# On Budz & Grabar's "Tutorial vs. Classroom" Study

Sarah Warshauer Freedman; Ellen W. Nold

*College English*, Vol. 38, No. 4. (Dec., 1976), pp. 427-429.

Stable URL:
http://links.jstor.org/sici?sici=0010-0994%28197612%2938%3A4%3C427%3AOB%26G%22V%3E2.0.CO%3B2-R

*College English* is currently published by National Council of Teachers of English.

ture. We are talking about freshmen, aren't we, Mr. Harp? Thus the antipathy of which you speak among students for the idea of liberal education may be understated. It is anathema to them, and for many minority students the *progymnasmata* you propose for freshmen writing is a new kind of hell invented for their torment in an institutional inferno already designed to increase their suffering.

FRANK D. ROSS

## *To Frank D. Ross*

Mr. Ross complains that a rhetoric course designed around the Great Books cannot be taught to freshmen; I respond that it has been and is being taught to them.

Let me restate some of the facts given in my essay. The program I described there at the University of Kansas enrolls annually 150 freshmen. It is a twenty-four credit course, requiring four semesters to complete; only six of those credits satisfy any college or university requirements. The students take the course as an elective—often the only elective they are able to fit into their undergraduate curriculum. Therefore, only student demand for the course permits it to continue.

So yes, Mr. Ross, we are talking about freshmen, *average* freshmen. Freshmen who come to the university from farms, inner cities, and suburbs. Freshmen who have sat through tedious fifty-minute discussions about *Zen and the Art of Motorcycle Maintenance* and Ginsberg's *Howl*. Freshmen for whom "Red lips are not so red/As the stained stones kissed by the English dead" strikes the very same contemporary note Virgil struck when describing the death and funeral of Pallas.

Is Mr. Ross's point that life is tragic and that the literature we teach must reflect this? Tragedy is so pervasive in the classics that it is not even necessary to cite examples. One modern commentator remarks that if it were not for Homer's "unwearying, unmoved speech . . . the *Iliad* would be a poem beside which the grimmest modern realism is child's play."

Mr. Ross says the progymnasmata "could become . . . counter-productive." Of course. But it has not. And it will not for anyone who teaches it seriously, delightfully—and with skill. The Great Books participate in the "institutional inferno" in one way only: they condemn the mediocrity and the inertia which reside there. In so doing they not only fan the flames but also light the match. Our students do not need to learn how to croak like frogs; rather, they must know how in this harsh world to draw their breath in pain and still be able to tell their story.

RICHARD L. HARP
*University of Nevada, Las Vegas*

## *On Budz & Grabar's "Tutorial vs. Classroom" Study*

We feel compelled to comment on the design and reporting of the study done by Judith Budz and Terry Grabar ("Tutorial versus Classroom in Freshman English," *CE*, March 1976, pp. 654-656), to make a plea for members of the profession who are interested in research problems to become more sophisticated in research design and to trust their intuition above the results of bad-

ly done "experiments." Not only did Budz and Grabar fail to report salient details of the design and statistics, but those details included revealed a study so flawed as to make the conclusions invalid.

The authors wanted to be able to determine whether the classroom method or the tutorial method was a more effective format for teaching composition. However, since the "classroom" method included an unspecified number of tutorials and the "tutorial" methods an indeterminate number of classroom hours, the distinction between the methods is, at best, tenuous. In their landmark study, *Research in Written Composition* (NCTE, 1963), Braddock, Lloyd-Jones, and Shoer warn against

> . . . the comparison of two methods which affords no descriptive details of the methods, merely a few generalizations which have different meanings for different people. It is essential in such cases that instructional procedures be described clearly (or made implicit in textbooks, exercises, films, machines, etc.) so that other investigators may reproduce at least the general nature of the experiment or replicate it in its entirety. (p. 27)

In order to understand the significance of the study, we need relevant details about the groups being compared: the mean and standard deviation of the number of hours of classroom time for both groups, the number of hours of tutorial time for both groups, the number of papers written by each group, the number of words written by each group, the number of revisions by each group. Besides these pedagogical variables, we need to know whether the mean SAT scores for the students randomly assigned to the two groups were significantly different and (because this point was important to the analysis) whether the *distribution* of the SAT scores was similar for the two groups. Lastly, we need to know how many of the students in each group elected to discuss exactly the same topic in response to the amorphous topic question given in both the pre- and post-tests: "write a well developed essay in which you attempt to convince an intelligent reader that your position on a matter of importance to you is reasonable." It seems plausible to us that students who gave themselves a second try on the same topic would have a better chance of improving from pre- to post-test.

Besides the difficulty of knowing whether there are actually two different methods of teaching and whether the two groups of subjects were indeed equivalently matched and tested, the next largest problem occurs in the reading of the pre- and post-tests. Though pre- and post-tests were read "blind" and together, the main reader was the student's teacher, another reader acting as a check on the teacher. The probability of biased decisions is very high in such a design: the teacher is influenced often by knowing (cued by handwriting and subject matter) the writer's identity, and the teacher can observe surface markers of pre- versus post-tests. For example, the teacher will recognize a student who has learned the correct use of the semicolon from a recent lesson and cannot help but be inclined to award that student a high grade. Because the teacher's opinion was deferred to in the matter of deciding which of two close grades were assigned, there was no real check on the reading. Paul Diederich in *Measuring Growth in English* (NCTE, 1975) gives guides for how schools can evaluate their English programs. The standards he sets are much less rigid than those for formal empirical research; still, he would find Budz and Grabar's reading scheme unacceptable for even his most informal inhouse evaluations. He recommends a scheme modeled in part after freshman English examinations at the University of Chicago, a scheme in which "two samples of each student's writing were

judged independently by four different teachers, selected at random" (p. 20). When grades of two readers differed significantly "readers, who did not know what grades these papers had received . . . would give each paper a third independent reading" (p. 20).

The design was possibly marked with a third serious problem. To negate teacher effects, *each* teacher should have taught a course in *both* formats. Otherwise, we could be witnessing the effects of a slightly more skilled group of teachers who happened to teach the "classroom" method. Budz and Grabar's description of their study gives no clear indication of who taught which method.

Finally, the interpretation of the data left much to be desired. The researchers assume that the mean score change of the two groups is valid as the sole basis for conclusions, but Braddock, Lloyd-Jones, and Shoer caution that

> . . . the progress of a range of students should not be examined only by mean scores, when average gains may be achieved merely by speeding up one end of the range; distributions of scores should be examined. (p. 25)

Besides the problematic focus on mean differences of group scores, the researchers insist on discussing differences that are *not* statistically significant, i.e., that cannot be confidently ascribed to something other than chance. Even the improvement *within* each group was not statistically significant. It makes little sense to hypothesize about reasons for differences that most likely are products of chance.

These researchers began with the intuition that teaching writing tutorially would be superior to teaching in the classroom. They were surprised by what they thought their numbers told them. On the basis of their statistical findings, they decided to give up the tutorial method, a decision that went against their basic intuitions. We hope to convince them and other readers that their intuition is probably a great deal more reliable than their conclusions from their unreliable study. They would be wiser to expend their energy on improving a teaching format that they tried for only one year, a format for which their students, not insignificantly, were "unanimously favorable" in their evaluations.

SARAH WARSHAUER FREEDMAN
ELLEN W. NOLD
Stanford University

---

## *Response to Freedman & Nold*

We thank Ellen W. Nold and Sarah Warshauer Freedman for their comments on our experiment in teaching composition. Any additional insight into techniques for scientific inquiry in the humanities is always helpful, and it is also amusing to be warned against "unreliable" experiments while being told to "trust [our] intuition."

Nold and Freedman ask about the distinctions between the two groups. Because the sampling was done correctly, the difference between the tutorial and classroom groups was clear and any apparent confounding is negligible. Both groups of students wrote the same number of papers. The number of students in the classroom group who had additional conferences with their teachers was minuscule, since, as most experienced composition instructors know, it is often difficult to lure students to individual sessions.

As the response points out, the probability of biased decisions in a design such as ours might be high. We controlled for this possibility by using our