Statistics as a Foreign Language—
Part 1: What to Look for in Reading
Statistical Language Studies*

JAMES DEAN BROWN
University of Hawaii at Manoa

This article is addressed to those practicing EFL/ESL teachers who currently avoid statistical studies. In particular, it is designed to provide teachers with strategies that can help them gain access to statistical studies on language learning and teaching so that they can use the information found in such articles to better serve their students. To that end, five attack strategies are advocated and discussed: (a) use the abstract to decide if the study has value for you; (b) let the conventional organization of the paper help you; (c) examine the statistical reasoning involved in the study; (d) evaluate what you have read in relation to your professional experience; and (e) learn more about statistics and research design. Each of these strategies is discussed, and examples are drawn from the article following this one in this issue of the TESOL Quarterly.

The TESOL Quarterly is currently the research journal of the organization, Teachers of English to Speakers of Other Languages. Ironically, many of the statistical studies on language learning and teaching that are found in the Quarterly may be incomprehensible to the very EFL/ESL teachers who make up the intended audience. Rather than bemoaning this situation (either by berating teachers for not knowing more about statistics or by criticizing researchers for producing articles that are frequently not accessible to teachers), this article will begin by accepting statistical language studies for what they are: legitimate investigations into phenomena in human language learning/teaching which include the use and systematic manipulation of numbers as part of their argument.

Notice that I purposely avoid terms such as empirical and experimental in referring to these statistical language studies. I am doing

* Part 2, scheduled to appear in Volume 26, discusses more advanced statistical procedures.
so for several reasons. First, there are other, nonstatistical studies
that could be called empirical (e.g., ethnographies, case studies,
etc.) since, by definition, empirical studies are those based on data
(but not necessarily quantitative data). Second, there are statistical
studies that are not exactly experimental in the technical sense of
that word (e.g., quasi-experimental studies, posttest-only designs,
etc.). Third, there are statistical studies that have little or nothing to
do with experimentation (e.g., demonstrations, survey research,
etc.).

Regardless of what studies are called, when confronted with
statistics, many readers will either skip an article entirely, or take a
rather cursory route through the paper. Such a route might include
skimming the abstract and the Introduction section, then skipping
over the Method and Results sections (with their tables, figures, and
statistics) to the Conclusions (and/or Discussion) section where they
look to find out what the study was all about. If this strategy sounds
similar to one that you use, you may be missing an opportunity.
Statistical reasoning is just a form of argumentation; by skipping the
Method and Results sections, readers not only miss the heart of the
study, but also buy the authors’ argument without critical evaluation.
Most of us would not surrender so easily if the form of argument were expressed in words rather than numbers. We would
read a prose article carefully and critically. We should not have to
surrender professional skepticism just because the form of argument may be a bit alien, that is, numerical.

The purpose of this article is to provide some attack strategies for
teachers to use in gaining access to statistical studies and in
understanding them better. In the process, examples will be drawn
from a study reported in the second article in this issue of the
TESOL Quarterly.

ATTACK STRATEGIES FOR STATISTICAL STUDIES

With the following strategies in hand, you may not understand
every word of a statistical study, but you will be able to gain access
to such studies and will have a purposeful way of grappling with the
content.

Use the Abstract to Decide if the Study has Value for You

Let us begin with the familiar and work toward the less familiar.
The portion of a statistical report that is probably most often read
is the abstract. An abstract typically contains about 150 words in the
TESOL Quarterly. Other journals may have somewhat longer or shorter abstracts. Regardless of their length, these handy summaries should contain enough information for the reader to know what the study was about, how it was conducted, and the general trend of the results. In other words, an abstract should tell the reader in a nutshell what is presented in the study and allow you to determine if an article is pertinent enough to your professional life and teaching situation to be interesting and worthy of your time.

Indeed, there is an overwhelming and increasing amount of information competing for our professional attention. Along with the TESOL Journal and the TESOL Quarterly, EFL/ESL teachers choose among other journals, such as the ELT Journal, the ESP Journal, Language, Language Learning, Language Testing, Modern Language Journal, Studies in Second Language Acquisition, and TESL Canada Journal/Revue TESL du Canada, to name just a few in the United Kingdom and North America, as well as Cross Currents and the JALT Journal (Japan), Prospect (Australia), RELC Journal (Singapore), System (Sweden), and many others.

Because of this plethora of journals, it is essential to use the abstracts to advantage. Consider the abstract associated with the example article that follows this one. Is there sufficient information in that abstract for you to decide whether the article is of interest to you?

Let the Conventional Organization of the Paper Help You

The TESOL Quarterly and many other journals in our field generally follow the format and organization described in the Publication Manual of the American Psychological Association (APA) (American Psychological Association, 1983). That manual advocates using the following general sections and subsections in reporting a statistical study:

- Introduction
  - Introduction to the problem
  - Background
  - Statement of purpose
- Method
  - Subjects
  - Materials (or Apparatus)
  - Procedures
- Results
- Discussion (and/or Conclusions)
- References
Typically in our journals, there are no headings for the Introduction or its subparts. However, beginning with the Method section (including any Subjects, Materials or Apparatus, and Procedure subsections) through the Results, Discussion, Conclusion, and References sections, you will generally find clear headings and subheadings. Since the general purpose of the headings and subheadings is "to help readers find specific information" (American Psychological Association, 1983, pp. 25-26), you should use them to help you find and organize the information that you need in order to understand the study. There is not space here to provide details about what each of these sections should contain. Indeed, such details are not necessary here because existing sources give ample information on this topic (e.g., Brown, 1988; pp. 43-62; Hatch & Farhady, 1982, pp. 33-38; Hatch & Lazaraton, 1991, pp. 107-126).

Nevertheless, there are a number of questions that you might want to ask yourself as you read through a statistical study. These questions should help focus the information contained in these key sections and help readers critically evaluate a study. Notice that section and subsection headings are listed in parentheses after each of the questions below. These are meant to suggest where you would typically find the information that would answer each question.

1. What literature is reviewed? Is the review current and complete? Where does the study fit into the field? (Introduction section)
2. What is the purpose of the study? (Introduction section, especially the Statement of Purpose subsection)
3. Who was studied and how were they selected? Were there enough people in the study to make the results meaningful? (Subjects subsection)
4. What tests, questionnaires, rating scales, etc., were used? What do they look like? And, are they reliable and valid for the purposes of the study? (Materials subsection)
5. What actually happened to the subjects during the data gathering process? (Procedures subsection)
6. How were the data organized and analyzed? (Results section)
7. Is there enough information provided to replicate the study? (throughout the Method section including the Subjects, Materials, and Procedure subsections)
8. What were the descriptive results? What other statistical results came out of the study? (Results section)
9. What were the answers to the research questions and what do the answers indicate? (Discussion section)

10. What are the implications of the results, and how do they relate to the field as a whole? (Discussion or Conclusions section)

11. Which conclusions follow directly from the results and which ones are more speculative? (Discussion or Conclusions section)

12. What questions arose in the course of doing the study that might be useful for future research? (Discussion or Conclusions section)

Since answers to these questions are important in understanding any study, you can use the conventional organization of statistical studies as represented by the sections discussed here to find your way around published research articles.

However, be warned that, even though the APA format and organization are well known, some authors do not use exactly the sections and headings listed above. Sometimes there are good reasons for such deviations. For example, in the example article, there is a separate Procedures subsection as advocated in the APA manual. However, in other studies, the same author has chosen to use a combined Materials and Procedures subsection because the two issues were inextricably intertwined to the degree that they made little sense if explained separately. Regardless of the specific sections and headings used in a particular study, you should find sufficient information somewhere in any statistical study to answer the twelve sets of questions listed above.

At this point, you may wish to turn to the example article. Without reading every word, try jumping from section to section while letting the above questions guide what you have to read in order to answer them. In the process, notice that everything is found just about where the APA manual suggests. Notice also that the organization and headings help you to find the information that you need and to generally understand the study.

Examine the Statistical Reasoning Involved in the Study

In order to understand the results of a statistical study, it is necessary to understand the statistical reasoning that underlies all such studies. There are several key concepts that are necessary parts of much statistical reasoning: (a) descriptive statistics, (b) statistical differences, (c) probability levels, (d) statistical tests, and (e) significance versus meaningfulness. These five concepts will be discussed in turn.
Descriptive statistics. Part of the content of any statistical paper describes what happened in the study. As mentioned above, such description is partially accomplished within the various parts of the Method section. However, description also occurs in the Results section, which describes what happened statistically. Descriptive statistics (a phrase used in contrast to inferential statistics) describe or summarize a data set but, by themselves, cannot tell us the extent to which they represent a larger population or other, similar sets. The descriptive statistics that are most often used are indicators of the central tendency and dispersion. The central tendency (which can be viewed as a typical value for a set of numbers) is commonly reported in terms of a statistic called the mean. (A statistic is any number that can be computed based on the observed data.) The mean is usually exactly the same as the arithmetic average that we use in daily life. Dispersion (which can be viewed as the variation of the numerical values away from the central tendency) is usually reported in terms of a statistic called the standard deviation. The standard deviation summarizes how much the numbers vary away from the mean, or how much they are spread out around the mean. Ordinarily, we would expect most scores (about 68% for a normal distribution) to fall within one standard deviation of the mean. For a normal distribution, 95% of the data fall within 2 standard deviations. For instance, if the data being described are a set of test scores, their standard deviation can be defined as “a sort of average of the differences of all scores from the mean” (Brown, 1988, p. 69).

You now have enough basic information about descriptive statistics to consider the central tendency and dispersion in a real study. Such descriptive statistics most often take the form of a table. Table 2 in the example study is typical. Notice that it describes the results in terms of the various groupings involved in the study, that is, for each group and/or subgroup in the study. The first column labels groups that were created by considering the raters in two faculties (English and ESL) separately as well as combined. Across the top of the table, you will also find labels for the groups that were created by considering the different types of students separately (English and ESL composition students) and combined. The most important thing to note is that the data being grouped and described are the ratings for two types of students (English and ESL) as rated by two types of instructors (from English and ESL faculties).

Notice that for each possible combination of Student Type, and Rater Faculty, the table provides three statistics: the number of students involved in the group (n), the mean (m), and the standard
deviation (SD). For instance, the descriptive statistics for the English students’ compositions as rated by the English faculty indicate that there were 112 compositions involved, that the mean was 2.46 (on a 0-to-5-point scale), and that the standard deviation was 1.11. This information is interesting in itself because it indicates that the test is fairly well centered (i.e., the mean is almost exactly halfway between the lowest possible score of zero and the highest possible score of five), and the scores are spread out to a reasonable degree (i.e., there is room for 2 standard deviations above and below the mean within the range of possible scores from 0 to 5).

However, the information is also useful for comparing groups to each other. Consider the fact that the means for all of the groupings are very similar in this table. That may indicate that there were no major differences among the groups of compositions as produced by the two types of students and rated by the two types of teachers. However, note that the standard deviations for the compositions written by the English students are generally higher than those for the essays written by the ESL students. This indicates that the scores for the English composition students were more spread out than those for ESL students, that is, there was greater dispersion in the writing scores of the English course students.

Statistical differences. As useful and informative as descriptive statistics can be, they are often not enough. There is a type of statistical reasoning that takes over at this stage in most studies: Inferential statistics investigate the extent to which descriptive statistics represent a larger population or other, similar data sets. This mode of reasoning hinges on the concept of significant differences. The significant differences most often of interest in statistical studies are the differences observed in comparing means, comparing frequencies, or comparing correlation coefficients to zero.

In comparing means (i.e., arithmetic averages), it is possible that any observed differences are purely accidental. After all, if you give a test to a group of students on several occasions, you would expect the means to be slightly different because human beings simply do not perform exactly the same on every occasion (e.g., some students’ scores might have been affected by the fact that they were sick, tired, depressed, etc., on one of the occasions). Indeed, it would be very surprising if the test results turned out to be exactly the same on successive occasions. The issue that researchers must grapple with is whether the differences that they observe between means are just such chance variations or are due to some other more systematic factor. The question being posed by the researcher and
answered through statistical tests is whether or not there is a significant difference between means.

For example, consider a hypothetical study in which the average number of correctly defined words on a vocabulary test is compared for two groups: one group that received lessons using language teaching Method X for 6 weeks, and the control group that received 6 weeks of instruction based on Method Y. The problem is that the two means will naturally vary to some degree by chance alone. The question that the researchers must resolve is whether there is a significant difference between the means (i.e., whether the observed difference between the means is systematic or occurred by chance alone).

If there is a significant difference, the researcher can say with a certain amount of confidence that the observed difference between the two means was not just accidental. This is an important issue because, if the group learning vocabulary under Method X has a higher mean than the other group, the higher number of vocabulary words learned can probably be attributed to the effects of Method X (provided the experiment was conducted properly). This would constitute an argument in favor of Method X, which might be interesting to other language educators responsible for teaching vocabulary. As described here, it is an argument based entirely on comparing the mean performances of the groups involved.

In *comparing frequencies*, it is also possible that any observed differences are due to chance variations. After all, if we tally the numbers of Taiwanese and Korean students in an ESL class on successive days, we would expect the resulting frequencies (also known as tallies) to be slightly different on different days. It might turn out that there are 7 Taiwanese and 11 Koreans on the first day, 6 Taiwanese and 12 Koreans on the second day, 7 Taiwanese and 9 Koreans on the third day, etc. Indeed, it would be surprising if the frequencies turned out to be exactly the same every day. Similarly, by chance alone, one would expect the numbers of Taiwanese and Koreans to vary from group to group because of chance variations in the proportions of different nationalities. The issue that researchers must come to grips with is whether any difference that is observed between frequencies is a chance variation or is instead due to some other more systematic factor. In short, the question being posed by a researcher who is comparing frequencies is whether or not there is a significant difference between the observed frequencies and the frequencies that would be expected by chance alone.

An example of this type of study might occur in performing a needs analysis for a language course. Perhaps the researcher is
interested in the frequency, or tally, of people who are interested in studying pronunciation as compared to those interested in studying grammar. The question that a researcher might pose is whether there is a significant difference between the existing frequencies and the fifty-fifty split that would be expected by chance. If the researcher can state that there is a significant difference, it will indicate that the observed difference is due to factors other than the chance fluctuations that would occur naturally.

Similar reasoning may be used in comparing correlation coefficients to zero. Correlation coefficients are indexes that represent the degree of relationship between two sets of numbers. Correlation coefficients can range from 0.00 (if there is no relationship) to 1.00 (if there is a very strong relationship). Consider the following data set:

<table>
<thead>
<tr>
<th>Students</th>
<th>EFL Test A</th>
<th>EFL Test B</th>
<th>Last EFL Study (years since)</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maria</td>
<td>100</td>
<td>97</td>
<td>0.0</td>
<td>26</td>
</tr>
<tr>
<td>Jaime</td>
<td>98</td>
<td>97</td>
<td>0.5</td>
<td>28</td>
</tr>
<tr>
<td>Carla</td>
<td>87</td>
<td>84</td>
<td>2.0</td>
<td>26</td>
</tr>
<tr>
<td>Jose</td>
<td>82</td>
<td>85</td>
<td>1.5</td>
<td>28</td>
</tr>
<tr>
<td>Juanita</td>
<td>77</td>
<td>74</td>
<td>4.0</td>
<td>27</td>
</tr>
<tr>
<td>Jimmy</td>
<td>55</td>
<td>52</td>
<td>3.5</td>
<td>27</td>
</tr>
</tbody>
</table>

EFL Tests A and B appear to be highly related in the sense that as one set of numbers goes up so does the other. The resulting correlation coefficient turned out to be a very high .99. If two sets of numbers are not related at all, the expected correlation coefficient is 0.00. This is the case for Age and Test B in the above example data where the correlation coefficient turned out to be very close to zero at 0.01.

It is important to note that correlation coefficients can also take on negative values anywhere between 0.00 (if there is no relation-
ship) and −1.00 (if there is a strong but opposite relationship). In
such cases, the two sets of numbers are related but in opposite
directions. For example, Test B and Last EFL Study in the above
example are fairly strongly related but in opposite directions. In
other words, as one set of numbers goes up, the other set goes
down. The result in this case is a high but negative correlation coef-
ficient of −.86. In short, correlation coefficients can range from
−1.00 (for strong, but opposite relationships) to 0.00 (for no rela-
tionship) to +1.00 (for strong relationships in the same direction).

One problem with correlation coefficients is that even for two
sets of random numbers some degree of correlation may be found
by chance alone. For instance, on successive sets of randomly selected numbers, correlation coefficients of .12, .07, −.17, −.01, etc., might be found by chance alone. Indeed, it would be surprising if a correlation coefficient of exactly 0.00 were found every time. The issue that researchers must deal with is whether the correlation coefficients that occur in a study are just such random (or chance) variations around zero, or rather are due to some more systematic relationship between the sets of numbers. The question being posed by the researcher is whether there is a significant relationship, i.e., a significant difference between the correlation coefficient that was observed in a study and a zero (or chance) correlation. If the researcher can state that there is a significant difference, the correlation coefficient observed in the study probably varied from zero for other than chance reasons. In simpler terms, it indicates that the relationship between the sets of numbers is probably systematic—not just a chance relationship.

**Probabilities.** Since there is always at least some possibility that differences are due to chance, researchers use statistical tests to compute a particular significant difference in terms of the probability that observed differences would occur by chance alone. In other words, when a researcher states that there is a significant difference (between the two means, between observed and expected frequencies, or between a correlation coefficient and zero), these results will always be stated in terms of the probability that the observed difference was due to chance fluctuations.

These probabilities are usually expressed as *p* values in statistical studies. They will normally be written as *p* < .01, or *p* < .05, or as exact figures, e.g., *p* = .9681. The *p* stands for probability. In straightforward terms, *p* is the likelihood that the researcher will be wrong in stating that there is a statistical difference (whether between means, between observed and expected frequencies, or between a correlation coefficient and zero) if, in fact, there is no difference. Thus, if *p* < .01, the probability (assuming chance alone) is less than 1% that the observed difference would be so large, giving strong evidence against pure chance. Similarly, if *p* < .05, the probability is less than 5% that an observed difference this large could have occurred by chance alone.

The probability levels of .05 and .01 (also referred to as alpha levels) are used by convention in most social science research to define the threshold of statistical significance. The choice between the .01 and .05 values is governed by how strict the researcher wants to be with regard to the conclusions that are drawn from a study. When a study is about an important medicine, we want to be very
sure that it will not hurt patients. Thus a conservative .01 value might be selected so that there is a 99% probability of nonchance results. If the study is about a new way of teaching reduced forms, the decision is perhaps not quite so crucial, and therefore, we can accept the .05 value, which indicates that we are willing to accept a less restrictive 95% probability of nonchance results.

The determination of significant differences and their associated probabilities takes many forms, but the most commonly reported types are the three that have to do with means, frequencies, and correlation coefficients. It is important to note that a study seldom compares only two means, or contrasts only two frequencies, or examines one correlation coefficient to see if it varies from zero. More commonly, there are a number of means involved, or a number of frequencies, or a number of correlation coefficients to complicate the picture. Nevertheless, the underlying processes of checking significant differences and determining probabilities are the same.

Statistical tests. The process of determining statistical significance as described in the Statistical Differences and Probabilities subsections above is referred to as performing a statistical test. The three most commonly reported types of statistical tests are used in the example study: mean comparisons, comparisons of frequencies, and comparing correlation coefficients to zero.

Example mean comparisons are discussed in the example study:

In short, the small differences among the means shown in Table 2 can only be interpreted as chance fluctuations, which are not attributable to systematic differences based on the variables used in this study.

In the above quote, the statistical reasoning for comparing pairs of means is explained. In this case, there were three comparisons of interest: (a) the difference between the mean scores for the two types of students, English composition or ESL composition students; (b) the difference between the mean scores assigned by the English faculty and ESL faculty raters; and (c) the difference between the mean scores that resulted from the two different orders in which compositions were rated (first or second). In all three comparisons, it turned out that there were no significant differences between the two types of students, the two types of raters, and the two types of orders.

The statistical test being used is the $F$ test, the results of which are based on the $F$ statistic reported in the second column from the right in Table 3 of the example study. The essential information is found in the column on the far right where the probabilities are
given in the column labeled \( p \). Notice that each of these \( p \) values has an asterisk next to it and that the asterisks refer to the statement at the bottom of the table indicating that \( p > .05 \). This is read as "the probability is greater than .05" and indicates that random chance alone could produce results like these more than 5% of the time; this is not convincing evidence against chance. Thus, the researcher would be wrong in stating that there was a significant difference. (In fact, the \( p \) values reported in the table are much higher than .05, so they indicate that the observed results are quite consistent with random chance.) The first, second, and fourth \( p \) values are associated with one \( F \) ratio each for the main factors in Table 3: Student Type, Rater Faculty, and Order. The quote cited above discusses what these three statistical tests indicate.

**Example frequency comparisons** are shown in Tables 4 through 7 of the example study. The statistical test being used is known as the chi-square test, or simply \( \chi^2 \). You will notice the asterisks in each table refer to the statement at the bottom of the table that \( p < .05 \). In this case, the statement would be read as "the \( p \) value is less than .05" and indicates that the researcher was justified (in those cases marked with an asterisk) in stating that there was a significant difference between the relative frequencies of English faculty and ESL raters who chose a particular feature, that is, the observed differences are unlikely to have occurred due to chance alone. Hence, we conclude that there are systematic differences between English faculty and ESL raters.

Notice that there are two steps involved in interpreting Table 4. First, there is the overall \( \chi^2 \) value to consider. This value, located in the bottom row, is found to be significant at \( p < .05 \) (as indicated in the line just below the table). This result simply suggests that one or more of the frequencies in the table differed from what would be expected, and more detailed analyses are justified. In order to investigate which of the specific pairs (English faculty and ESL raters) of relative frequencies might be contributing to the overall significant difference, the \( \chi^2 \) values for pairs were also calculated. These \( \chi^2 \) values are reported in the column furthest to the right. They indicate that there was a significant difference (at \( p < .05 \)) for the English faculty raters and ESL raters on Cohesion, Organization, and Syntax (i.e., the frequencies observed for the English and ESL raters on Cohesion, Organization, and Syntax were significantly different). In contrast, the frequencies of response for English and ESL faculty were not significantly different from expectations for the rating categories of Content, Mechanics, and Vocabulary. Similar two-step interpretations can be drawn from each of the other frequency tables (Tables 5-7).
Example comparisons of correlation coefficients to zero are shown in Table 1. The statistical test results are based on the Pearson product-moment correlation coefficient, or simply r. The asterisks in the table once again refer to the statement at the bottom of the table that \( p < .05 \). The statement would be read as "the \( p \) value is less than .05," which, in this case, indicates that the researcher was justified in each case in stating that the correlation coefficient was significantly different from zero. More specifically, random chance produces such strong correlations less than 5% of the time. Based on this evidence against pure chance, the researcher was justified in stating that there was a significant difference between each correlation coefficient and zero. For instance, the results in Table 1 show that, even though it is relatively low in magnitude, the correlation coefficient of .37 between Groups A and B of the English faculty raters differs from zero for reasons other than chance. The same is true for all of the other correlation coefficients in this table. Such is not always the case. Other studies may well find correlation coefficients that are not significant, indicating just chance differences from zero with no systematic association between the sets of numbers involved.

**Significance versus meaningfulness.** It is important to recognize that a statistically significant difference is just that, and no more. Significant differences, whether working with means, frequencies, or correlation coefficients, simply indicate that we have concluded that the observed differences are due to other than chance factors. In other words, the differences are systematic in some way. It does not indicate that the differences are necessarily interesting or meaningful. In fact, a difference can be statistically significant, yet be so small that it is not at all meaningful or interesting.

For instance, the correlation coefficient of .37 was found to be statistically significant at .05 (i.e., the correlation coefficient is probably different from 0.00 for other than chance reasons), but the meaningfulness of the relationship between the two sets of numbers is a separate issue. In this case, the numbers are scores assigned by two raters and the low correlation coefficient indicates that there was some association between the two sets of scores but there are other important factors that are still not accounted for. The weakness of agreement found here is worrisome because it indicates that the scores may not be very reliable. Thus, this is an example of a correlation coefficient that is statistically significant, but not very meaningful in magnitude.

Similarly, if two means are statistically different at \( p < .05 \), yet only differ by two points out of 100, then the result might not be at
all meaningful. Likewise, if a set of observed frequencies differs from expected frequencies at $p < .01$, yet differs to an uninteresting degree, the results may not be meaningful. Thus, it is always important to examine the descriptive statistics in any study and think about any statistical tests in terms of descriptive statistics so that you can determine whether any significant differences are also meaningful.

The important thing to remember, then, is that meaningfulness is a separate issue from statistical significance and that meaningfulness will depend on all of the factors involved in the situation in which the study was conducted. When reading a statistical study, you might want to check to make sure that the researcher has kept separate these two issues of significance (i.e., can we rule out chance?) and meaningfulness (i.e., is the difference large enough to be interesting?) and interpreted them clearly.

It is important to remember that statistical studies are no more likely to be infallible than any other form of argumentation. Authors make errors, and computers make errors. However, if a study is properly carried out and the results are adequately described and systematically explained, such studies can help us to view the important issues in our field in new and useful ways.

**Evaluate What You Have Read in Relation to Your Professional Experience**

So far, you have used the organization of the abstract to decide if you wanted to read the study, used the organization of the paper to help you understand how the study was conducted, and used some basic concepts to interpret the statistical reasoning of a study. You are probably now at a point where it makes sense to pull away from the study a bit and think about it more critically. There are six types of questions that may prove useful in thinking about the article after having read it. These questions will enable you to know, comprehend, analyse, apply, synthesize, and evaluate what you have read. (These six categories are taken from Bloom's [1956] taxonomy. Note that they are presented here in a slightly different order from the original.) In short, after reading an article, try to recall basic information about the article by asking yourself Questions 1–3 below; then try to relate the article to your professional life by asking yourself Questions 4–6:

1. **Know:** Who wrote the article? When? In what journal? (Useful for identifying the study when referring to it)
2. **Comprehend:** In a sentence, what was the article about? (Useful for briefly summarizing the study)
3. **Analyze:** What sections was the article divided into? (Useful for recalling the overall structure of the study)

4. **Apply:** How can you apply what you learned in the article to your professional EFL/ESL teaching situation? (Useful for determining whether the article is applicable to your teaching experience)

5. **Synthesize:** How does the article relate to other professional books or papers that you have read? (Useful for seeing how the study fits into the professional literature)

6. **Evaluate:** How good was the quality of the article internally (in terms of style, organization, reasoning, etc.)? How good was it externally (i.e., in terms of everything else you know about the profession)? (Useful for evaluating the overall quality of the article)

Going through these questions (or similar ones) will help you to remember which article you read, comprehend its essential message, analyze the constituent parts of the article, apply what was learned in the article to your professional situation, synthesize what you found in the article with other points of view in the profession, and evaluate the quality of the article (both internally and externally).

**Learn More about Statistics and Research Design**

Having gone this far in the process of understanding statistical studies, you may now be intrigued by the prospect of learning more. For instance, you may have heard about ANOVAs, regression analyses, factor analyses, and other analyses not directly covered in this article. It is only by learning more that you will be able to understand some of these more complex analyses. In fact, it is only by learning more that you will be able to decide whether the author of a given study chose the correct statistical tests at all, or whether the assumptions that are required for many statistical tests were met.

There are a number of ways to learn more about statistics and research design. In addition to Part 2 of this discussion, there are books specifically designed to help language teachers do statistical research: Butler (1985), Hatch and Farhady (1982), Hatch and Lazaraton (1991), Seliger and Shohamy (1989), Woods and Fletcher (1986), etc. Another book, Brown (1988), is designed to help readers who are only interested in reading (rather than doing) statistical research. If the topics that interest you are more closely related to
the statistics and research in the area of language testing, it may be more appropriate to read references such as Bachman (1990) and Henning (1987). If you have no idea which book to choose, it might be useful to read Hamp-Lyons' (1989, 1990) book reviews which describe a number of the volumes listed above.

I am not advocating that every EFL/ESL teacher read and internalize all of the knowledge in these books. However, I am suggesting that a number of strategies are available to teachers: (a) for some teachers, a thorough reading of one or two of the books listed above may be just what is needed; (b) for other teachers, it may prove useful to use several of the books listed above as references to explore topics in more depth as need arises in reading statistical studies; (c) still other teachers may be more comfortable with the structure provided by taking an organized course in basic research design and statistics at a local college or university.

Regardless of the strategy that is used, learning more about statistical research can help not only in understanding the statistical studies in the professional literature but also in grappling with the research that is reported in the lay media, much of which is done in the same statistical research paradigm that is used in our field. (For an excellent and easy to read treatment of how numbers, figures, and tables are used to fool the general public, you may want to read a book appropriately titled *How to Lie with Statistics*, Huff & Geis, 1954. Yes, it is still in print.) Armed with such knowledge, teachers can then defend themselves against numbers, and understand the reasoning that surrounds their use.

CONCLUSION

This article set out to provide attack strategies for EFL/ESL teachers to use in gaining access to statistical studies. These strategies include using the abstract and conventional organization of statistical papers to guide reading, examining the statistical reasoning, critically evaluating what the results signify to each reader, and learning more about statistical studies. There are a number of reasons why I hope that some readers will find these suggestions useful. First, if the studies that appear in the *TESOL Quarterly* have a larger informed readership, such studies will have greater impact on the field. All of us must use all available information about language learning and teaching to improve the ways that we serve our EFL and ESL students. Second, it is only by having an informed readership that the quality of the statistical studies in the *TESOL Quarterly* can be assured. Though the review process for selection of articles is thorough and fair, there are no guarantees that the

584
articles that appear in print are 100% correct or uncontroversial. It is therefore our responsibility to read any articles that interest us as carefully and critically as we can so that the interface between teaching and research can be strengthened.

ACKNOWLEDGMENTS

I would like to thank Kathleen Bailey, Graham Crookes, Thom Hudson, Andrew F. Seigel, and Ann Wennerstrom for their insightful comments and suggestions on an earlier version of this paper.

THE AUTHOR

J. D. Brown is on the faculty of the Department of ESL at the University of Hawaii at Manoa. He has published numerous articles on language testing and curriculum development, and a book on critically reading statistical studies (Understanding Research in Second Language Learning, 1988, Cambridge University Press).

REFERENCES