items. In my discussion, I will focus on evaluation of courses as a whole. However, the basic notions are easily expandable to include a general evaluation of instructors and of the content of courses. The brevity of such an instrument would have two immediate advantages. First, we often hear the complaint that evaluation instruments take up valuable class time, especially near the end of

[1] Part of a symposium presentation entitled "A Plan for the Comprehensive Evaluation of College Teaching" at the American Educational Research Association Convention in New Orleans, Louisiana, February 27, 1973.

the semester when the instructor is hurrying to cover material he would have already covered had he planned better. Obviously a brief instrument would expend less time.

Secondly, and probably more importantly, if, in the extreme, every course is evaluated by students every semester, the process of filling out questionnaires quickly becomes repetitious and tedious and, of course, we are at the mercy of the veracity and care with which students respond ~~to such instruments.~~ Although I have no data to back up this claim, I suspect that students would find a short form less objectionable and, therefore, treat it more seriously.

To further pursue the feasibility of an instrument like the one described above, and to further illuminate its features, I would like to discuss it in terms of the content domain, reliability, face validity and university-wide applicability.

The content domain. If we were interested in teacher behaviors, the content domain from which we would have to sample would naturally contain a large number of potential items. Furthermore, the items within that domain would not be inherently related. For example, the items, "The teacher spoke in a loud and clear voice." and "The teacher made the course objectives explicit." may be positively correlated, but there is no inherent reason to expect them to be.

However, if we want a general evaluative instrument to assess how students feel about the global goodness of a course, there is no reason to select other than very broad general evaluative items. As contrasted with teacher behavior, the potential number of general evaluative items is quite small. There are only so many ways to ask how good the course was. Furthermore, these variants would be expected to show strong positive correlations, and our experience has consistently indicated that they do. For example, the average off-diagonal intercorrelations among the eight items which make up the General Course Attitude subscale of Form 66 of the Illinois Course Evaluation Questionnaire (CEQ) is about .85, when correlations are computed over sections. Thus, I feel a relatively small sample of items can adequately generalize to the entire domain, as I have delimited it.

Reliability. Reliability and a small number of items are usually inconsistent notions. Indeed, we frequently advise instructors or researchers to add more items to a test or to a questionnaire to increase its reliability. However, I wish to demonstrate that lack of reliability is not a shortcoming of a brief evaluative instrument by making the extreme claim that each item, taken individually, will have acceptable internal-consistency reliability in most situations.

The reason why we typically use and advise others to use large numbers of items is that each item represents one measurement of something and one measurement by itself is not usually very reliable. However, when we take additional measurements, typically by the use of related items, the reliability of the sum or average increases, essentially following the Spearman-Brown prophesy formula.

In contrast, for a course evaluation item, multiple measurements are already built in because many raters (students) respond to each item. For example, the reliability of a student's attitude score as measured by a twenty-item-scaled-instrument is exactly analogous to the reliability of an instructor's average rating on one item by a class consisting of twenty students.

Let me turn to a specific example. The following Likert-type item appears on the CEQ, "Overall, the course was good." The response categories for this item are: Strongly Agree (SA), Agree (A), Disagree (D) and Strongly Disagree (SD). I chose this partiuular item because it typifies the item type which I feel is appropriate for the short instrument. I could have chosen other items with equivalent results.

I took two samples of 200 sections whose instructor chose to use the CEQ at the University of Illinois (U. of I.) fall semester of 1971-72. My only restriction, beyond the volunteer nature of the example, was that neither the same instructor nor the same course could appear in either sample more than once.

The intraclass correlation for an average individual rater was computed to be .222 for one sample and .246 for the other sample. This is a reliability estimate for the average single rater. If the analogy introduced above is continued, this is analogous to saying that the average off-diagonal correlation among items is .222 and .246 for two samples. Now by applying the Spearman-Brown formula, we can see what <u>item</u> reliability we can expect for classes of various sizes. The reliability as a function of class size is plotted in Figure 1. For the purpose of this example, the intraclass correlation was rounded down to .22. As can be seen in this figure, the curve climb very rapidly so that at a class size of nine, the reliability surpasses .70. At fourteen, it has climbed to .80. Finally, with a class of thirty-one or more, we can expect a reliability at or above .90.

If we expand our instrument to include a few other general items of equivalent reliability and moderate to high interitem correlations, adequate reliability will be achieved for all but the smallest classes.

Face validity. By face validity we mean, of course, do the items look like they are asking what it is we want to know. Thus, the item, "Overall, the course was good." seems to have face validity for ascertaining if the students felt that overall the course was good. However, you might quarrel with the response categories, Strongly Agree to Strongly Disagree. Since it is not always clear what the difference between, say, strongly agree and agree is, perhaps you might prefer the item:

Overall the course was:

A. Excellent
B. Good
C. Fair
D. Poor
E. Atrocious

Again, using data from the two samples of 200 sections mentioned above, the CEQ also contained the item, "Grade the course in comparison with other courses you have had this semester." The response categories were the well-known A, B, C, D and E. The correlations between this item and the "Overall, the course was good." item, computed over sections, were .925 and .911 for the two samples. Thus, we have two objectively scorable items of high face validity, but each using quite different response categories which yield highly similar results. In general, it looks like as long as the response categories are appropriate, ordered, have roughly equal intervals and span the response space, which specific ones are chosen does not matter much.

University-wide applicability. The classical statement, some variant of which anyone working in course evaluation has heard, goes something like the following: No standard instrument can be valid for my course, because my course and the way

I teach it is different from every other course. It seems to me that there is enough truth in that statement to make measurement people appropriately uncomfortable. Furthermore, as an instrument attempts to gain more specific information, the problem becomes more acute. However, I do not see this as a legitimate complaint, in and of itself, against broad general items. Indeed, it would be essentially arguing against a good-bad dimension, which argues against a presupposition fundamental to evaluation.

Conclusion

To conclude, I believe a brief instrument can be devised which will provide administrators with one or a few numbers which are reliable, face valid and have university-wide application and comparability. This number or these numbers will accurately indicate how a set of students generally feel about a given course and/or instructor and/or content. But that is all they will indicate. They will not indicate how much the students have learned, although the two may be correlated. They will not indicate how lenient the instructor graded, although the two may be correlated. They will not indicate how inherently interesting the course was, independent of the instructor's contribution, although again the two may be correlated. Importantly, the number or numbers will not provide any hint to an instructor as to what he is doing well and what he is doing poorly, although they may provide the impetus for him to find out. And they will certainly not tell an administrator if students should evaluate instruction in the first place, although they may cause him to give serious thought to the proposition. Thus, it is essential that an evaluation instrument like the one I have proposed be accompanied by additional inputs, adjunct instructional services and wise and tempered judgment.

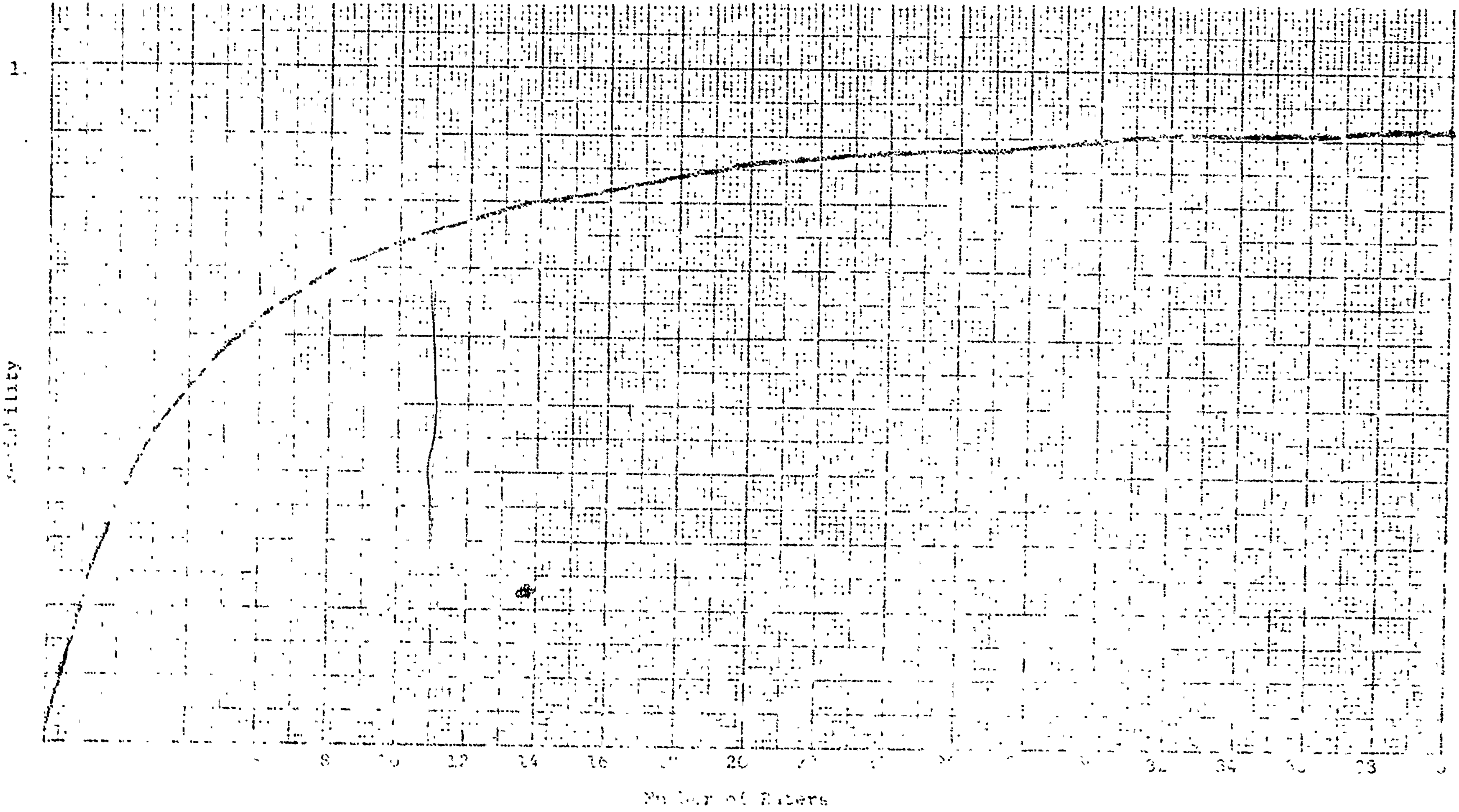

Figure [illegible]. The Reliability of an Item Whose Intraclass Correlation for Individual Raters = .2, as a Function of the Number of Raters