DOCUMENT RESUME

ED 123 391 CE 007 127

AUTHOR Helwig, Carl

TITLE The Five Point Scale on Graduate Students Evaluations of Instructor Performance.

PUB DATE [Apr 76]

NOTE 5p.; Paper presented at the Adult Education Research Conference (Toronto, Ontario, April 7-9, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.

DESCRIPTORS *Graduate Students; *Rating Scales; *Statistical Analysis; Statistical Data; *Statistical Studies; *Student Evaluation of Teacher Performance; Student Opinion; Teacher Evaluation; Teacher Rating

IDENTIFIERS *Five Point Scale

ABSTRACT

The five point scale is the most frequently applied scaling used in the current practices for evaluating instructor classroom performance through graduate student observations. Hence, the investigation addressed itself toward determining, through a series of 55 computerized exact randomization tests, at what degree of mean differences would several graduate student reported classroom means produce statistical significance at alpha .05 on a one-tailed test in either direction. Obviously, the primary intent was to sort out nonrandom from random reported observations so when an instructor compared two means on himself from two classes or a comparison between two instructors and their reported means were compared, such evidence was to be nonrandom rather than random as usually required in behavioral theory and analysis. The results indicated that an instructor would have to have a mean difference between 2.25 and 2.50 to assure himself reasonably that the reported observations on him were nonrandom. The simulated results revealed several severe constraints with the five point scale, making its practical application and interpretation most questionable. Two references are listed. (Author/JR)

THE FIVE POINT SCALE ON GRADUATE STUDENTS EVALUATIONS OF INSTRUCTOR PERFORMANCE

by

Carl Helwig
Old Dominion University
Norfolk, Virginia 23508

ABSTRACT

The five point scale is the most frequently applied scaling used in the current practices for evaluating instructor classroom performance through graduate student observations. Hence, this investigation addressed itself toward determining, through a series of computerized exact randomization tests, at what degree of mean differences would several graduate student reported classroom means produce statistical significance at alpha .05 on a one-tailed test in either direction. Obviously, the primary intent was to sort out nonrandom from random reported observations so when an instructor compared two means on himself from two classes or a comparison between two instructors and their reported means were compared, such evidence was to be nonrandom rather than random as usually required in behavioral theory and analysis.

The simulated results revealed several severe constraints with the five point scale, making its practical application and interpretation most questionable.

Graduate student evaluations of instructor classroom performance seem now to be routine procedures under the current quantification notions of behavioral accountability, including instructor-stated "instructional objectives" with this teaching to be assessed through graduate student observations through some "rating scale." Particularly, student evaluation forms usually contain the rather flabby, non-operationalized item, "rate the overall teaching ability of this instructor" or "considering everything, how do you rate the teaching ability of this instructor" on a five point scale. From these questionable observations, means, standard deviations and other statistics are computed and then surreptitious administrative comparisons between an instructor's own courses as well as between two instructor's courses are accomplished.

To say nothing of these "apples and oranges" comparisons, the proverbial 0-5 point scale itself was subjected in this simulation investigation to statistical mean differences comparisons through the computer-programmed Lohnes and Cooley (1968) exact randomization test. As the authors claimed, this program computed a replacement random sample of 200 points from the possible t-test outcomes of assigning n scores in two groups (samples) in all combinations of n things assigned n/2 at a time.

Randomization tests, according to Siegel (1956), were the most powerful non-parametric techniques whenever measurements were so precise as to give the scores numerical meanings. Did graduate student class means on an instructor's classroom performance have such numerical meaning? And these means, foremost, had to be non-random in their outcomes according to behavioral theory to obtain such numerical meaning. Therefore, in this investigation, the level of statistical significance was set at the proverbial alpha .05. Because the exact randomization test used all the information in its non-randomly selected samples, for two independent samples, the exact randomization test had a power efficiency of 100 per cent (Siegel, 1956). Now, at how much difference would Professor Everyman have to realize with two of his class means at alpha .05 to convince himself that the two obtained means on him were statistically significant, even if one course he taught was in educational statistics and the other in educational history, thus leading to a possible "apple and oranges" comparison nevertheless?

For the computer runs, whose results are reported in the table, an n of eight was selected on the premise that an instructor had a normal teaching load of twelve contact hours with four classes. He thus had four classes one semester and four the next. Each score fed into the computer with the n of eight was greater than zero and less than five, Obviously an infinite number of scores between zero and five were possible, but the data in the table do, it is believed, establish reasonable limits for the comparison by a given instructor of his overall performance for one semester against another and, at the same time, to effect a comparison between two instructors despite the "apples and oranges" limitations. After all, the principal intent in this investigation was to find mean difference limits. Thus for a second insight, what mean difference would be required to assert that Professor Excellent's overall mean was statistically significant at alpha .05 from Professor Poor's mean, despite the fact that one might be in the physics department, while the other is in engineering?

The range of means, as indicated in the table, was from .25 to 4.75 on the five point scale. The total possible number of outcomes for an n of eight (four course means on either side) resulted in:

$$n! \ / \ (n/2!)^2 \ \text{ or } \ 8! \ / \ (4!)^2 \ \text{ or } \ 40{,}320/576 = 70$$

Therefore seventy computer runs were possible. Fifty-five were actually completed for the mean difference 2.25 - 2.50 established the zones between statistical significance at alpha .05 on a one-tailed test in either direction. A few runs shown in the table represent duplication for confirmation as well as a few runs, the interchange of $M_1 - M_2$ for $M_2 - M_1$, to check on direction in the randomization.

The Lohnes and Cooley program produced 200 t-distributions on each run. According to the authors: "A nice thing about 200 outcomes is that .01 times the order number (or rank) of the randomization outcome equal to or closest to (on the small side) the absolute value of the obtained t is the two-tailed probability of the actual outcome of the experiment on the null hypothesis that randomization alone explains the group difference."

Mean differences of 2.0, more often than not, produced non-significant probabilities at alpha .05 on a one-tailed test. Obviously, the size of the standard error of the mean difference in the formula, t=mean difference / standard error of the mean difference, was somewhat controlling. This cited formula is more often known as the expression $t = M_1 - M_2 / \sqrt{s.e._{M_1}^2 + s.e._{M_2}^2}$.

Data on Fifty-five Exact Randomization Tests

M_1	M_2	M.D.	S.E. M.D.	t - test for obtained scores	p (one-tailed test)
1.00	1.75	-.75	1.436	-0.522	.240
4.00	3.00	1.00	.816	1.225	.075
3.50	4.50	-1.00	.408	-2.449	.065
3.25	1.50	1.75	.559	3.130	.040
3.50	2.00	1.50	1.258	1.192	.200
3.50	2.00	1.50	1.323	1.134	.150
4.00	2.00	2.00	.816	2.449	.075
4.00	2.00	2.00	.913	2.191	.035
2.00	4.00	-2.00	1.080	-1.852	.115
2.00	4.00	-2.00	.816	-2.449	.085
2.00	4.00	-2.00	.816	-2.449	.085
3.75	1.75	2.00	.354	5.657	.001
1.75	3.75	-2.00	.354	-5.657	.020
1.50	3.75	-2.25	.901	-2.496	.045
1.75	4.00	-2.25	1.109	-2.092	.065
2.00	4.25	-2.25	.946	-2.377	.090
2.00	4.25	-2.25	.946	-2.377	.055
4.25	2.00	2.25	.854	2.635	.055
4.25	2.00	2.25	1.031	2.183	.065
4.25	2.00	2.25	.854	2.635	.035
4.25	2.00	2.25	.946	2.337	.030
4.25	2.00	2.25	.854	2.635	.010
4.25	2.00	2.25	.854	2.635	.060
4.00	1.75	2.25	.629	3.576	.005
4.00	1.75	2.25	.629	3.576	.035
4.75	2.50	2.25	.382	5.892	.010
4.75	2.50	2.25	.901	2.496	.045
2.50	4.75	-2.25	.901	-2.496	.055
2.25	4.50	-2.25	.382	-5.892	.035
4.75	2.50	2.25	.990	2.274	.075
3.75	1.50	2.25	.382	5.892	.010
2.00	4.25	-2.25	.479	-4.700	.025
2.25	4.50	-2.25	1.407	-1.599	.100
3.75	1.50	2.25	.382	5.892	.010
3.75	1.50	2.25	.382	5.892	.010
4.25	1.75	2.50	.354	7.071	.015
4.25	1.75	2.50	.354	7.071	.020
1.50	4.00	-2.50	.500	-5.000	.010
1.75	4.25	-2.50	.540	-4.629	.025
1.75	4.50	-2.75	.382	-7.201	.015
1.75	4.50	-2.75	.382	-7.201	.010
1.25	4.00	-2.75	.479	-5.745	.010

M_1	M_2	M.D.	S.E. M.D.	t - test for obtained scores	p (one-tailed test)
0.25	3.25	-3.00	.354	-8.485	.010
1.25	4.25	-3.00	.354	-8.485	.035
1.50	4.50	-3.00	.408	-7.348	.010
1.50	4.50	-3.00	.408	-7.348	.010
1.50	4.50	-3.00	.408	-7.348	.020
1.25	4.50	-3.25	.382	-8.510	.035
1.50	4.75	-3.25	.382	-8.510	.035
1.25	4.75	-3.50	.354	-9.899	.010
1.25	4.75	-3.50	.354	-9.899	.035
0.25	4.25	-4.00	.354	-11.314	.035

On the other hand, mean differences of 2.5 on the five point scale produced statistically significant results at alpha .05, with the obtained probabilities being .01 or .02 in either direction on a one-tailed test. As shown in the table, mean differences greater than 2.5 produced statistical significance, while mean differences less than 2.0 did not.

Mean differences of 2.25 seemed to be in the penumbra area, producing probabilities from .02 to .08 on a one-tailed test in either direction. Thus the zones in which Type I and Type II errors were, in general, being produced were also somewhat identified.

What would all the above in part indicate? An instructor would have to realize a mean difference between 2.25 - 2.50 or greater to assure himself reasonably that the reported observations on him were non-random.

Thus, if one class mean were 4.5, the lower mean would have to be 2.25, that is, 4.5 - 2.25 = 2.25. Or with a mean difference of 2.5, if the higher mean were 4.5, the lower mean would have to be 2.0, that is, 4.5 -2.5 = 2.0. The same could be asserted for the more questionable comparison between Professor Excellent and Professor Poor, where, it is held, the 'apples and oranges" comparison would be further magnified because of situational differences, including course content, class size, disciplines, and so on.

At my institution, the reported graduate student data I have seen over a four year period have indicated that graduate students are reluctant to use the higher end as well as the lower end of the five point scale, therefore not many 4.5 to 5.0 nor 0.0 to 2.5 means are produced. As a matter of fact, I have never seen a mean of 3.0 or less. Since under behavioral theory statistical significance must be insisted upon in order to separate random from non-random propositions and/or outcomes, the accountability advocates might take another look at scaling and behavioral numbers game in their efforts to quantify teacher performance through student observational data.

References

Lohnes, P.R. and Cooley, W.W. Introduction to Statistical Procedures with Computer Exercises. New York, John Wiley and Sons, Inc., 1968.

Siegel, S. Nonparametric Statistics for the Behavioral Sciences. New York, McGraw-Hill Book Co., 1956.