Statistics and Research Design: Essential Concepts for Working Teachers

By Ivan Lowe (Tunisia)

The careful interpretation of numbers is an integral part of being a teacher. Every time a teacher marks an examination and students interpret the marks, they are using statistics. Teachers need to know that words such as *average* have several meanings and that authors can slant their argument by using the definition that best suits them. Unfortunately, introductory textbooks of statistics usually start at too high a level and neglect to teach the underlying concepts.

This article presents 10 keys for teachers seeking to understand statistical reasoning and research design. The first 6 keys cover basic definitions in statistics. Keys 7 to 10 help teachers grasp the principles underlying the design and interpretation of research. None of them requires more than simple arithmetic.

Key 1. Establish definitions and circumstances

The most important key to using statistics is understanding the definitions of what is being measured. A memorable example of how a change of definitions and circumstances can affect numbers happened after the Second World War in England, when the divorce rate went up. I ask students to suggest reasons for the increase and they usually comment about men coming back from the war after years of separation from their wives. That long separation may well have been a contributing factor (a *variable*); but actually the law had also been changed with the effect that divorce was somewhat easier to obtain. Therefore, prewar and postwar divorce rates are not directly comparable.

A document with statistics may not clearly state crucial definitions or may use definitions that are different from what is generally expected. For example, in a census, the average age of death could be calculated using the age of death of those who reached adulthood. Adulthood would require a specific definition, such as people who have reached the age of 18. Or the calculation could include children. Obviously, an average age of death calculated the second way would be much lower. Such crucial definitions can vary among authors or different editions of a publication.

Another important point about definitions concerns formulas. When a formula is given, the meaning of each of the letters used must be stated because conventions change or may not be widely known.

Key 2. Understand the meaning of average

There are three kinds of average: the *mean*, the *median*, and the *mode*. All three can legitimately be called an average. They are often similar, but they are not always equal. With this series of numbers: 7, 7, 7, 4, 4, 3, 3, the mean is 5 (35÷7), the median is 4 (the middle number when the

series is in incremental order), and the mode is 7 (the most frequently occurring number in the series).

An average can be chosen to suit a specific purpose. In a survey of salaries, the average level of pay can be distorted upward by choosing the mean if there are a few very high wage earners. In this example, the mode (the wage earned by the largest subgroup of workers) would be a fairer choice. As Woods, Fletcher, and Hughes (1986:32) point out, the median is particularly useful when there are a few atypical examples at one of the extremes of the range of values, since it is relatively unaffected by them. A large difference between the median and the mean is also a clear indication that there is not a symmetrical distribution of scores on either side of the mean and the data may be skewed.

In the TEFL context I work in, there is interference from French in my students' understanding of statistics. I often hear them say, "I did not get the average." That statement is incorrect, not because of the grammar but because of the meaning of average. The rough equivalent of average in French, *la moyenne*, includes all three averages above, with the addition of what is often called in English the *passmark*, the minimum passing score. In Tunisia, the passmark is almost always fixed in advance at 10 out of 20, so a student with 9/20 has failed to get the *moyenne*. The idea that the passmark can vary according to the subject, or that the passmark can be fixed after exams have been marked, as is commonly done in other countries, would appear to students and teachers alike as a form of manipulation of the exam results. They assume that exam marks are objective, comparable, and reliable, which brings us to the third key.

Key 3. Avoid pseudo-precision with numbers

Huff astutely pointed out years ago, "[a] difference is only a difference if it makes a difference" (1954:56). Superficially, numbers look definite and absolute and give a deceptive impression of accuracy. We can reduce deception if we understand the following concepts.

Know the difference between a decimal place and a significant figure

Because of the limits of accuracy in measuring techniques, scientists usually measure in *significant figures* (s.f.). The concept is easily illustrated by considering the length of a piece of furniture such as a table. With a normal tape measure, an accuracy to three significant figures is easy to obtain, that is, to the nearest centimetre (for example, 130 cm). But to measure to the nearest millimetre requires an accuracy of four significant figures, for example 1301 mm, which is not as easy to achieve as it sounds. *Decimal places* (d.p.) are the number of figures beyond the decimal point. If the table were measured in centimetres to three s.f., the measurement would have no decimal places (130 cm). If the same table were measured to four s.f., the measurement would have one decimal place, that is, 130.1 cm. Usually, providing significant figures is more important than simply providing decimal places. There must be no more stated precision than what the measuring tool can reasonably provide, even if longer sequences of numbers are actually generated from calculations.

Round correctly

Pseudo-precision (Matthews, Bowen, and Matthews 1996) can be reduced by rounding properly. For instance, 11256 divided by 460 equals 24.4695652173913043478…. This number (a *quotient*) can be rounded to four significant figures, or to two decimal places, at 24.47. Rounded

to only two significant figures, it is 24. If it were rounded first to 24.5 and then up to 25, it would be incorrect; it should be rounded down to 24 instead. Whenever the last digit is a 5, the missing digit to its right needs to be identified to know whether rounding should go up or down. When calculated numbers are combined, the rounding should happen only once; otherwise errors can be multiplied.

Certain important points of style must not be forgotten. While it is true that whole numbers do not need the decimal point followed by a zero, if a table of figures lists 14.1, 15.3, 17.4, and 18.1, then 17 would be listed as 17.0 (not 17) to follow the pattern. This shows that the number genuinely has one decimal place and is not a number rounded to zero decimal places.

Use percentages wisely

The larger the sample size, the more reasonable it is to present results as a percentage. However, the sample size (or *population*) must be specified. This is especially important when the sample size or number of subjects is less than 50 (Matthews, Bowen, and Matthews 1996). Both a percentage and the number of subjects can be quoted. This rule exists for a good reason: whenever a percentage is calculated based on small samples, small differences become magnified. Using percentages increases the likelihood that a small difference will be interpreted as a significant one.

Let's take an example in which the examination pass rates of two classes are compared. In class A, 55 percent of the students passed, and in class B, 60 percent passed. The difference of 5 percent looks invitingly significant. Can we conclude that class B is better than class A? Not on the basis of the information provided, because the size of each class was not stated. If class A had 20 students and class B had only 10 students, then 11/20 (55%) of class A students passed and 6/10 (60%) of B passed. The possible variables contributing to this outcome are numerous and the numbers involved are small; the 5 percentage point difference is due to only one student in class A. Even if only 9/20 (45%) of class A passed, we still could not conclude that B is a better class.

Key 4. Go beyond stating the mean

How can 20°C in Cairo in the winter be considered cold, whereas in England in the summer it is warm? Ignoring the differing psychological attitudes to temperature in various climates, there is a problem inherent in the way the data in world temperature charts is presented. There is also the question of the number of hours during which the temperature stays at 20°C, which in Cairo in winter is likely to be only two or three hours, whereas the same English summer temperature is likely to last most of the day. This example illustrates the need to quote the *range of the mean*, as is done in some forecasts which state the minimum and maximum expected daytime and nighttime temperatures.

A mean without an accompanying indication of the range of values is almost always almost meaningless. Statisticians have provided several ways of expressing that range, which include the standard deviation and the standard error of the mean (well explained in Rowntree 1981). Other, simpler ways to indicate the range of the mean include stating the mean with a simple plus

or minus and a value. This is useful where the absolute minimum and the absolute maximum are known.

Another way is to state the range of values that would cover 90 percent of the sample, the so-called percentile lines. These are commonly seen on growth charts of the expected weights and heights for boys and girls in their first five years. I have found that most of my students know about these charts, and therefore they are a good starting point for discussing ranges of the mean.

Key 5. Anticipate false positives and false negatives

These two concepts may be familiar to students who know about diagnostic testing for AIDS or screening tests for cancer. When testing for disease, a result that is positive when it should not be is alarming, but rarely fatal, because people who test positive are usually retested to either confirm the initial result or expose a false positive. On the other hand, a false negative could allow a disease to go undetected for a long time, and could mean the difference between early and late treatment, and even between more years of life or an early death. Disease screening programmes seek to reduce both types of error but especially false negatives, rather than the false positives, since it is the false negatives that have potentially fatal consequences.

It is a challenge to encourage students to apply these concepts to their own situations and studies. In language testing, it is important to ask which error is more significant: the false positive or the false negative. Is it better to let pass those who should fail (the false positives) or to fail those who should pass (the false negatives)? The answer requires us to apply key 1, because the circumstances of the test are critical. Usually, teachers try not to fail students without good reason, so they work against false negatives.

But sometimes it is more important to fail a large number than to accept anyone who is lacking in competence. Take, for example, examinations that are set in two sessions, an arrangement allowing those who fail the first attempt to sit for the exam again. The concern in the first session is to minimize the number of students who undeservedly pass. The worst that can happen with a false negative in this scenario is that a student has the pain of resitting. In the second session, however, the concern is the reverse: there must be no false negatives.

Since the burden of proof is on researchers to show that their results are significant, most experiments are designed to prevent false positives. Once students master the concepts of false positives and false negatives, they will begin to grasp one of the most difficult areas of basic statistics: significance levels. A significance of $p = 0.05$ means that the possibility of a false positive is 1 in 20 (or 5%) or less.

Key 6. Tabulate students' marks using simple procedures

A common procedure for a teacher is collating and inspecting marks of students, sometimes comparing several classes that have been taught the same curriculum by two or more teachers. To make this comparison effective, I build a frequency table. In Tunisia we mark based on a maximum score of 20 points, which gives 20 possible marks. I begin by writing 1 to 20 in a column down the page. I count the number of students who received each mark and place a tick in each corresponding row of the column of 20 possible scores—a visual frequency table of the

number of students who got each mark. On most exams I would expect to see a peak somewhere in the middle, though it may vary.

After tabulating students' individual results, I use the frequency table to calculate the class mean. This is easily computed because I already know how many students got each mark. I also count the number of students who passed outright and the number who got 8/20 or more, because I expect around two-thirds of my students to score 8 or higher. I can then take all three measures and compare classes, looking for anything unusual. Where double-marking is not practiced, these quick calculations and inspections are particularly important in checking for errors in marking or unexpected results.

Key 7. Select controls and identify variables carefully

One of the most common design procedures in educational research is to take two groups, do a pretest on both, expose one group to an experimental procedure, and then retest both groups to see whether there is a difference which can be attributed to the experimental procedure. The group which had no extra experimental procedure is the *control*. The concept and use of controls lie at the heart of the methodology of systematic investigation, but they are not always easy to understand or apply.

One way to explain controls is to use an example from bacteriology, where lab tests are done by placing substances (which may or may not contain bacteria) in a growth liquid and then observing what organisms develop. The problem is to ensure that the liquid used to cultivate the bacteria is sterile. When tests are done, there is always the possibility that the test tubes used were never completely sterilized and any growth could be due to contamination and not to the bacteria being studied. As a control for this, usually one test tube without the substance is placed in the warm incubator, along with the test tubes that contain the substance. If preparation was imperfect and unsterile, then all the test tubes will be contaminated, including the control test tube. Then the researchers will know they must repeat the test. What is being controlled for is the presence of contamination. The control test tube becomes the standard for comparison against the other test tubes used in the experiment. At the design stage, any researcher must consider how to control for and prevent "contamination" of the results so that only the effect being studied is responsible for the measured result.

Controls are also important to measure the full range of variables that are naturally possible before an experiment took place. A basic prerequisite of research in most experiments is that when two groups are compared, they should be identical in every way except for one feature, which is the experimental (or *dependent*) variable. The most common variables in educational research are age, gender, and the social and economic background of the participants (also called *subjects* or *population*) in the study. Cohen and Manion (1985) have a chapter on controls with helpful examples.

This natural variation, which could interfere with the results, needs to be identified and managed. A variable can be made constant; for example, by restricting the population to females only. Otherwise the experimental and control groups should include equal proportions of each variable (Scholfield 1995). Every additional variable considered demands an increase in the population of

the study. In most cases, only one variable at a time should be changed. Brown (1992) has a very accessible discussion of variables.

In experiments, the measured result must not be attributable to natural variation within a population, nor should it be due to interference (contamination) from other factors. That is why all good experimentation involving human subjects should begin with a careful specification of the characteristics of the population. Ideally, a researcher should assign subjects to the control group or the experimental group randomly, or by using matched pairs. Unfortunately, both of these techniques for placing subjects in groups are often unrealistic in the educational settings where most language acquisition research takes place. Nevertheless, the experimental and control groups must be as similar as possible (Cohen and Manion 1985).

Finally, for the research results to be amenable to significance testing, each group needs to have at least 30 subjects. For instance, a random occurrence of 5 students having a sleepless night before an important exam early the next morning could be disastrous in a group of only 20 students but would have less impact on a group of 100 students taking the exam.

Key 8. Compare like with like

It is all too easy to compare the incomparable or to make an unfair comparison. The use of controls (key 7) is based upon fair comparison, along with an appraisal of all the variables. For years, the standard of comparison for an L2 speaker has been the monolingual native speaker of the L2. That comparison is now acknowledged as being unfair (Widdowson 1994; Wiley and Lukes 1996). A better comparison is the L2 speaker with other multilinguals, since L2 speakers are by definition multilingual.

Commonly we teach students who are doing their first piece of serious research to specify the question and then to ask what data could be collected to answer the question. Another approach is more time-consuming but rewarding. It is to study a situation in depth using methods from ethnography, and to obtain a thorough grasp of the population being studied first, before proceeding to study one question in greater detail. With or without a preliminary study, the variables, those factors that influence the data, must be identified and accounted for. That way like can be compared with like and the data collected can be interpreted in context.

Qualitative investigations are important in enabling fair comparisons to be made. They help ensure we understand the way variables are interrelated, interpret the experimental data in their context, and make valid comparisons and generalizations. Burgess's textbooks and case studies provide good examples of how to combine the qualitative with the quantitative (1982, 1983, 1984).

Key 9. Use the most demanding research design

One of the most neglected principles of research is that the design of an investigation should ensure that the data is collected in a way that works against the hypothesis. This principle is common in the physical and biological sciences, where I began my career.

Experimental procedures frequently sensitize people to the problem being studied. One example is from my research in chemistry terminology (Lowe 1992). I had already established that the morpheme order within names in organic chemistry was different in English and French. I wanted to see if students of chemistry in French could understand the English names of chemical structures. I made two versions of a test, one using the names of the structures in French and the other using the names in English but presented in a different order. The question was, which test should I administer first? Whichever was given first would give extra practice to the students, and would therefore sensitize them to the second version. I had to use the order that worked against detecting a difficulty in comprehension to make the results more convincing. Therefore, I administered the French version first.

Inherent in the design of any research is the need to collect the data in a way that avoids bias and to evaluate all possible explanations for any data collected. In the case of my chemistry terminology research, there were three intervening variables. The first one was the students' skill in drawing the structure of the chemical, which I assessed as being neutral or posing negligible interference, because the students had had extensive practice in this basic skill. The second one was the sensitization due to using the same questions in both versions of the test. Different questions would have introduced a more complicated variable: the level of difficulty of the questions. The third variable was the sensitization due to administering both tests on the same day. At the planning stage I identified these intervening variables, estimated their impact, and made sure that the research design worked against the detection of a significant difference. Therefore, any difference I did find would be more credible.

Key 10. Identify all possible explanations for the results

One year, after collating and inspecting a set of exam results from my students and those of another teacher, I noticed there were two bad classes and two good classes. There were only two teachers, so could it be that one of us was significantly worse than the other? That conclusion could easily be drawn, but there are other possible reasons for the difference. To my relief, I found that each of us had a good class and a bad class. But even if one of us had had two bad classes, that in itself would have proved nothing about us as teachers.

As we know from keys 1 and 7, circumstances and variables are important. It turned out that the two bad classes had their lessons on Friday afternoon and Saturday morning, respectively. It would have been easy to explain away the results as being caused by the lesson time. How could evidence be accumulated to test this hypothesis? One way would be to change the timetable so that the good classes received the Friday afternoon or Saturday morning time slot, and the bad classes received the desirable time slots at the beginning of the week. If simply the day and time caused the results, then this change should show up in the next set of tests.

The idea is attractive, but would not necessarily work, for there are still other factors that could influence the change. As soon as a school starts changing classes, it introduces a new variable that was not originally there. The moment a teacher notices the difference between classes, that teacher will have a natural tendency to teach the poorer class differently: either to teach better to help students improve or to teach with less effort because of lowered motivation. With simply a change in the teacher's attitude, the poor classes could improve or get worse. The students

themselves, noticing they have done badly, could make an improvement or a slide further into mediocrity.

Another method of testing the hypothesis would be to look for similar scenarios in other subjects. Is there a trend of poorer results, regardless of the class and teacher, for classes taught at inconvenient times? Is it a trend that repeats from year to year?

This example illustrates well the problem of making inferences from numerical data. The crucial concept here is the variables. When teaching this example, I also include the related topic (and well-known source of confusion) of the difference between *cause* and *correlate*. For interested readers, I recommend books on critical thinking, such as Warburton (1996), Thomson (1996), and Thouless (1974).

Conclusion

All teachers should have a critical feel for statistical reasoning and research design. Teachers use statistics, and an understanding of these fundamental keys should be part of their skills repertoire. A grasp of the concepts outlined in this article should enable any teacher to handle quantitative data.

Using statistics correctly is part of the wider issue of arguing fairly and therefore is the concern of all who rely on research. Tukey (1962:13–14) reminds us to think about why we are even using numerical data: "Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise."

References

Brown, J. D. 1991. Statistics as a foreign language. Part 1: What to look for in reading statistical language studies. *TESOL Quarterly, 25*, 4, pp. 569–586.

———. 1992. Statistics as a foreign language. Part 2: More things to consider in reading statistical language studies. *TESOL Quarterly, 26*, 4, pp. 629–664.

Burgess, R. G. (ed.). 1982. *Field research: A sourcebook and field manual.* London: Allen and Unwin.

———. 1983. *Experiencing comprehensive education: A study of Bishop McGregor School.* London: Methuen.

———. 1984. *In the field: An introduction to field research.* London: Allen and Unwin.

Cohen, L., and L. Manion. 1985. *Research methods in education.* (2nd ed.). London: Croom Helm.

Fitz-Gibbon, C. T., and L. L. Morris. 1987. *How to analyse data.* Thousand Oaks, Calif.: Sage Publications.

Huff, D. 1954. *How to lie with statistics.* Reprint. London: Penguin Books, 1991.

Lowe, I. 1992. *Scientific language at pre-university level between French and English.* Unpublished doctoral thesis, University of Surrey, U.K.

Matthews, J. R., J. M. Bowen, and R. W. Matthews. 1996. *Successful scientific writing: A step-by-step guide for the biological and medical sciences.* Cambridge: Cambridge University Press.

Nunan, D. 1989. *Understanding language classrooms: A guide for teacher-initiated action.* Hemel Hempstead, U.K.: Prentice Hall.

———. 1992. *Research methods in language learning.* Cambridge: Cambridge University Press.

Rowntree, D. 1981. *Statistics without tears: A primer for nonmathematicians.* London: Penguin.

Scholfield, P. 1995. *Quantifying language: A researcher's and teacher's guide to gathering language data and reducing it to figures.* Clevedon, U.K.: Multilingual Matters.

Thomson, A. 1996. *Critical reasoning: A practical introduction.* London: Routledge.

Thouless, R. H. 1974. *Straight and crooked thinking.* London: Pan Books.

Tukey, J. W. 1962. The future of data analysis. *Annals of Mathematical Statistics, 33*, 1, pp. 13–14. Cited in C. C. Gaither and A. E. Cavazos-Gaither. 1996.

Warburton, N. 1996. *Thinking from A to Z.* London: Routledge.

Widdowson, H. G. 1994. The ownership of English. *TESOL Quarterly, 28*, 2, pp. 377–389.

Wiley, T. G., and M. Lukes. 1996. English only and standard English ideologies in the U.S. *TESOL Quarterly, 30*, 3, pp 511–535.

Woods, A., P. Fletcher, and A. Hughes. 1986. *Statistics in language studies.* Cambridge: Cambridge University Press.

Author's note: Teachers can benefit from one of the many introductory course books in statistics. A good book, although slightly outdated in content and style, is the frequently reprinted classic by Huff, *How to Lie with Statistics* (1954). My favourites among the basic texts are Rowntree's *Statistics without Tears* (1981) and Fitz-Gibbon and Morris's *How to Analyse Data* (1987). I also recommend Nunan (1992), which has a helpful section on the logic of statistical inference, and Nunan (1989), whose appendix summarizes the basics of research methods. Brown (1991) is particularly helpful in explaining the key steps of statistical reasoning. Brown (1992) clearly explains two major areas: variables and the choice of statistical method.

Ivan Lowe is a lecturer in the Faculty of Human and Social Sciences at the University of Tunis and specializes in ESP.