Some Troublesome Points in Quantitative Analysis of Survey Results

H. P. L. Molloy, Asia University

There are many excellent guides to choosing statistics for analyzing survey results. The books, articles, and web sites referred to below are good places to start, and I refer the readers to these for specific details on how to choose analyses and how to do the actual analysis. This short summary serves simply to collect some of the points that frequently cause difficulties in survey analysis. It was born of my efforts to analyze my own surveys and to understand the analyses of others. The points I make simply strike me as ones that have been underemphasized hitherto.

STATISTICS AND THE SURVEY USER

A functional knowledge of statistics is a matter of fundamental literacy for anyone who uses research in applied linguistics. If you do research of any sort using surveys or base any decisions on surveys you read, then knowing what the statistics mean is essential. Statistical literacy is just as important for the qualitative researcher as it is for the quantitative researcher: the choice to concentrate on qualitative research can only be justified when you know the limitations and assumptions of more widely used research approaches.

Fortunately, basic statistical concepts are very simple. Most commonly used statistics require mathematical sophistication at only the junior-high school algebra level. Like any discipline, statistics can be complicated and (at times) contentious, but almost any introductory statistics text can familiarize you with at least the most common descriptive statistics and the ideas behind them. In applied linguistics, the many writings of J. D. Brown (*e.g.,* 1991, 1992, 1996, 2001, 2005) provide clear explanations of various statistical procedures and, in the longer works, examples that can be followed easily with a spreadsheet. The now out-of-print guide by Hatch and Lazaraton (1991) has a good selection of examples from published applied linguistics studies. Unfortunately, the text is marred by numerous mistakes. On the other hand, if you catch the mistakes, you'll know you are on your way to understanding. The Hatch and Lazaraton text is one of the few to cover both major families of statistical tests (see below). The second edition of the text by Siegel and Castellan, Jr. (1988) is the standard guide for small-scale analyses. The Internet offers a seemingly endless number of resources. A good place to start looking is in the excellent StatPages web site (Pezullo, 2004), which contains links both to downloadable texts and interactive statistics study pages and to sites offering free statistical programs. For me, indispensable among the latter is the free Excel add-in PopTools (Hood, 2003), which comes with many examples and fixes some flaws in Excel.

I mention that a "functional" knowledge of statistics is essential. Even if you specialize in statistics, you will not learn everything about the topic. The solution is to ask someone who does. Luckily, Japanese universities have an abundance of statistical experts: it always pays, even if only with a sense of confidence, to check with someone who knows more than you or might have a different viewpoint than you do. The Internet offers resources that are easy to exploit: the statistics department at UCLA, for example, offers a free statistical consulting service on the Web (http://www.stat.ucla.edu/).

Applied linguistics is a young discipline, and many of its analytic techniques were borrowed from other areas, such as psychology, sociology, linguistics, and anthropology. Many of the statistical techniques used in applied linguistics derive ultimately from work in probability theory, derived from studies of gambling, and from agricultural studies. The assumptions of many techniques, then, may not be appropriate for work in applied linguistics. The youth of our field also means that readers cannot be sure the statistics in published works have been interpreted correctly or have met the assumptions of the particular statistic used.

Besides violation of assumptions, two common problems seem to obtain in many applied linguistics studies: the interpretation of statistical significance, and multiple comparisons.

CHOOSING A *P* LEVEL

The *p* level is the cut-off point for determining what numbers in your data reflect a real phenomenon in the world. The *p* level is also referred to as an *alpha* (α) level and as "sig" (in the computer program SPSS). A value from your data that is smaller than your chosen *p* level is considered to be reliably different from random noise. Such a difference is known as "statistically significant."

Both words in the phrase "statistically significant" are in quotation marks because "statistically significant" does not at all mean the same thing as the everyday word "significant." Further explanations of this point are available in Rozeboom (1960), Cartmill (1980), or McCloskey (1985). "Statistically significant" does not mean "important" or "influential." It only means unlikely to have been generated by random noise.

How unlikely? The *p* value result given by statistical tests shows how unlikely. A *p* value of 0.16, for example, shows that there is a 16 of 100 (or 16%) chance that the set of numbers you have could have been generated simply by using random numbers. Before you begin collecting data in a study, you should choose a *p* level. The simplest way to view your choice of a *p* level is to consider it a kind of bet: what chance are you willing to take that any information you find in your results doesn't reflect what is really there? If a 1 in 20 chance is acceptable, then choose the most common $p = 0.05$ level; for a 1 in 100 chance, choose $p = 0.01$.

If you have chosen a *p* value of 0.05 and your results indicate that your numbers give a value of 0.049 (*i.e.,* less than 0.05), then go ahead and say that your data probably reflect some phenomenon in the real world.

If you have chosen a *p* value of 0.05 and your results indicate that your numbers give a value of 0.051 (*i.e.,* more than 0.05), then you cannot say your data probably reflect some phenomenon of the real world. This is one of the most difficult points to deal with in using statistics: the question of statistical significance is binary. Either your results are lower than or higher than your chosen *p* level, and that is all. If you have chosen a *p* level of 0.05 and the results of your test are 0.75, your result is not statistically significant. If you have chosen a *p* level of 0.05 and the results of your test are 0.05001, your result is also not statistically significant. The value of 0.05001 is not "almost significant," "approaching significance," or showing some kind of "trend": it is simply not significant. In this, statistical significance resembles betting on a horse to win: it doesn't matter if your horse meanders in three minutes after the winner or is beaten by a nose. You lose either way.

Another important note is that the *p* level you choose applies to the entire set of numbers you are studying, which brings us to the next common danger in using statistics for research.

THE PROBLEM WITH MULTIPLE COMPARISONS

If you gamble on one number on one spin of a roulette wheel, you have a one in fifty chance of winning. You probably won't win. If you keep playing the same number over and over, however, eventually you will win. Betting many times in one session increases your chances of winning in that session, even though the odds of winning stay the same for any single play.

An analogous situation obtains with using statistics. If you set a *p* level of 0.05 and run a statistical check on your data, you have a one-in-twenty chance of "losing," of finding a pattern that may not reflect one in the real world. If you run the same statistical check on your data again, you are decreasing your chances of "losing" and increasing your chances of "winning." In other words, you increase the likelihood that it will look like you've found something when you actually have not. A more thorough explanation of this point is available in Siegel and Castellan, Jr. (1988, pp. 168-169).

Should this happen, you will end up misleading yourself and your readers and perhaps basing decisions on things that are not really there.

The simplest solution for avoiding this problem is to divide your *p* level by the number of times you are doing a statistical check and using the result at the *p* level for each check. For example, if you are using a *t* test nine times with the same data set, as did Matsuura, Chiba, and Hilderbrandt (2001, pp. 75, 77), you should divide the overall 0.05 *p* level by 9 and apply the resulting 0.005… *p* level to each of the tests. It can be seen in the table the authors present on page 77 that without this adjustment the authors would have "found" six reliable differences between "native speaker" English teachers and Japanese English teachers. On page 75, they correctly note that their data show only two differences after the correction is done.

An example of the trouble multiple comparisons can cause can be found in the studies on a similar topic presented by Chiba and Matsuura (2004). In this case, the authors analyzed the results of two surveys, but apparently did not use the correct adjustment. In the first survey, for example, six differences between "native" and "nonnative" teachers' ideas are reported (Table 1). A rough recalculation using the multiple comparisons adjustment (dividing by the number of tests used) seems to indicate that only three real differences exist. Likewise, in the comparison of "native" and "nonnative" teachers presented in Table 2, 15 differences are reported, but probably only 9 are real differences. [1] Similar problems with multiple comparisons can be found in (using an example I happen to have at hand) Wharton (2000).

CHOOSING STATISTICS TO PRESENT AND USE FOR ANALYSIS

All research in applied linguistics implicitly involves comparisons. This is true even in the case study, where ostensibly only one thing is studied: the researcher and consumer in case studies must distinguish the case in comparison with cases in the same category.

With largely descriptive surveys, explicit comparisons are possible but not necessary. Researchers or consumers can compare responses from one person with those of another. For example, Connolly (2004) presented responses of all his participants, which allowed the survey consumers (that is, the readers) to compare answers of one participant with another's or with the consumer's own. The comparison as such, however, was not a part of Connolly's research design.

In surveys involving larger numbers of participants, comparisons are inevitable parts of the research design, and the use of statistics to facilitate or make clear the comparisons is necessary, even when the researcher is ostensibly only describing a population. For example, Ito (2004) conducted a survey to investigate parents' reasons for choosing French immersion programs in Canada. Although the purpose of the survey was only to describe parents' reasons, Ito implicitly presents comparisons by ranking responses.

[1] Unfortunately, the schedules of Drs. Chiba and Matsuura did not allow them to address these questions in detail for this paper.

In surveys that involve some sort of non-binary scale (such as a Likert, Likert-type, or ranking scale), comparisons are usually explicit. When researchers choose to use a scale of some sort, they imply that the construct or behavior of interest is (a) something that can be measured, (b) that differences in the construct are important, and (c) that the possible responses on the scale represent real differences in the construct.

For example, Sawada (2004) chose a Likert-type scale to measure reactions to prompts related to motivation to study English. The first prompt is “I study English because I like it” (translations are mine and have not been back translated). The response scale is marked with “I strongly disagree” at 1, “Neither agree nor disagree” at 3, and “I strongly agree” at 5. The reader or participant assumes that 2 and 4, respectively, stand for “disagree” and “agree.” The consumer of the research presumes that Sawada contends that there is a meaningful difference between “I strongly agree” and “I agree,” and Sawada’s report includes an analysis that takes into account that difference, the analysis of variance (ANOVA) test to check for differences between groups. In another example, I asked participants (Molloy, 2004a) if their answer to a question if their answers to a question were “appropriate/*fusawashi*” before and after instruction. Because I do not believe appropriateness can be meaningfully be measured with a scale, I used a binary response format and analyzed the results with a statistical test that can check binary responses, the chi-square test.

When choosing statistics to use for analysis, it is necessary to think about what you can meaningfully measure and how it can be measured. (This has been covered in the companion article.) Details on exactly which tests should be used are available in the references I have given, but there are two questions that should be considered first, as most statistical textbooks I have seen (with the exception of Hatch and Lazaraton, 1991) do not cover a wide range of statistical texts.

Statistical procedures have traditionally been divided into two principal types: parametric and nonparametric. Parametric statistics are more powerful and more often reported, but can only be used when fairly strict criteria have been met. Non-parametric statistics are less powerful (that is, cannot detect very small, but real, differences), but have less strict criteria. The two most important considerations for choosing between the parametric and nonparametric families in survey analysis are sample size and the type of scale used in the survey.

HOW MANY PARTICIPANTS WERE THERE?

All parametric statistical tests have minimal sample size (number of participants) requirements, along with other requirements. The sample size requirements exist because of the way most parametric statistics work. They work by comparing the mathematical properties of your data set against the properties of known sets of numbers that resemble your data set. For example, the popular *t*-test works by comparing average scores from two groups to see what the chance the difference in scores could have been generated by random numbers. The test can work with an absolute minimum of 30 participants in each group, but only as long as the scores from each group are normally distributed (that is, in a bell-shaped curve), something you can check with, for example, the histogram function provided with Microsoft Excel. If you have fewer than 30 members in each group or if the scores are not normally distributed, you cannot use the *t*-test. Hence, with the first study reported by Chiba and Mastuura (2004), the *t*-test would seem to be inappropriate because there were only 16 participants in each of the two groups compared. In the second survey reported, there are enough participants (41) in each group, but at least some of the responses do not seem to make a bell-shaped curve.

Other parametric tests have higher (or much higher) participant number requirements. For example, the popular factor analysis procedure used for survey analysis has high minimal requirements. The requirements vary according to which statistical text consulted, but they range from 50 participants per question on the survey to the square of the number of questions plus 50. Hence, Ito (2004) would seem to have been mistaken to have attempted to use factor analysis for the analysis of a 32-question survey with only 262 participants: a liberal estimate of a minimal sample size would be 1074 participants.

If you have not met the minimal sample size requirements for a test, you either have to get more participants or choose a nonparametric test and consult Siegel and Castellan, Jr. (1988).

WHAT KIND OF SCALE DID YOU USE?

In applied linguistics, response scales used in surveys can be divided into three general types: nominal, ordinal, and interval. (A fourth major division used often in the physical sciences is the ratio scale, which has a true zero level, but ratio scales almost never exist in applied linguistics.)

A nominal scale is a scale on which participants can chose from mutually exclusive categories. These include multiple-response checklists (*e.g.,* Ito, 2004), binary-response or yes/no formats (*e.g.,* Molloy, 2004a), and the textbook topic classification scheme used by Yamanaka (2004). Comparing answers from a nominal scale can only be done with nonparametric statistics.

An ordinal scale is a scale in which participants rank items in order or choose one of a ranked set of responses. An example of item ranking can be found in the survey used by Ito (2004). Choosing from ranked responses can be found in the Likert-type scales used by Chiba and Matsuura (2004) or Sawada (2004).

An important idea about ordinal scales is that the distance between any two ranks is unknown. For example, if I were asked to respond with my liking for various kinds of healthful foods on a five-step Likert-type scale (with 1 as great liking and 5 as great dislike), I might respond to a “carrot” prompt with a 3, a “cauliflower” prompt with a 4, and to an “egg” prompt with a 5. I am indifferent to carrots, don’t like cauliflower much, and have a strong, visceral loathing of eggs. The difference in my feelings between carrots and cauliflower is a great deal different the difference between cauliflower and eggs. The scale I am responding to, however, does not capture this difference. A researcher might take an average of my responses to characterize my general liking of healthful foods, but this average would be inaccurate because the scale does not allow the capture of my extreme response to eggs.

Likert-type response scales are ordinal scales, but researchers often treat them as if they are interval scales, as if the distance between responses is equal. This is a common and accepted practice, but the researcher should be ready to defend the decision to use inappropriate statistics. Properly, Likert-type scale surveys should be analyzed with nonparametric statistics, but in some cases parametric treatment is tolerated.

Interval scales are ranking scales in which the distance between the response categories is known and the same between each possible pair. An example would be asking participants to write down the number of books they have read in the past month as a way of measuring love of reading. A participant who has read four books can properly be said to love reading twice as much as a participant who has read two books. True interval scales are seldom seen in applied linguistics studies (*e.g.,* Molloy, 2004b), but it is only with true interval scales that the powerful parametric statistics can be used without considerable theoretic buttressing. (A ratio scale is an interval scale that has a zero point.)

VALIDATION AND ERROR

How much do I weigh? Four hours before this writing I stepped on scale and got a reading of 69.7 kg. Forty-eight hours before that, I got a reading of 68.9 kg. Over the last four years, my weight has varied from a high of about 71 kg to a low of about 66 kg. What, then, is my weight? “About 69 kg” is a good enough answer for most purposes.

In analyzing results from a survey, however, “about” some units of one measure or another is not an acceptable answer. There are standard procedures for reporting the variation of any measurement. In the case of many cases being used to make a typical measure of one thing (that is, an average or mean), the standard deviation is usually the accepted indication of error. Standard deviation can be thought of a measure of how different any randomly selected measure is likely to be from the average.

In the case of my weight, over the past four years my weight has averaged 68 kg, but I have gotten many different weight measurements, depending on the time of day, how much clothes I have on, how much exercise I’d got, how much I’d eaten, and so on. About two thirds of the time, my weight reading has been in the range 67 to 69 kg. About 90% of the time, it has varied from 66 to 70 kg. Only once or twice over the past four years have I got readings of less than 66 or more than 70.

One way to report my weight would be to say that it has a mean of 68 kg with a standard deviation of 1 kg; this means that about 66% of the times I’ve checked my weight it’s been between 67 and 69 kg.

The standard deviation is the correct way of reporting the amount of variation in your data if you have a used a survey with numbers that are averaged and the numbers have a bell-shaped distribution. The formula for the standard deviation can be found in the help file of Microsoft Excel (1999). Showing results with averages is the most common way of reporting results in surveys that involve ordinal or interval scales, but it is not the only way.

Ito (2004), for example, used an ordinal ranking scale and showed percentages of participants responding to the various stimuli presented (in Tables 2 and 3, for example). In this case, Ito should have included an error column in the tables. The correct error measurement here would have been the standard error of proportions, the simple

$$\sqrt{\frac{p\times(1-p)}{n}}$$

where *p* is the proportion and *n* is the total number of objects being counted.
This can be done in Excel with =SQRT((*R**(1-*R*))/*n*), where *R* is the cell reference for the proportion for which you which an error calculation and *n* is the number of participants.

The reader can note in Ito’s Table 2 (2004), for example, that 3.3% and 2.5% of 122 participants were unable to tell if they were satisfied or not with their children’s French immersion programs or were not very satisfied, respectively. Four participants said they couldn’t tell, and 3 said they were not very satisfied: a change of one person would have reversed the rankings. If Ito had presented the standard error of proportions in an additional column, the reader would have been reminded of this important point.

Besides gaining and losing weight, I get different weight readings on different days because the scale I use gives slightly different weights for the same things under different circumstances. (The label on the scale says the weight shown is accurate to within 0.1 kg.) This is a matter of measurement error, not a matter of variation in my weight data.

Surveys are crude instruments for measuring mental constructs in applied linguistics, perhaps analogous to weighing oneself with a scale meant to weigh trucks. Hence, ideally surveys should not be used as a lone research instrument in a study and measurement error is usually fairly great with a survey instrument.

One factor that lends to the crudity is the error that attends your treatment of your data. Any abstraction or data reduction, necessary in survey analysis, involves inaccuracies. An average or a percentage can be a fair representation of your data, but necessarily involves error. It is essential, then, to include some sort of error or variation measurement in your data reporting so that readers can see how well the summary figures you report characterize your data.

RELIABILITY

Besides the error involved in abstracting or summarizing data, any research also involves error introduced by variability in the research instrument. In an interview, for example, the things the interviewee says can be affected by any number of factors, such as the time of day, mood, the interviewer, the setting, whether the interview is paid, the topic, and myriad others. The scale I use inn checking my weight does not always give the same measurement. Surveys as measurement instruments as well inevitably have some measurement error. This error is usually characterized as reliability.

WHAT IS RELIABILITY?

Reliability is the tendency for a measurement instrument to measure the same thing the same way each time it is used. Some instruments are more reliable than others. For example, if you measure a length twice with a ruler and twice by simply pacing off the object, you are likely to find that the steel ruler gives two measurements more alike than the two pacing measurements. The ruler than can be considered more reliable a measuring instrument. Likewise, if you measure something twice with a ruler marked by millimeters and twice again marked only by centimeters, you will likely find the millimeter ruler more reliable. Reliability is the tendency for an instrument to give the same results each time it is used.

Fortuitously, survey reliability is a well-studied matter, and a number of simple and well-known methods to measure survey reliability exist. When you use a survey, you should always check and report the reliability.

HOW RELIABLE DOES A SURVEY HAVE TO BE?

Reliability is usually reported as a proportion of 1, with 1 denoting a perfectly reliable survey. Perfectly reliable surveys are not impossible, but they are unlikely to yield any useful information. If I give a survey (in Japanese) to all of the members of one of my classes asking them how many hands they have, I could get exactly the same results if I gave the survey twice and a perfect reliability statistic, but I wouldn't learn anything interesting. Surveys that are used to garner useful information will be to some degree or another unreliable.

The rule of thumb for survey reliability is 0.80: any survey at least that reliable would be considered acceptably reliable for reporting in the literature. A reliability of 0.90 is very good; survey reliability of more than 0.95 is almost unheard of in applied linguistics.

The measurement analogies I gave above can be used to illustrate the two major factors that affect survey reliability: consistency of measurement and fineness of measurement.

With the measurement analogy, a steel ruler is more likely to be accurate because it will vary less from time to time than one's pace will. With surveys, instruments with well designed, unambiguous, and well tested prompts will be more likely to consistently elicit the same reactions every time they are administered. The measurements with a steel ruler and with well-written prompts will be more reliable because the instruments are less likely to change.

Comparing measurements with a millimeter-resolved ruler and a centimeter-resolved ruler can be considered analogous to measuring with a survey with 20 questions per topic or measuring with a survey with 1 question per topic. With surveys, all other things being equal, the more prompts there are the more accurate the survey will be. With surveys, of course, all other things are never equal, as using an inordinate number of prompts will fatigue the participants.

DIFFERENT FORMS OF RELIABILITY

Measuring the reliability of surveys can be done in several ways; each has advantages and disadvantages.

TEST-RETEST RELIABILITY

Test-retest reliability simply means using the same instrument with the same participants two times and comparing the answers. This is how you might test the reliability of a sphygmomanometer or a ruler. In applied linguistics, test-retest reliability is seldom used, for three principal reasons. First, it is expensive and inconvenient, necessitating twice as much time, twice as many copies, and generally twice as much trouble. Second, participants are likely to remember prompts, which may affect the results, and may become annoyed at or bored with having to respond to the same prompts twice. Third, if you wait long enough between administrations for participants to forget, you are unlikely to have precisely the same participants the second time, as they will have learned new things or, more generally, changed in differing ways.

EXTERNAL RELIABILITY

External reliability means comparing the measurement made by one instrument with that made by another. For example, Yamashita (1996) compared measurements made by three different surveys with each other and with other measurement instruments to assess the reliability of different ways of measuring pragmalinguistic and sociopragmatic ability in Japanese as an L2. External reliability checks have the advantage of allowing the researcher and the consumer a more thorough view of the phenomenon of interest. The disadvantages concern mainly expense: using more than one instrument can involve much work and time.

One form of external reliability that is underused in applied linguistics is parallel-forms reliability. In this reliability check, two different surveys, each measuring the same thing, are given to participants, one to each half of the group of participants. Responses can then be compared across groups: if the two surveys actually do measure the same thing, the results should be roughly the same. The disadvantage to parallel-forms reliability is that it necessitates developing and testing twice as many prompts. An example of a recent use of parallel forms can be found in Molloy (2004b).

INTERNAL RELIABILITY

Mainly because of the difficulty of using test-retest or external reliability measure, the most frequently used reliability measurements used with applied linguistics surveys are internal reliability measurements. (For further details and procedures, see Brown, 1996, 2001, or 2005, or Molloy, 2004c.) Internal reliability measurements work by comparing responses to certain prompts with to responses to all prompts concerned with a given construct or by comparing responses to half the questions concerned with one construct with the other half of the questions concerned with that prompt. The disadvantage of internal reliability measure is that they necessitate fairly large numbers of prompts to work properly (see Nichols, 1999). The very great advantage of internal reliability measurements and the reason for their popularity in applied linguistics is that they are easy to calculate and necessitate the least work of all reliability measures.

PRESENTING RESULTS

In presenting the results from a survey, according to APA style, you should present enough information so that the reader can "corroborate the analysis" (American Psychological Association, 1999, p. 112) or "reasonably replicate your study" (p. 15). With surveys, this involves following the requirements of APA style or similar style guides.

Among other things, you should be sure to report the reliability of your survey and the correct measuring of variation in responses, as well as the overall *p* level you have chosen and why you have chosen it. The APA style guide (particularly pp. 7-21 and 111-119 of the 4th edition) is a good place to begin to determine what is necessary and appropriate. A statistics textbook that covers the statistical procedure or procedures you use can tell you what information must be included. Unfortunately, novice researchers cannot rely on copying the information presented in published sources, as the literature is frequently deficient in reporting (*cf.* Wilkinson *et al.*, 1999).

The APA style guide is the bane of beginning applied linguistics researchers or graduate students. The sections that (in my experience) receive the most attention are, oddly, the least important: the sections on style, particularly as regards references to other published works. Matters of reference style, however, are trivial. They can be easily checked and corrected by copy editors or computer programs, and oddities in reference style merely cause annoyance for readers. With surveys, the most important sections are those that concern minimal reporting requirements (the first and third sections, not including the extensive material on references and mechanics). The minimal reporting requirements must be met in reporting the results of a survey, or the survey will not be understandable. Even if you are able to be published, you will nevertheless be doing a disservice to the applied linguistics community if you do not include enough information to allow readers to evaluate your research effort.

In some ways, an ideal way to present the results from your survey would be to publish all of the raw data along with all of your calculations and your interpretation of the results. However, most readers are busy, paper is expensive, and extreme detail is of limited interest to all but specialists whose interests are identical to yours. As a compromise, researchers should make efforts to make their data available to interested parties who wish to examine them. If possible include a notice to this effect in your report.

References

American Psychological Association. (1999). *Publication manual of the American Psychological Association (4th ed.).* Washington, DC: American Psychological Association.

Brown, J. D. (1991). Statistics as a foreign language—Part 1: What to look for in reading statistical language studies. *TESOL Quarterly, 26,* 569-586.

Brown, J. D. (1992). Statistics as a foreign language—Part 2: More things to consider in reading statistical language studies. *TESOL Quarterly*, 26, 621-664.

Brown, J. D. (1996). *Testing in language programs.* Upper Saddle River, NJ: Prentice Hall Regents.

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

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment.* New York: McGraw-Hill.

Cartmill, M. (1980). John Jones's pregnancy: Some comments on the statistical-relevance model of scientific explanation. *American Anthropologist*, 82, 382-385.

Connolly, M. (2004). Revisiting the special challenges of teaching lower-level Freshman English. *CELE Journal*, 12, 5-23.

Hatch, E. & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics.* Boston: Heinle & Heinle.

Hood, G. M. (2003) *PopTools version 2.5.9.* [WWW document]. URL http://www.cse.csiro.au/poptools.

Ito, H. (2004). Parental evaluative perceptions of immersion education in Canada. *JACET Bulletin*, 39, 123-135.

McCloskey, D. N. (1985). The loss function has been mislaid: The rhetoric of significance tests. *The American Economic Review*, 75, 201-205.

Microsoft Excel 2000 [Computer software]. (1999). Redmond, WA: Microsoft.

Molloy, H. P. L. (2004a, 3 November). Short report on pragmatics instruction and confidence [1369 words] [E-mail]. *JALT Prag SIG mailing list.*

Molloy, H. P. L. (2004b). Is appropriate appropriate? An investigation of interpersonal semantic stability. *Proceedings of the 2003 JALT PAN-SIG Conference.* Tokyo: Japan Association of Language Teachers.

Molloy, H. P. L. (2004c). How reliable is the Asia University Freshman English Placement Test? A classical internal reliability study. *CELE Journal*, 12, 64-86.

Nichols, D. P. (1999). My coefficient α is negative! *SPSS Keywords, 68.* [WWW document]. URL http://www.ats.ucla.edu/stat/spss/library/negalpha.htm.

Pezzullo, J. C. (2004). *StatPages.net* [WWW document]. URL http://members.aol.com/johnp71/javastat.html.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, 57, 46-428.

Sawada, M. (2004). Adult EFL learner motivation: Learning English as lifelong learning. *JACET Bulletin*, 39, 59-71.

Siegel, S., & Castellan, N. J., Jr. (1988). *Nonparametric statistics for the behavioral sciences.* New York: McGraw Hill College Division.

Wharton, G. (2000). Language learning strategy use of bilingual foreign language learners in Singapore. *Language Learning*, 50, 203-243.

Wilkinson, L., & The APA Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.

Yamanaka, N. (2004). An evaluation of English textbooks from the viewpoint of culture based on the 2003 Ministry of Education's Course of Study guidelines. *JACET Bulletin*, 39, 87-103.

Yamashita, S. O. (1996). *Six measures of JSL pragmatics* (Technical report 14). Honolulu: Second Language Teaching & Curriculum Center, University of Hawaii at Manoa.