Communicating Our Research Findings: Do We Say What We Mean and Mean What We Say?

Ratnawati Mohd Asraf
Department of Education
International Islamic University

James K. Brewer
Department of Educational Research
College of Education
Florida State University

In the field of linguistics and language teaching, the trend in research over the past 25 years has been toward the quantitative (Henning, 1986), which is commonly understood by researchers to be those studies that use statistics in the description and analysis of data. Indeed, not only are more quantitative studies being published in our journals, there is a distinct trend, according to Henning, toward greater use of inferential as opposed to descriptive statistics. Although Henning sees this as bringing language acquisition research into the realm of scientific inquiry, many have expressed concern regarding its misuse (Bakan, 1967; Brewer, 1991; 1996; Carver, 1978; Hays, 1994; Oakes, 1986; Shaver, 1993). Most of these concerns have to do with researchers' extrapolating their findings beyond what statistics can actually say, thereby having serious implications on the meaningfulness of their findings and on the validity of the conclusions reached. Concern has also been expressed about the inconsistencies in researchers' adherence to conventions of quantitative research methodology (Brown, 1991; Henning, 1986)

While research in linguistics and applied linguistics has increased tremendously in the past two decades, the findings, as Celce-Murcia and McIntosh (1979) have noted, have not always filtered down to educators. Hence, they have urged that there be greater co-operation and communication between researchers and educators. Clear communication of research findings is essential if they are to be fully utilised by teachers in helping them make decisions pertaining to the kinds of materials to use and language teaching practices to adopt. However, before teachers can benefit from research findings, both researchers and teachers need to have a basic literacy in research design and statistical concepts—especially if these findings are the result of statistical analyses (Dunkel, 1986; Flynn, 1985; Lazarton, Riggenbach, and Ediger, 1987)

There are indications, however, that such may not be the case. Brown (1991), for instance, notes the lack of literacy in research design and statistical concepts among teachers of language who are the primary audience of research journals such as *The TESOL (Teachers of English to Speakers of Other Languages) Quarterly* while Lazarton et. al. (1987), in a survey of university professors and researchers active in the field of ESL in the USA concluded that "there is a considerable range in the degree of familiarity with the concepts and procedures associated with empirical research" (p. 275) among their respondents. Crookes (1991) has also expressed the same sentiment regarding researchers' understanding of other central concepts in statistics.

While basic literacy in statistical concepts and research methodology is important in clearly communicating research findings, equally important is the careful attention that should be paid to the proper choice of words. Words have to be carefully chosen to convey the meaning we intend to convey. This is especially so when communicating quantitative research findings because words to which we have ascribed certain meanings in the English language have assumed a more specific meaning or interpretation when used in statistics or quantitative research. Careless use of these words would render a meaning different from that which we had intended.

This paper focuses on how researchers communicate their quantitative research findings. Although there is more to research than the quantitative, and that quantitative methodology is not necessarily the best means to address research questions in our field, the fact that the bulk of our research seems to be dominated by the use of statistics (Henning, 1986) warrants a discussion of the logic underlying its use as well as how it has been misused by researchers and language professionals. More specifically, the purpose of

this paper is to highlight how researchers might unintentionally misrepresent their research findings because of their choice of words and also possibly because of their misconception of certain statistical concepts. Finally, this paper discusses how results should be reported and conclusions written so that the words we have chosen are true to the meaning that we intend to convey

The Theoretical Background

The Message Model of linguistic communication as applied to written communication can be summarised as follows: Linguistic communication is as successful as it is because messages have been conventionalised as the meaning of words or expressions. Thus, by sharing knowledge of the meaning of a word or an expression, the *reader* can understand a *writer's* message (Akmajian, Demers, Farmer, and Harnish, 1990).

However, as is commonly known, words can also be defined to satisfy the purpose of the individual who uses them (Savory, 1953; Smith and Ennis, 1961). Because of this, many controversies arise over the meaning of terms. Such situations—where the same word can have different meanings for different people—can cause problems in communication. This is especially so when we are dealing with terms in statistics, many of which have been borrowed from our everyday vocabulary, but imbued, however, with a new and special meaning. The use of these "special" words in the context of research or statistics is likely to lead to confusion if we use them in a manner to which we are accustomed and not in the specialised manner in which they are intended to be used.

Because research is a communication process—although not always thought of as such—it is important that researchers pay careful attention to the proper choice of words when communicating their research findings. This communication requires a preciseness of language far more exacting than that demanded of everyday conversation because the words we choose have implications as to the truth properties of the statements that we are making. To highlight the importance of being precise in our use of terms, it would be important, at this point, to enter into a discussion of the logic underlying the use of statistics because it is this logic and the purpose for which statistics is put that will determine the words that we could use to report our procedures and findings.

The uses of statistics

Statistics, in the minds of most people, has to do with a description of "how things are" (Hays, 1994) Indeed, one of the uses of statistics is to make sense of and to summarise large amounts of quantified data in order that we may describe their general characteristics. This would include such things as describing the average age of students in a particular language classroom or describing the percentage of errors made by a particular group of students in the use of certain irregular past tense forms. Such would constitute the *descriptive* use of statistics, and the tools required to achieve this purpose, *descriptive statistics*.

Very often, however, we are concerned not just with the characteristics of one particular set of (observed) data, but rather, in going beyond this data to make general statements about a large body of unobserved data, of which the data collected are but a sample. For instance, we might be interested in estimating (with a certain degree of confidence and accuracy), using the sample percentage, the true percentage of teachers in Selangor who feel that grammar should be emphasised in the teaching of English in schools. Or we might be interested in testing, for example, the hypothesis that the true average scores of two groups of learners, taught by two different methods, do not differ on a test of language proficiency. The construction of confidence intervals that would be involved in the former and the testing of hypotheses that would be involved in the latter constitute the *inferential* use of statistics.

Because the using of statistics for an inferential purpose rests on the (conceptual) possibility of repeated observations made under essentially the same conditions, there will always be *uncertainty* connected with observation of any given object or phenomenon (Hays, 1994). Hence, inferential statistics is a theory about uncertainty, or the tendency of outcomes to vary when repeated observations are made under essentially the same conditions. In other words, because the use of statistics for an inferential purpose involves the making of inferences from a *limited* sample of scores to that of *the entire collection* of unobserved scores (population), that inference will be subject to error; that is, there is the *probability* that any statistical inference that we make—whether it involves confidence intervals or conclusions reached as a result of hypothesis testing—might be in error Hence, if we are estimating, using our sample percentage, the percentage of teachers in Selangor who feel that grammar should be given greater emphasis in schools, that sample percentage we have obtained will not necessarily reflect the true population

percentage (that is, the percentage of teachers whom we did not survey as to how they feel about the same issue) In fact, what is commonly done, in estimating population values (parameters) from sample values (statistics) is to specify the degree of confidence in the estimate and the accuracy of the estimate.

Suppose we wish to estimate, using our sample percentage, the true (population) percentage using 97% confidence intervals with a certain prespecified accuracy (for example, an accuracy of ± 5%) The probability that the true population percentage will be captured by intervals like the one we produce from our sample percentage will be .97, while assuring that, in the long run, sample percentages will be no further than .05 from the true percentage. The true percentage will fail to be captured with probability .03 (i.e., 1- confidence). In essence, we can say it is a good, safe "bet" that the true percentage is contained within our specific sample confidence interval even though the probability of containment applies to *all* possible intervals that could be calculated using our sample size. To enable us to construct confidence intervals with a prespecified level of accuracy and probability of error in the estimate, however, we would have to meet the minimum sample size required to justify those values.

Another statistical inference technique, and one that is perhaps more familiar to researchers and more frequently used—though not always appropriately—is hypothesis testing. As with confidence intervals, hypothesis testing also involves the probability of error For studies of any given sample size, there will always be the *probability* of making Type 1 error; i.e., the *probability of rejecting* the null hypothesis when the null hypothesis is indeed *true* (in other words, the probability of incorrectly rejecting the null hypothesis). There is also the *probability* of making Type 2 error; that is, the *probability of not rejecting* the null hypothesis when the null hypothesis is indeed *false*. Hence, if we have rejected the null hypothesis that "There is no difference in the true average scores of two groups of learners taught according to methods A and B" and concluded that "There is a difference in the true average scores of two groups of learners taught according to methods A and B", the conclusion we reached could be wrong. We *could* have committed a Type 1 error; that is, *rejecting* the null hypothesis that "There is *no* difference in the true average scores of two groups of learners taught according to methods A and B" when, in fact, there *really* is no difference in the true average scores of two groups of learners taught according to methods A and B. On the other hand, if our data *did not* allow us to reject the null hypothesis, we could have made a Type 2 error; that is, *not rejecting* the null

hypothesis (that there is *no* difference in the true average scores of two groups of learners when taught under methods A and B) when we should have done so—that is, when in fact, there *is* a difference.

In reaching our conclusions we have no way of knowing if our conclusions are right or wrong as the "truth" lies in the population, which we are unable to obtain, or measure. Thus, in statistical inference we can never be *certain* as to the truth or falsity of our conclusions; only the probability that they may be wrong or right. In hypothesis testing, the two probabilities of error in our conclusions are *alpha* (the *probability* of *incorrectly* rejecting the null hypothesis) and *beta* (the probability of *not* rejecting the null hypothesis when the null hypothesis is indeed *false*), while the probability of *correct* rejection (i.e., *rejecting* the null when the null is indeed *false*) is known as *power* In constructing confidence intervals, the probability of being correct (i.e., capturing the parameter) is given by the confidence level.

When making statistical inferences, we cannot eliminate the probability of error because we are dealing with a sample, not a population of scores. However, in conducting their studies, researchers can—and in fact, should—try to reduce the probability or likelihood of error in their estimates or conclusions (Brewer, 1991; 1996) In constructing confidence intervals, for instance, one could specify the probability of error in the estimate at the level that one wishes, and similarly, for hypothesis testing, one could set the probability of making Type 1 error (alpha) and the probability of making Type 2 error (beta) at the lowest level of error with which one is comfortable, but in order to do so, one would have to meet the minimally adequate sample size required to *justify* the levels that one had set for either statistical inference procedure (with all else held equal, the lower the probability of error set, the higher the minimum sample size needed). Meeting the minimum sample size requirements for constructing confidence intervals and hypothesis testing, however, is something that many researchers fail to do when conducting their research (Brewer, 1991; Cohen, 1965; 1988; Crookes, 1991)

The purpose of the preceding discussion is to highlight the logic underlying the use of statistics. When we use statistics for a descriptive purpose, we are using statistics to describe a particular group; not to generalise to other groups or situations. Hence there is no probability involved since the descriptive statistic is not used to infer to any population value. When we are using statistics for an inferential purpose, however, there will always be the *probability* of error involved, as we are using a sample of scores to

estimate or conjecture about the population value. The purpose for which statistics is put will thus have important implications on how we should conduct our study and on the words we could use to report our procedures and results. And this forms the basis of our discussions in the sections that follow

The problem with definition and meaning

One of the problems with meaning that is sometimes encountered in quantitative research reports has to do with the operational definition, which is a definition based on the observable characteristics of that which is being defined (Tuckman, 1988). Because many variables that are to be investigated are abstract, the first thing the researcher has to do before embarking on his or her study is to arrive at a clear, precise, and exact definition of these variables. He could, of course, provide a conceptual definition, which would identify the variable in terms of conceptual or hypothetical criteria. However, because empirical investigations deal with the observable world, conceptual definitions would not help in providing a means to investigate the variable or construct. Thus, for the purposes of investigation, concepts need to be defined operationally This involves identifying the specific behaviours recognised by current theories in the field as realising these variables (Seliger and Shohamy, 1990)—in other words, specifying how the variables could be "measured"; for it is these (measurable) behaviours that go towards the operational definition of these terms. For instance, if a researcher wishes to investigate the construct, language proficiency, he or she would, first of all, have to operationally define the term. Thus, he or she may, for example, decide to define English language proficiency as "getting between 80% to 100% on a particular teacher-made grammar test". Another researcher, however, may wish to define English language proficiency as "getting a score of at least 550 on the TOEFL", while yet another may define it as "getting a grade of A1 or A2 on the English GCE O' Level paper"

The problem that sometimes arises with regard to operational definitions is that they do not always succeed in "capturing" the construct. In other words, the construct may have been operationally defined in such a manner that it does not correspond to the theoretical or even common-sense notion of what the contruct is or ought to be. For instance, the researcher may have defined the construct of language proficiency as "getting between 80% to 100% on a particular teacher-made test on certain grammar points", while the reader may conceive of language proficiency as encompassing a wider

range of behaviours, such as being able to speak the language with a certain degree of fluency, being able to write grammatically, and so on. Or a researcher, in studying motivation, may have operationally defined a motivated student as "one who manifests persistent school attendance", which may not correspond to our notions of what motivation is. In coming up with an operational definition aimed at bringing concepts and variables to a sufficiently concrete state for examination, researchers may provide operational definitions that are so exclusive, that is, that are so specific in terms of the behaviours that go towards the definition of the concept that it restricts its generalisability. Hence, if the exclusiveness of an operational definition is carried to an extreme, it would limit the usefulness of that definition to only that of the research situation. Because of this, the researcher should take particular care in coming up with an operational definition that corresponds as closely as possible to what would generally be considered as defining the construct because the meaningfulness of the concept or construct would depend on how adequate or appropriate the operational definition is. If the operational definition is so far from what we would consider to be a decent definition of a construct the reader may be sceptical of the definition, and hence, the results. Thus, it is also important for the reader to keep the operational definition in mind as he or she reads the results of research studies so that the results will not be misinterpreted. The researcher might have (operationally) defined the construct in a very restricted sense—for the particular purposes of the study And although the reader might have objected earlier to the operational definition given by the researcher, she might forget her objection to it by the time she reaches the "Results" section. Instead of restricting the meaning of the construct to that of the operational definition, she now understands the construct according to its conceptual definition, or according to the way the concept is usually understood. Hence, she may interpret the results differently from the way they should have been interpreted because of the restriction in the meaning of the construct as brought about by the operational definition. One of the things researchers could do to reduce the likelihood of miscommunication would be to mention, once again, the operational definition as they go into a discussion of the results. In this way, the reader would be reminded of the definition, and would have the same conception of the construct as the one held by the researcher, and would take that meaning into perspective when evaluating and interpreting the results of the study

The misuse of terms in research reports

Random sample or representative sample?

One of the terms used in quantitative research that is sometimes open to a lot of interpretation is the term "random" When a researcher states, for example, "The corpus for this paper consists of thirty articles selected at random from six different language teaching journals published from 1985 to 1991" or "A total of thirty abstracts were extracted from randomly selected theses..." what exactly does he or she mean? In what sense is the word "random" used by the researcher? Is he or she using it to mean, "lacking aim or method; without careful choice or plan; purposeless; haphazard" (Webster's New World Dictionary) the way the lay person would use the term, or is he or she using the term the way it is (meant to be) used in statistics? Unless the researcher has actually described in greater detail how he or she had collected the sample other than the fact that "it was randomly collected" or that it had been collected "at random", it would be difficult to determine the sense in which the term is used.

Far from implying "without careful choice" or "haphazardly", the way we usually use the term in our everyday discourse, as in, "I picked at random the students that I would like to take part in the march", randomisation as used in statistics requires careful attention to the method of selection; as a random sample is defined by the method by which it was procured (Brewer, 1991, Glass and Hopkins, 1984; Hays, 1994) The idea of randomness is based on the concept of simple random sampling, which is defined as "a method of drawing samples such that each and every distinct sample of the same size N has exactly the same probability of being selected for the sample" (Hays, 1994; p. 53). The consequence of this definition is that each observation (score) within the sample also has an equal chance of being selected from the population of interest (Brewer, 1991).

When using statistics for inferential purposes, it is important that researchers collect a random sample because randomisation is essential in judging the *validity* of the inferences made from the sample to the population; i.e., the generalizability of the results (Glass and Hopkins, 1984). This is because randomisation "will ensure, *within a certain known margin of error*, representativeness of the samples" (Glass and Hopkins, 1984, p. 177) Indeed, randomisation is a fundamental assumption underlying statistical inference (Brewer, 1991; Glass and Hopkins, 1984; Hays, 1994; Shaver, 1993), the violation of which may totally invalidate any study (Hays, 1994). As Shaver (1993) argues, "without randomness, the result of the test of statistical sig-

nificance is meaningless or, at best, its relevance to a statement of probability is indeterminate" (pg. 299)

Thus, it would be helpful to the reader if the researcher could describe in greater detail how the sample had been selected; because only by knowing how the researcher had selected his or her sample would the reader be able to determine whether this fundamental assumption of statistical inference has been met. It would certainly enable the critical or informed reader to have greater confidence in the results. Unfortunately, however, many studies using statistical inference do not carry out randomisation (Henning, 1986; Shaver, 1993); nor do they even describe how the sample was collected (Henning, 1986).

While failure to describe in greater detail how one's "random sample" had been selected leaves the reader wondering as to whether it is, in fact, a random sample, a term that is erroneously regarded by some researchers as being synonymous to the random sample is the term representative sample. In fact, it is also one of the most frequently misused terms in the research literature (Brewer, 1991; Hays, 1994; Shaver, 1993). The following statement, taken from a research manual for researchers in applied linguistics, misleads the reader as to a requirement of statistical inference: "When we want to generalize from our sample to the population, we must be *certain* that the sample is truly *representative*" (Hatch and Lazarton, 1991, p. 234). Similarly, Shaver (1993) has also found it baffling that Thompson (1987) would assert that "significance testing imposes a restriction that samples must be representative of a population, but does not mandate that this end must be realized through random sampling" (p. 299) What is misleading about Hatch and Lazarton's statement—and Thomson's as well—is that statistical inference does not require that one be certain that the sample is *truly representative*. In order for us to do so, we would have to ensure that the sample characteristics match *exactly* the population characteristics on the variables of interest—which would mean that we would have to know *exactly* what those population characteristics are. And if we would like to know exactly what those population characteristics are, then we would have to have the entire population of scores. And if that could be done, then there would be no need for statistical inference for the purpose of statistical inference is to generalise, from a limited sample of scores, or from sample characteristics, the entire population of scores, or population characteristics which we are unable, for some practical reasons, to obtain. It should be pointed out at this juncture that what is required in order to do statistical inference is randomisation. This is because randomisation "will ensure, *within a certain*

known margin of error, representativeness of the samples" (Glass and Hopkins, 1984, p. 177). In other words, random sampling addresses solely the representativeness of samples in the long run; it provides *no guarantee* that the sample is representative of the population (Brewer, 1991), that is, "that all of the characteristics of a particular [random] sample, including the dependent variable(s) under investigation will be the same as those of the population, only that...they will differ only by chance from the population characteristics" (Shaver, 1993; p. 296). Thus, contrary to what Hatch and Lazarton assert, we need not be *certain* that our sample is representative of the population when we want to make generalisations to that population. Certainty cannot be assured even with a random sample. It would therefore be misleading to claim that what we have is a "representative sample" It is a random, not a "representative" sample that we should obtain in order that we may justifiably do statistical inference.

Certainty or Confidence?

In the research literature, one would occasionally come across statements that reflect certain misconceptions regarding statistical inference. The following statement (in addition to the one that was discussed in the previous section) reflects this misconception. "The most powerful [statistical] test allows us to be sure that when we reject the null hypothesis we are correct or that when we accept the null hypothesis we are correct" (Hatch and Lazarton, 1991; p. 239).

Contrary to what the authors claim, even the most powerful statistical test will *never* allow us to be sure that we have correctly rejected or incorrectly rejected the null hypothesis. As we have previously mentioned, just as there is the *probability* of correctly rejecting the null hypothesis (power), when doing hypothesis testing, there is also the *probability* of incorrectly rejecting the null hypothesis (that is, the *probability* of rejecting the null hypothesis when the null hypothesis is indeed true) Because hypothesis testing is a statistical inference technique, probability will always be involved. Thus, we could never be *certain* as to whether we had correctly rejected or incorrectly rejected the null hypothesis.

Even if we utilise confidence intervals to estimate a true parameter as in the section on "The Uses of Statistics", there is always a probability (1 - confidence) that intervals like the ones produced from the researcher's sample will not capture the parameter of interest. To illustrate, consider the statement: "The researcher has 95% confidence that if all the English teach-

ers in Selangor were to be surveyed, between 65% to 75% of them would want to see greater emphasis given to grammar in the teaching of English in secondary schools" What this statement means is that the researcher's sample provided a sample percentage of 70% and a subsequent interval (65%, 75%) which estimates the true percentage of teachers who feel this way about grammar The true percentage may or may not be in this particular interval, but it is a good bet that the true percentage is in this interval since the odds of capturing the true percentage before any sample was taken were 19 to 1 in favour of capture. There is still a 1 in 20 chance that any set of intervals (including this one) selected at random will not contain the true percentage.

Significance or Importance?

Another interesting example of how ordinary English words have taken an entirely different meaning in statistics is the use of the terms, *significance* and *importance*. In fact, most readers of quantitative research reports—and some researchers as well—find it difficult to distinguish between the two terms. While many may be somewhat familiar with the use of the term *significance* in the context of statistics, not as many are familiar with the use of the term *importance*. In the statistical context, significance has to do with rejecting the null hypothesis. To put it very simply, the term "significant", when used in the statistical context only means that the null hypothesis was rejected. Thus, when one has rejected the null hypothesis, one might say "the findings are (statistically) significant" or "the findings have reached significance" It would make no sense, however, to make statements such as, "the difference in morphological errors is only weakly significant", that "the results are approaching significance" or that "the difference in syntactical errors is highly significant"—although these statements are widely found in the literature. This is because significance has to do with rejecting the null hypothesis at an alpha level (probability of making Type 1 error) set by the researcher prior to the study, on which the minimum sample size was partly based.

Statistical significance is thus a statement about the likelihood of the observed result, given the null hypothesis is true; nothing else (Brewer, 1991; Hays, 1994) It does not mean that something important, or valuable, or meaningful has been found. However, many researchers often consider results that are statistically significant to be important or meaningful as well. But what then, is significance if not importance, and importance if not significance?

While statistical significance has to do with rejecting the null hypothesis, importance has to do with whether observed values or differences are of any practical value. Let us consider the following situation. Suppose a researcher is interested in testing the (null) hypothesis that the true average scores of two groups of learners, taught by two different methods, do *not* differ on a test of language proficiency versus the alternate hypothesis, that the true average scores do differ Importance of the findings has to do with whether or not the difference in the observed mean scores—that is, the difference in the mean score as calculated on Group A and the mean score as calculated on Group B (should there be a difference in the *observed* mean scores) is viewed as meaningful, worthwhile or important to the researcher. This observed difference is called "post hoc effect size", meaning that it was *sample* differences observed after the data was collected. This is to be distinguished from "a priori effect size" (before data are collected), which is the researcher's judgement or expectation of what the smallest true difference should be in order for it to be called important. The latter effect size (a priori) is one of the criteria for determining minimal sample size. (For a fuller discussion of effect size, see Brewer, 1991, 1996; Cohen, 1988). For example, the researcher might have stated, prior to collecting any data, that a difference of at least 20 points between the true mean scores of the two groups would be important or meaningful. After the data had been analysed, he found that he was able to reject the null hypothesis, and thus had statistically significant results. He also found a difference of 6 points in the sample mean scores between the two groups. Consequently, he decides that his results are not (practically) important because the difference in the sample mean scores was not close enough to the population expectation of what an important true mean difference would be, i.e. 20 points. In other words, the difference in observed means scores is not large enough—as judged by the researcher's standards—to be considered important. Thus, he got *statistically* significant findings (and was thus able to conclude that performance on a language proficiency test—as judged by mean scores—of two group of learners taught by two different methods *does* differ) but that this difference is *trivial* (that is, the difference in mean scores of the two groups is too small to be considered important) However, some researchers and readers have the mistaken notion that if a result is statistically significant, then it must be practically important as well (Bracey, 1991, Brewer, 1991; 1996; Shaver, 1993) It has to be remembered that in the statistical context, significance and importance are two entirely different concepts. Significance has to do with rejecting the null hypothesis, while importance has to do with whether the difference in the *observed* mean scores (using this particular example) is large enough as judged by the criterion set by the researcher prior to the study The reader will notice that the researcher has to decide if what was observed

in the sample is to be viewed as trivial or not. Not every situation is as clear-cut as the above example.

To prove or to provide evidence?

On reading the journals in the field of linguistics or applied linguistics, one might occasionally come across a statement such as the one that has been italicised below:

> Although the apparent relationship between type of theoretical orientation, years teaching ESL, and the prominence of a particular methodological approach is based solely on descriptive data, one might speculate that the sources of ESL teachers' theoretical beliefs may stem from the methodological approaches that were prominent when they began teaching ESL. *If this speculation can be proven through future empirical research, it may have important implications for second language teacher education programs* (Johnson, 1993).

Some questions that one might raise with regard to statements such as the one above are: Can one "prove" a speculation through empirical research? Indeed, can empirical research ever "prove" anything to be true? If it cannot, what are we allowed to say as a result of our empirical investigations?

In the social and behavioural sciences, and in fact, in the language sciences as well, the focus of inquiry seems to be on generalising from a particular set of observations to all the potential observations that might be made under the same conditions. In other words, we are interested in "going from what is true of *some* observations to a statement that this is true for *all possible* observations made under the same conditions" (Hays, 1994; p. 4). However, does this mean that the outcome of statistical inference allows us to state, with absolute certainty, that our conclusion is indeed true? How do we know that the results have not have been reached in error, or are the product of chance variation in conditions over which we have no control? Would we get similar results if our study were repeated over and over again? Our common-sense reaction might be to say that we couldn't be *certain* that conclusions reached as a result of statistical inference reflect the state of truth. And yet this is precisely what is implied when we state that something could be proven through empirical research.

In the physical sciences like Physics and Chemistry, the making of general statements about physical phenomena from observations of limited numbers

of events is not an issue because it is usually possible to exercise precise experimental control to remove a substantial amount of variation among observations. However, in the social sciences, the sources of variability among human beings are extremely difficult to identify and measure; let alone be subject to precise experimental controls. Hence, the drawing of conclusions in the social sciences involves a great deal of uncertainty Hays puts it very succinctly:

> Faced with only a limited number of observations or with an experiment that can be conducted only once, the scientist can reach conclusions only in the form of a "bet" about what the true long-run situation actually is like. Given only sample evidence, the scientist is always unsure of the "goodness" of any assertion made about the true state of affairs. The theory of statistics provides ways to assess this uncertainty and calculate the probability of being wrong in deciding a particular way. Provided that the experimenter can make some assumptions about what is true, the deductive theory of statistics tells us *how likely* particular results should be...Regardless of what one decides from evidence, it *could* be wrong; but deductive statistical theory can at least determine the probabilities of error in a particular decision. (Hays, 1994; p. 4)

Thus, the conclusion that one makes as a result of statistical inference cannot be stated in such a manner that would exclude the probability of error, because it is possible that the conclusion could be wrong. Hence, a researcher is allowed to say, "There is evidence to *conclude* that learners taught according to method A perform better than learners taught according to method B", because he or she is saying that the conclusion is made on the basis of available evidence, which does not preclude the possibility that it could be wrong. However, statistical logic would not allow him or her to say, "This *proves* that students taught according to Method A perform better than students taught according to method B" Unlike mathematical theorems, which can be proven by logical deduction, conclusions resulting from statistical inference cannot be used as a "proof" of the "correctness" or "wrongness" of something, or the "superiority" of one method over another because statistical inference involves the probability of error In fact, we consider that even to use the term "shows" when stating a conclusion, as in, "This research *shows* that students taught according to Method A perform better than students taught according to Method B" would be stating the conclusion too strongly or too definitely, while Eskey (1987) goes even further to say that it would be dishonest to make such a claim.

Thus, when stating our conclusions or reporting the conclusions of others, we have to be very careful not to state them as if they were facts, as in the

following example, in which Pica (1985) reports the conclusions of three separate studies conducted by Larsen-Freeman (1975, 1976a, 1976b): "It *has been shown* that the factor most critical to production accuracy is not a morpheme's linguistic complexity, but rather the frequency with which it occurs in the input that the learner receives" (p. 214) Although Larsen-Freeman might have indeed concluded that such is the case, again, it is possible that the conclusion could be wrong and that other factors or chance might, in fact, be responsible for the observed results. Besides, reporting the conclusions of the three studies of one researcher does not justify using the expression, "it has been shown", which would imply that this conclusion has consistently (if not always) been reached in the studies that looked at the same phenomenon. It has to be realised that for a particular conclusion that has been reached, there are other plausible explanations. Thus, it is common to find statements such as the following in our research literature: "It is premature to address the question of what aspect of SLA is influenced by cognitive style. The existing research does not *conclusively show* that it is a major factor where success is concerned" (Ellis, 1990) The question is, given the many uncertainties in the field of the social and behavioural sciences—of which the language sciences are a part—should we even expect research to "conclusively show" a certain result or happening?

One of the reasons why researchers sometimes state their conclusions as if they were facts, as shown by their use of the expressions, "prove" or "research has shown" may be that they believe that the procedures involved in quantitative research are indeed able to do that. This misconception is perpetuated by some textbooks, as illustrated by the following statement, taken from a textbook on communication. "A hypothesis is a statement to be proved or disproved by research" (Treece, 1989; p. 362). There also seems to be a general feeling, among some researchers, that experimental research is able to achieve this, as reflected in the words of this author:

> Experimental research is highly valued in the social sciences because it can establish cause-and-effect relationships. To test their hypotheses, researchers divide the environment into treatment and control groups, administer treatments, and assess the results with measurement instruments that, it is hoped, are valid and reliable....Then, if the treatment groups perform significantly better, the treatment is said to have caused the difference....The strength for the validity of the claim for cause and effect still lies on the validity of the measurements used. (Connor, 1987; p. 11).

Even if the measurements are valid, and even if the treatment groups perform significantly better, one would be hard pressed to argue for a cause-and-effect

relationship, because in hypothesis testing, which involves statistical inference, there always exists the probability of error. While the probabilities of error for a particular study could be set by the researcher *a priori* and consequently justified by the meeting of a minimally adequate sample size, a problem in many quantitative studies is that they are not (Brewer, 1991; Cohen, 1988; Crookes, 1991) While some researchers specify the probability of Type 1 error (alpha) for their studies at the conventional .05—although the setting of alpha is actually a subjective judgement on the part of the researcher (Brewer, 1991)—many do not. Very few, if any, specify the probability of Type 2 error (Crookes, 1991), which would consequently give us the power of the test. Thus, we have many research situations in which the probabilities of error are unknown, since the researcher has not specified them and since no justification is made for the sample size used. Indeed, many researchers base their sample size on the belief that "30 is sufficient" (Crookes, 1991) With studies having such small sample sizes, we may have situations in which the power of the test (that is, the probability that the statistical test and tests like that one would *correctly* reject the null hypothesis) is, in fact, very low and in which the probabilities of error would be high (Brewer, 1991) Added to this is the fact that most studies are not able to take into account or control for all the relevant variables that might affect the outcome of the study Given all these circumstances, it would be very difficult to argue for the conclusiveness of the results of (even) experimental studies. Thus, using the terms "prove" or "it has been shown" in reporting conclusions that are a result of statistical inference is totally inappropriate—and in fact, misleading, because it implies that we are absolutely certain as to the truth of our conclusions. If the use of these terms is inappropriate, how then, should we report our results?

Reporting Our Results: Saying What We Mean and Meaning What We Say

When we are making generalisations or reporting conclusions that are the result of statistical inference, it is important that we not state them as if they were facts, because there is always the possibility that the conclusion could be wrong. They should, instead, be stated in "soft" terms as shown in the following example: "The results of the present study *offer evidence* that conscious attention to form in the input competes with conscious attention to meaning, and, by extension, that only when input is easily understood can learners attend to form as part of the intake process" (Van Patten, 1989). By stating that the results of the study offer evidence in support of the stated conclusion, the researcher is not making any claim that there is absolute

certainty in the "truth" of the conclusion, but rather, that as evidence (for which there would also be counter-evidence), they would be used as pieces that would go towards the final "picture" It would also be appropriate to state one's conclusions in the following manner· 1) "In my own research, both input and L1 factors *appeared* to contribute to the patterns of emergence, development, and—particularly—to the fossilisation of particular forms" (Lightbown, 1985b). 2) "The results in this article *provide further support* for the hypothesis that form-based instruction within a communicative context contributes to higher levels of linguistic knowledge and performance." (Lightbown and Spada, 1990) 3) "The results of the study *suggest* that classroom instruction has a distinct impact on the acquisition and production of a second language (Pica, 1985) By writing their conclusions in this manner, the researchers imply that it would be possible to arrive at conclusions other than the ones they had reached, and are thus not making claims as to the infallibility of their conclusions.

Conclusion

The importance of being precise in our choice of words when communicating the procedures, results, and conclusions of our quantitative research findings cannot be overemphasised. This is especially so because many of these words, which we are so accustomed to using in the everyday context, have taken a more specific meaning when used in statistics. If we continue to use these terms in the unrestricted, everyday sense, it is likely to lead to confusion in meaning. This will have implications not only on how readers would interpret the findings, but also on the extent to which other researchers would be able to replicate our study if they so wished.

The question of how research findings are interpreted by the reader is a very important matter indeed. As Lightbown (1985a) and Tarone, Swain, and Fathman (1976) have noted, there are great expectations on the part of some teachers and researchers to apply the results of language acquisition research to the classroom. Although one should be cautious about making specific recommendations about language teaching on the basis of research in language acquisition (Lightbown, 1985; Tarone, Swain and Fathman, 1976), nevertheless, these recommendations continue to be made. Lightbown has noted that many of these recommendations have been premature, based on research that was extremely narrow in scope, and based on overinterpretation of data. It should be stressed that research in linguistics and applied linguistics is still at its infancy, and is not without its limitations. Among some of the limitations associated with the research are that the methodology used

in the collection and analysis of data are still in a developmental state, and that few of the studies have been replicated by researchers (Tarone, Swain, and Fathman, 1976) For those studies that utilise statistical inference, many do not meet the fundamental requirements of the procedure such as randomisation and a minimally adequate sample size. In addition to this, one has also to ask to what extent the studies are properly designed, and to what extent they have been able to account for the other variables that might affect the outcome of the study Because of all these factors, together with the fact that the nature of social science research (especially those that utilitise statistical inference) is such that any result obtained will never be free from the probability of error, the conclusions that we reach should be considered tentative at best. Hence, results of available research should be considered suggestive rather than definitive, and care should be taken by researchers not to communicate anything to the contrary. It is of utmost importance that research findings are critically evaluated and properly interpreted so that the conclusions reached are meaningful in light of the findings. These issues should be noted especially by those who review and interpret research for teachers and syllabus designers, for such reports have the potential to influence policy and curricular decisions as well as classroom practices. It would be unfortunate indeed if entire curricula or a large bulk of pedagogical practices were based on findings and conclusions that have been communicated as "facts", for these conclusions might not, in fact, reflect the true state of affairs. As researchers and writers, it is of utmost importance that we communicate research findings and conclusions clearly and precisely, and not discount the possibility of arriving at other conclusions. It should be borne in mind that research conclusions are subject to logical pitfalls and errors, which form part of any human endeavour

References

Akmajian, A., Demers, R. A., Farmer, A. K., & Harnish, R. M. (1990). *Linguistics: An introduction to language and communication* (3rd ed.) Massachusetts: The MIT Press.

Bakan, D. (1967) On method: *Toward a reconstruction of psychological investigation.* San Francisco: Jossey-Bass.

Brewer, J. K. (1991). *Introductory statistics for researchers* (5th ed.) Minnesota. Burgess International Group, Inc.

Brewer, J K. (1996) *Everything you always wanted to know about statistics, but didn't know how to ask* (2nd ed.) Dubuque, IO· Kendall/Hunt Publishing Company.

Brown, J. D. (1991) Statistics as a foreign language—Part 1. What to look for in reading statistical language studies. *TESOL Quarterly,* 25(4), 569-585.

Celce-Murcia, M., & McIntosh, L. (Eds.). (1979). Teaching English as a second or foreign language. Rowley, MA: Newbury House.

Carver, R. P (1978) The case against statistical significance testing. *Harvard Educational Review,* 48(3), 378-399.

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 (1988) *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ· Erlbaum.

Crookes, G. (1991) The forum: Research issues. *TESOL Quarterly,* 25(4), 762- 765.

Dunkel, (1986) Review of statistics in linguistics. *TESOL Quarterly,* 20, 549-542.

Ellis, R. (1985) *Understanding second language acquisition.* Oxford: Oxford University Press.

Eskey, D E. (1987) Comments on Devine. In Devine, J., Carrell, P L., & Eskey, D. E. (Eds.), *Research in reading in English as a second language* (pp. 86-87) Washington, D. C: Teachers of English to Speakers of Other Languages.

Flynn, (1985). Review of research design and statistics for applied linguistics. *TESOL Quarterly,* 19, 155-158.

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

Hatch, E., & Lazarton, A. (1991) *The research manual. Design and statistics for applied linguistics.* Los Angeles: Newbury House Publishers.

Hays, W L. (1994) Statistics (5th ed.). Fort Worth: Harcourt Brace College Publishers.

Henning, G (1986) Quantitative methods in language acquisition research. *TESOL Quarterly,* 20(4), 549-552.

Johnson, K. (1993) The relationship between teachers' beliefs and practices during literacy instruction for non-native speakers of English. *Journal of Reading Behavior,* 24 (1), 83-107

Larsen-Freeman, D. (1975) The acquisition of grammatical morphemes by adult ESL learners. *TESOL Quarterly,* 9(4), 409-419

Larsen-Freeman, D (1976a) A explanation for the morpheme accuracy order of learners of English as a second language. *Language Learning.* 26(1), 125-135.

Larsen-Freeman, D. (1976b). ESL teacher speech as input to the ESL learner *UCLA Workpapers in Teaching English as a Second Language,* 10, 45-49

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

Lightbown, P (1985a). Great expectations: Second-language acquisition research and classroom teaching. *Applied Linguistics,* 6 (2), 173-

Lightbown, P (1985b) Input and acquisition for second-language learners in and out of classrooms. *Applied Linguistics,* 6 (3), 263-274.

Lightbown, P., & Spada, N (1990) Focus-on-form and corrective feedback in communicative language teaching. *Studies in Second Language Acquisition,* 12, 429-448.

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

Pica, T (1985). The selective impact of classroom instruction on second-language acquisition. *Applied Linguistics,* 11, 214-223

Savory, T.H. (1953) *The language of science.* London. Andre Deutsch Limited.

Seliger, H. W., & Shohamy, E. (1990) *Second Language Research Methods.* Oxford: Oxford University Press.

Shaver, J.P (1993) What statistical significance testing is, and what it is not. *Journal of Experimental Education,* 61(4), 293-316.

Smith, B.O., & Ennis, R. H. (1961) *Language and concepts in education.* Chicago: Rand McNally & Company

Tarone, E., Swain, M., & Fathman, A. (1976). Some limitations to the classroom applications of current second language acquisition research. *TESOL Quarterly,* 10(1) 19-33

Treece, M. (1989) *Communication for business and the professions* (4th ed.) Boston: Allyn and Bacon.

Tuckman, B. (1988) *Conducting educational research* (3rd ed.) New York: Harcourt Brace Jovanovich.

Van Patten, B. (1990) Attending to form and content in the input. *Studies in Second Language Acquisition, 12*(3), 287-301