Hoye contends that we are arguing for mainstreaming. We are not. We are rather raising three central questions for the profession to address, consider, and debate. No syllogisms of the sort Hoye puts forth underlie our article. Although social equality may well contribute to language learning, we are not so naive as to think that mainstreaming, in and of itself, will result in social equality. Social equality will be achieved only when all individuals share equal political and economic access. The school, along with other social institutions, plays a role in either supporting or undermining social equality. When social equality is attained, individuals and social groups will be able to select the language programs that best meet their needs.

We envision our article being used in the following manner. A community would set the priorities they believe important based on the questions we raised and others. If a community were to decide that its priorities included native language development, academic achievement, and social integration, then program development would proceed with these priorities in mind. Setting priorities would encourage planners to design innovative programs responsive to community needs, rather than automatically selecting a preexisting model. Our article advocated no approach. We argued rather for a careful weighing of priorities before designing or accepting any approach to minority education.

Research Issues

The *TESOL Quarterly* publishes brief commentaries on aspects of qualitative and quantitative research. For this issue, we asked two researchers to address the following question: What is the importance of power and effect size for second language research?

Edited by GRAHAM CROOKES
University of Hawaii at Manoa

Power, Effect Size, and Second Language Research

A Researcher Comments . . .

ANNE LAZARATON
The Pennsylvania State University

When we engage in testing hypotheses in our research, we hope we will be able to reject our null hypotheses (e.g., that there is no difference

between two groups) and accept our research hypotheses (e.g., that there is a difference). This decision is based on a test of significance, where some observed statistic is compared to a critical value at a prespecified level of probability. Because of the nature of significance testing, however, some of these decisions to reject the null hypothesis will be incorrect. We may incorrectly reject a true null hypothesis (H_0)—claim that an effect or relationship exists when it does not—and commit a Type I error. The likelihood of this is equal to our prespecified level of probability, alpha (typically .05). A Type II error, beta, occurs when we claim there is no effect or relationship when there is, and retain the null hypothesis when it should be rejected. Alpha and beta are inversely related: When we choose a more conservative alpha value, such as .001, to minimize the likelihood of a Type I error, at the same time we increase the chance of committing a Type II error by retaining a false H_0. The reverse is also true. As researchers, we strive for the optimal balance between the possibility of committing either type of error. The statistical test which best achieves this balance is the most powerful test we can choose.

Power, the ability of a statistical test to detect a false null hypothesis, is therefore highly desirable, because by minimizing the likelihood of making an error in evaluating a null hypothesis, we increase confidence in our findings. How do we know how powerful a test is? While it is possible to estimate power in a particular study (Cohen, 1988; Shavelson, 1988), for those of us who engage in small-scale research or who are consumers of research, it is more realistic to be aware of factors which affect power and ways we can increase the power of a statistical test. Power is a function of four factors: significance level, variability within the population, sample size, and effect size.

Conventionally, the significance level for testing the null hypothesis is often .05. But in attempting to control for Type I error by selecting a low alpha level, we increase the likelihood of Type II error. We might, therefore, consider raising the .05 level to .10 to ensure that we will be able to detect a false null hypothesis. We should also think about the trade-off between Type I and II errors, and their consequences. It may be that the conservative .01 level is appropriate if a serious decision rests on the outcome; however, in a pilot study of new materials, .10 may be reasonable. As researchers, we must choose significance level carefully. It should be selected before the study is conducted, and results reported relevant to it. This highlights the common misconception that .01 or .001 values (and the double and triple asterisks that often accompany them) indicate importance: Lower values mean only that we can be more certain that effects or relationships exist, not that they are important. For example, a statistically detectable difference of only a fraction of a point between two groups may have no practical implications for teaching.

A second factor affecting power is the variability of the population from which the sample is drawn and to which conclusions will be generalized. The less variability in the sample, the more powerful our test, because any true effect or relationship will be more easily detected when it is less obscured by random differences. One way that the effect of variability

can be decreased is to increase sample size, a third facet of power. As more observations are made, the less variability there is in our statistical summaries because these now reflect a larger body of combined information. Also, as more observations are made, the closer our sample distribution will approach the (assumed) normal distribution of the population itself. (That it is normal is an assumption which underlies many powerful statistical procedures.) The larger the sample, the more powerful our test, and it goes without saying that we should use a large sample whenever possible. If true differences exist, they are more likely to be detected in a sample of 100, rather than one of 20. It is true, though, that we often work with data from "general learners," which tend to represent a heterogeneous population. This often leads to large within-group variance, and thus to nonsignificant differences. Consequently, even larger sample sizes are needed in these cases.

Finally, effect size is a critical component of power, although it is one of the least familiar concepts of statistical inference (Cohen, 1988, p. 10). Effect size refers to the magnitude of the difference between the population means under the null hypothesis (H_0) and the research (H_1) hypothesis. The hypothesized difference in means must be expressed exactly in the research hypothesis, not just as greater or less than the mean under the null hypothesis. The magnitude of the difference is often expressed as a z score, a "small" difference being .2 of a z score, a "medium" difference .5, and a "large" difference, .8 (Cohen, 1969, cited in Shavelson, 1988, p. 295). Unfortunately, even in well-designed experimental studies, an exact alternative hypothesis—necessary for the calculation of effect size—cannot be stated (Henkel, 1976). Though we may be unable to specify effect size in advance, we can estimate and report it after the fact, with an appropriate strength of association measure. Eta-squared, omega-squared and phi give an indication of the importance of obtained results in terms of strength of treatment effect or relationship. These measures tell us how much variability in the dependent variable can be accounted for by the independent variable, or how much information variables share in a given sample. While the measures cannot speak to strength of association in the population, they provide vital information about a study that is unavailable when just alpha levels are reported (see Hatch & Lazaraton, 1991, for more on these measures).

To summarize, power is a function of significance level, variability in the population, sample size, and effect size. Decisions about significance level and sample size, like those about hypothesis formulation, data collection, and data analysis, cannot be avoided by the researcher. It is tempting to "let the computer (or the consultant) decide," but ultimately the responsibility for our work rests with each of us, and the integrity with which we carry out this work *is* judged by our fellow applied linguists if not the larger educational community.

ACKNOWLEDGMENTS

I wish to thank Patricia Dunkel and Evelyn Hatch for insightful comments on this article.

THE AUTHOR

Anne Lazaraton is Assistant Professor of Speech Communication at The Pennsylvania State University. She publishes and does research in conversation analysis, oral proficiency testing, and research methodology in applied linguistics.

REFERENCES

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. New York: Newbury House.

Henkel, R. E. (1976). *Tests of significance*. (Quantitative Applications in the Social Sciences Series.) Beverly Hills, CA: Sage.

Shavelson, R. J. (1988). Statistical reasoning for the behavioral sciences (2nd ed.). Boston, MA: Allyn & Bacon.

Another Researcher Comments . . .

GRAHAM CROOKES
University of Hawaii at Manoa

The most obvious propositions which come to mind when discussing power and its companion, effect size, in a second language (SL) research context are: (a) the use or provision of power and effect size estimates is extremely desirable, and (b) they are almost never used. This may lead us to asking why there is this disuse, and what can be done about it.

Because power and effect size are unusual in SL research (the former was ranked 22 out of 23 in terms of researchers' "self-knowledge" [Lazaraton, Riggenbach, & Ediger, 1987]), I begin with a simple (and partial) definition. In a simple two-group experiment, effect size is the extent to which the mean of the experimental group differs from that of the control group, standardized in terms of the standard deviation of the two groups combined. (Several other measures exist—Murray & Dosser, 1987; Rosenthal & Rosnow, 1984—and the concept is applicable outside an experimental context, but consider this, for simplicity.)

Why is the use of this statistic desirable? First, it provides a measure which indicates whether the result of an experimental treatment was substantial or not, regardless of sample size. Second, if the likely effect size of an investigation can be estimated before the study is undertaken (via a pilot or previous inquiry) it can be used, in a "power study," to ascertain how many subjects will be needed to have a specific probability of rejecting the null hypothesis. In other words, an existing effect size estimate can tell the investigator how large a sample will be needed to obtain statistical significance at a given level of power.

Power studies are essential if we are to keep to a minimum the two main errors that can be made in statistical inference testing (Types I and II;

cf. P. Cohen, 1983): rejecting the null hypothesis when in the whole population a difference does not exist; and accepting the null hypothesis on the basis of the results obtained from the groups sampled when in the whole population a difference does exist. Most of the time, investigators use statistical testing conditions which are quite unlikely to register (by showing a statistically significant difference) an effect which actually exists and have no idea of their chances of obtaining statistically significant differences. Nevertheless, there is a persistent tendency to rely on the conventions (a) "30 subjects is sufficient," and (b) control of Type I error at 1 in 20, with no control of Type II error.

Why, one may ask, is this the case, whose fault is it, and why has nothing been done? One reason for this situation is the institutional inertia in the educational research community as a whole, and particularly in the SL research field (cf. Lazaraton et al., 1987), where until recently many senior investigators had training in linguistics rather than educational research. The situation is exacerbated by the exclusion of recent developments in statistics from statistical research texts, which present statistical inference as subject neither to debate nor change (cf. Gigerenzer & Murray, 1987; Oakes, 1986). This is indicated in the lack of reference to original works in many such texts (exemplary exceptions: Glass & Hopkins, 1984; Keppel, 1991). The one person whose fault it is not is Jacob Cohen, who with various associates has been providing ringing calls to redress this situation for several decades now, on whose work I am drawing extensively here (e.g., Cohen, 1962, 1965, 1977, 1990).

There is no good reason why nothing has been done, but the following weak excuses may be noted. (a) The SL research community cannot unaided overcome the neglect of the topic within standard statistical courses and texts. (b) Until recently, power analyses required the use of fairly complex formulae and tables (J. Cohen, 1977; Kraemer & Thiemann, 1987; but see Lipset, 1990). (c) Since so much SL research is groundbreaking, there are almost never the preliminary estimates of effect size necessary for power analyses. (d) SL research is labor-intensive in nature—no undergraduate classes of Psychology 101 with their required participation for grade credit and 200-strong *n* sizes are available to us! (e) Power and effect size programs are not to be found in the standard mainframe statistical packages.

Much of this constitutes the unfortunate history of this topic, but little can serve as justification. A simple computer program (Borenstein & Cohen, 1988; cf. Borenstein, Cohen, Rothstein, Pollack, & Kane, 1990) can handle the necessary calculations. There is a growing literature which documents the weaknesses of underpowered studies (e.g., Frieman, Chalmers, Smith, & Kuebler, 1978; Lynch, 1987; cf. Cohen & Hyman, 1979; Sedlmeier & Gigerenzer, 1989). SL researchers are increasingly able to handle original sources in quantitative methods. We now recognize that many published SL studies are no more than pilot studies, which would have been greatly improved if seen as such and followed up by doing the actual study, properly and with a decent *n* size. But in addition, in some

areas, accumulations of studies do provide the possibility of preliminary work which can be used in power studies, so that more definitive results can be obtained. Having said this, I would have to acknowledge from my experience that the labor-intensive demands of applied linguistics research still carry practical, if not logical, weight.

Nevertheless, it is essential that those of us doing quantitative work which involves the use of statistical inference commit ourselves to not publishing pieces unless they can make a substantive contribution to the field, as opposed to our resumes—we must resist the temptation of the MPU (minimal publishable unit). Just as journal editors and article readers have conventionalized (for good or bad) the .05 alpha level, it will be up to them to seek more substantive indications from those who wish to be published in the future that the results they provide are not just a chance effect of the law of small numbers. Power studies and effect size measures can help to provide this substantiation.

THE AUTHOR

Graham Crookes is Assistant Professor, Department of ESL, University of Hawaii at Manoa.

REFERENCES

Borenstein, M., & Cohen, J. (1988). *Statistical power analysis: A computer program.* Hillsdale, NJ: Lawrence Erlbaum.

Borenstein, M., Cohen, J., Rothstein, H., Pollack, S., & Kane, J. (1990). Statistical power analysis for one-way ANOVA: A computer program. *Behavior Research Methods, Instruments and Computers, 22,* 271-282.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal Social Psychology, 65,* 145-153.

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95-121). New York: McGraw-Hill.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*(12), 1304-1312.

Cohen, P. (1983). To be or not to be: Control and balancing of Type I and Type II errors. *Evaluation and Program Planning, 5,* 247-253.

Cohen, S. A., & Hyman, J. S. (1979). How come so many hypotheses in educational research are supported? (A modest proposal). *Educational Researcher, 8*(12), 12-16.

Frieman, J. A., Chalmers, T. C., Smith, H., & Kuebler, R. R. (1978). The importance of beta, the Type II error and sample size in the design and interpretation of the randomized control trial. *New England Journal of Medicine, 229,* 690-694.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics.* Hillsdale, NJ: Lawrence Erlbaum.

Glass, G. V., & Hopkins, K. D. (1984). *Statistical methods in education and psychology* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Keppel, C. (1991). *Design and analysis* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Kraemer, H. C., & Thiemann, S. (1987). *How many subjects? Statistical power analysis in research*. Newbury Park, CA: Sage.

Lazaraton, A., Riggenbach, H., & Ediger, A. (1987). Forming a discipline: Applied linguists' literacy in research methodology and statistics. *TESOL Quarterly, 21*(2), 263-277.

Lipset, M. W. (1990). *Design sensitivity: Statistical power for experimental researchers*. Newbury Park, CA: Sage.

Lynch, K. B. (1987). The size of educational effects: An analysis of programs reviewed by the joint dissemination review panel. *Educational Evaluation and Policy Analysis, 9*(1), 55-61.

Murray, L. W., & Dosser, D. A. (1987). How significant is a significant difference? Problems with the measurement of magnitude of effect. *Journal of Counseling Psychology, 34*(1), 68-72.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.

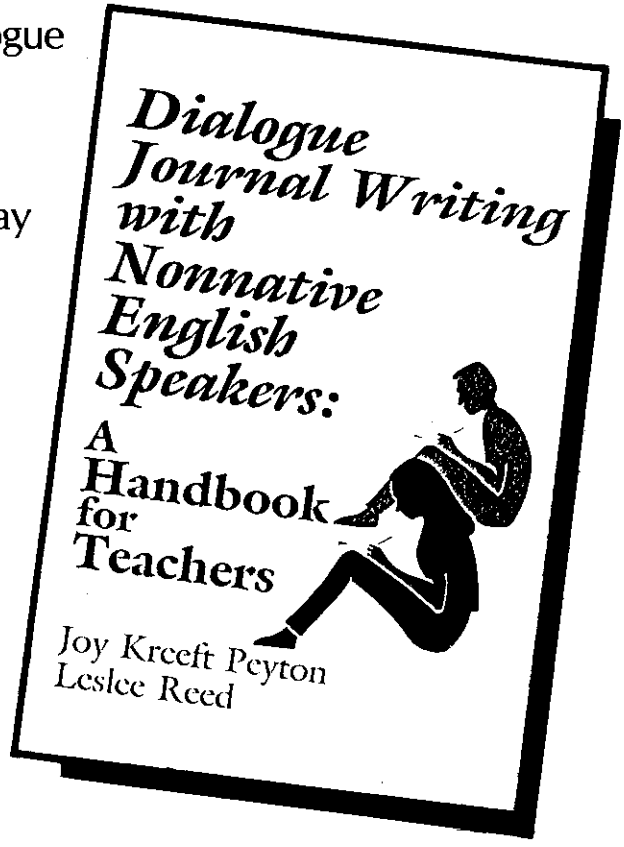
Rosenthal, R., & Rosnow, R. L. (1984). Considerations of power. In *Essentials of behavioral research* (pp. 355-365). New York: McGraw-Hill.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309-316.

Prepare for class with books from TESOL

Even if you have no experience with dialogue journal writing, Joy Kreeft Peyton and Leslee Reed will get you started. Between them, they have 30 years of firsthand experience with this activity. They explain how to:

- effectively start dialogue journal writing with students
- maintain the dialogue once begun
- deal with typical problems that may arise



This handbook offers practical advice in an easy-to-use format free of jargon. It is recommended for teachers of nonnative English-speaking students in mainstream, bilingual, or ESL programs, from kindergarten through high school. It also has direct application to native English-speaking, gifted and talented, learning disabled, and special education students.