Shiken: JALT Testing & Evaluation SIG Newsletter. 10 (2) December 2006 (p. 21 – 24)

Statistics Corner: Questions and answers about language testing statistics

Generalizability from second language research samples

by James Dean Brown (University of Hawai'i at Manoa)

Question: It seems one of the most common mistakes by novice researchers is making statements about a large population on the basis of a small sample. Is there any rigorous way to compute how a classroom research sample of 40 might actually be able to generalize to a Japanese undergraduate university population of 2,809,000? What precautions should novice researchers take when attempting to generalize their studies to larger populations?

Answer: You seem to be asking several different questions simultaneously: one about sampling and *generalizability*, and the others about sample size and statistical *precision* and *power*. I will deal with the first (generalizability) question in this column and the others (precision & power) in subsequent issues of *Shiken*.

Generalizability is usually defined as the degree to which the results of a study based on a sample can be said to represent the results that would be obtained from the entire population from which the sample was drawn. In other words, generalizability depends on the degree to which the particular sample in question can be said to be representative of the population. To explain that, I will first have to define and discuss (a) samples and populations, (b) random samples, (c) stratified samples, and then come back to the issue of (d) what constitutes adequate generalizability to a population.

Samples and Populations

A *population* is the entire group of people that a particular study is interested in. For example, political polls in the US are often focused on the entire voting population of the country. Such polls are typically based on samples taken from that population, but the polls would not be very useful if inferences about the entire voting population were not possible. In second language studies, we are often interested in the population of all ESL students in the US, or all university-level EFL students in Japan, or some such population. However, few language researchers have the resources to study these entire populations. So researchers use *samples*, that is, subgroups of the students are drawn from the population to represent the whole population. Researchers use samples for a variety of reasons: usually some combination of making the data collection and analysis cheaper, more practical, efficient, and/or effective. Properly sampled data should represent what would result if data for the entire population were used. In other words, the results of the study should be representative of results that would occur if the researcher were able to investigate the entire population. A number of strategies are used to accomplish this *representativeness*, but the two most common are called random samples and stratified random samples.

Random Samples

Random samples are created by making sure that each person in the population has an equal chance of being selected into the sample. This can be achieved by clearly defining the population that is the focus of the study, listing all the members of that population, assigning a separate ID number for each member of the population, and randomly selecting the members of the sample on the basis of random numbers generated in a spreadsheet or taken from random numbers tables (which are found in the back of many statistics books). By using random numbers to decide who should be in the sample, the selection is made dispassionately, thereby minimizing any conscious or unconscious biases in the results of the study. It is assumed that a sample made up of a large number of randomly selected people drawn from the population (i.e., a *random sample*) will probably represent the population from which it was drawn. Researchers widely accept this assumption that random samples are representative (for more information on random sampling, see Brown, 1988, pp. 111-113; Brown, 2001, pp. 72-74, Thompson, 2006, pp. 12-13).

Stratified Samples

Stratified samples are created in a slightly different way: by clearly defining the population that is the focus of the study, identifying *strata* (i.e., salient characteristics of the population and/or particular characteristics of interest to the researcher), selecting members from each of the strata in the population (perhaps using random numbers as described above) so that the resulting sample has about the same proportions of each characteristic as the whole population. For instance, in the population of students studying English at Tokyo University, a researcher might choose to use strata within the population for: gender (male, or female), home prefecture, academic status (graduate or undergraduate), and type of major (e.g., science and humanities). With correct information about how many students fall into each of these categories, the stratified sample would be created by selecting students from each of the strata in proportion to those same strata in the total population.

As I point out in Brown (2001, p. 73), three factors influence whether random or stratified sampling procedure is more appropriate.

1. If the population is comparatively heterogeneous, a stratified sampling strategy may be more appropriate, since a random sample might not supply an adequate variety of people from each stratum.
2. If the sample will be small or the groups formed within the study will have unequal sizes, a stratified sampling may be more appropriate.
3. If the researcher wants to use the strata as variables in the study, stratified random sampling may be more appropriate.

However, if the opposite is true, that is the sample is relatively homogeneous and large with equal sized groups, and the strata are not important to the analysis, a random sample may prove better because it removes the need to define the strata in the population and sample proportionately from each stratum.

What Constitutes Adequate Generalizability?

In most of the second language studies I have read, the sampling has been neither random nor stratified. The samples have instead typically been "samples of convenience", that is, made up of "the students at our university who took the placement examination", or students in "my class and my friend's class". This can be problematic when language professionals think of such "samples of convenience" as representing some larger population.

There are two problems with this way of thinking. The first problem is that the "some bigger population" cannot be defined. What is the population of ESL students in the US or university EFL students in Japan? In the first case, there are literally hundreds of nationalities and language background students studying ESL in the US. At each language center in the US, I'm willing to bet that there is a different balance of language backgrounds (not to mention differences in the balance for genders, educational backgrounds, IQs, socio-economic status, etc.). In order to sample from such a population we would need to define it. We could do so by tracking down and listing every ESL student in the US. Unfortunately, even if we could find them all, by the time we had our list, the population would have changed. Thus defining such a population is nearly impossible. Similar problems would arise in trying to define the population of all EFL students in Japanese universities (even if we could get permission to do such a study).

The second problem with samples of convenience is that often they should not be thought of as samples at all, but rather should be viewed as intact populations, perhaps "populations of convenience". That is to say, since these groups of students are so poorly selected that they cannot be said to represent anything larger than themselves, they should be considered populations of convenience that represent themselves.

The mistake that researchers make is in generalizing from such populations of convenience to larger populations. For instance, a researcher might do a study of 40 undergraduate students at Tokyo University and then try to generalize the findings to "a Japanese undergraduate university population of 2,809,000" (as you put it). The problem, of course, is that students from Tokyo University (or those from any other university) can in no way be said to serve as a sample that represents the population of all Japanese university students. Another team of researchers might do a large scale study of the trends in EFL scores for high school students in Ibaraki prefecture and then make the mistake of generalizing from that population to the larger population of all high school students in Japan. The problem of course is that even using the population of all the high school students in Ibaraki prefecture does not justify saying that they constitute a random, stratified, or any other sort of representative sample of all the high school students in Japan. It would be much better to simply discuss such a study in terms of it representing the population of all high school students in Ibaraki prefecture (during such-and-such a period of time).

Given the fact that many of the sets of students used in second language studies are samples of convenience and the fact that samples of convenience themselves are typically the populations beyond which it is irresponsible to generalize, perhaps we should be less concerned with generalizability and more concerned with the transferability of the results of a study (as suggested by the discussion in Lazaraton, 1995, p. 465). *Transferability* is the demonstration of the "applicability of the results of a study in one setting to another context, or

other contexts" (Brown, 2001, p. 226). In other words, given that we very often cannot generalize our results beyond the population of convenience, perhaps we should abandon the notion of generalizability and, instead, describe the groups of students in these populations of convenience *thickly* (i.e., in considerable detail) so other researchers and the readers of our studies can decide for themselves if the results are *transferable* to the settings that they are dealing with.

References

Brown, J. D. (1988). *Understanding research in second language learning*. Cambridge: Cambridge University.

Brown, J. D. (2001). *Using surveys in language programs*. Cambridge: Cambridge University.

Lazaraton, A. (1995). Qualitative research in applied linguistics: A progress report. *TESOL Quarterly*, *29*, 455-472.

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. New York: Guilford.