Conducting Tests of Hypotheses: The Need for an Adequate Sample Size

Ratnawati Mohd Asraf and James K. Brewer
International Islamic University Malaysia and Florida State University

Abstract

This article addresses the importance of obtaining a sample of an adequate size for the purpose of testing hypotheses. The logic underlying the requirement for a minimum sample size for hypothesis testing is discussed, as well as the criteria for determining it. Implications for researchers working with convenient samples of a fixed size are also considered, and suggestions are given about the steps that should be taken when they are not able to obtain a large enough sample. Finally, the implications of not having an adequate sample size for hypothesis testing are discussed to highlight the importance of determining sample size prior to conducting one's study.

Introduction

A very important requirement of hypothesis testing that is often neglected by educational researchers is the collecting of a sample of an adequate size (Brewer 1972, Cohen 1962, 1965, 1988, Olejnik 1984, Kraemer and Thiemann 1987). In addition to having to meet the assumption of random selection, a sample has to meet minimum size requirements before we can have any confidence in findings based on it. However, many researchers are unaware of the importance of sample size, and how they obtain a sample is influenced more by 'local tradition … unaided intuition … negotiation (the latter usually between doctoral candidate and sponsor, or author and editor)' (Cohen 1962, p. 145), and availability (Brewer 1972, Brewer and Sindelar 1987, Cohen 1962, 1988, Olejnik 1984) than on a set of underlying principles or criteria. In addition, misconceptions abound regarding the size of the sample to be used. Many base the sample size required for inferential procedures such as hypothesis testing on that required for confidence intervals – although the requirements may be drastically different (Brewer and Sindelar 1987) – while others rely on the rule of thumb that '30 is enough' (Crookes 1991). Indeed, a perusal of the inferential studies conducted in the behavioral sciences will reveal the lack of awareness on the part of researchers of the importance of adequate sample size as a requirement of statistical inference, judging from their lack of explanation or justification for the sample size they used. This observation has also been noted by others, such as Cohen (1962), Olejnik (1984), and Sedlmeier and Gigerenzer (1989).

This article addresses the importance of obtaining a sample of an adequate size for the purpose of statistical inference through hypothesis testing. The logic underlying the requirement for a minimum sample size for this statistical procedure will be discussed, as will the criteria for determining it. Finally, implications for researchers working with convenient samples of a fixed size will be considered, and suggestions given as to the steps that should be taken by researchers who are not able to obtain a large enough sample.

It will be assumed throughout this paper that researchers using, or contemplating on using, hypothesis testing as a statistical inference technique are doing so because they are comfortable with the method and they consider it an appropriate technique for their research purposes. If such is the case, it would be important to pay heed to the fundamental assumptions and requirements on which the method is based. The discussion will focus on minimum sample size as merely one of the requirements of hypothesis testing. Other requirements of hypothesis testing, such as random selection, and other assumptions of statistical tests such as the scale on which the variables were measured and the nature of the population from which the sample was drawn, although just as important, will not be discussed. It is also important to note, at this point, that the requirement for an adequate sample size applies only when we use statistics for *inferential* purposes, such as in hypothesis testing and interval estimation. There is no requirement for a minimum sample size when we use statistics for *descriptive* purposes, that is, when we are not interested in generalising our results beyond that of our sample. However, before going into a discussion of the importance of adequate sample size, it is important to consider briefly what hypothesis testing entails.

Hypothesis testing

Hypothesis testing is basically concerned with the rejection (or not) of a null hypothesis. For instance, a researcher may believe that the true (or population) state of affairs pertaining to English language instruction is that ‘Students who undergo instruction under the communicative approach to language teaching will perform *better* on a test of English proficiency than those who undergo instruction under the grammar-based approach’. To gather support for his or her hypothesis (commonly termed the alternate hypothesis) the researcher has to collect sample evidence, which would involve the collecting of English proficiency scores from two different groups of students, one group undergoing instruction under the grammar-based approach, and the other under the communicative approach. These data will then be used to reject or not reject the null hypothesis, that 'There is *no difference* in the true average English proficiency scores of students who undergo instruction under the communicative approach and students who undergo instruction under the grammar-based approach'. In essence, hypothesis testing thus involves the collecting of sample data to make inferences or assertions about the hypothetical true state of affairs, that is, about data relating to the population or populations that we are unable to obtain or measure.

The need for an adequate sample size

Because statistical inference involves making inferences from a *limited* sample of scores to that of the *entire collection* of unobserved scores (the population), that inference will be subject to error; that is, there is some *probability* that the conclusions that we reach as a result of hypothesis testing are in error.

There are two types of error that we could make in hypothesis testing. The first is Type I error, which is rejecting the null hypothesis when the null is indeed true, and the second is Type II error, which is *not rejecting* the null hypothesis when the null is indeed *false*. Thus, if the null is in fact true, that is, if the true state of affairs is that 'there is *no difference* in the population English proficiency scores of students who have undergone instruction under the communicative approach and students who have undergone instruction under the grammar-based approach', but we have *rejected* the null, and concluded that 'Students who have undergone instruction under the communicative approach *perform better* on a test of English proficiency than those who have undergone instruction under the grammar-based approach', then we have committed a Type I error. Similarly, if the true state of affairs is that the null hypothesis is, in fact, *false*, that is, that the method *does* make a difference in the population, and that students who have undergone instruction under the communicative approach *do*, in fact, perform better than students who have undergone instruction in the grammar method, but we have *failed* to reject the false null hypothesis, we have then made a Type II error.

In hypothesis testing, as in any kind of statistical inference, we will never know whether or not we have made an error in our conclusions, because the 'truth' lies in the population data, which we are unable to obtain. However, we are, or should be, concerned with the *probability* of making errors. The first of the two probabilities of error in hypotheses testing is *alpha*, which many researchers are familiar with as the significance level. Alpha is, in fact, the *probability* of making a Type I error; that is, the probability of rejecting the null hypothesis when the null is indeed true. In other words, it is the probability of *incorrectly* rejecting the null hypothesis. Although many researchers adopt the conventional alpha of .05, alpha can be set at any level, as it is a subjective decision on the part of the researcher (Brewer 1988, Cohen 1962, 1990, Labovitz 1968, Rosnow and Rosenthal 1989). In fact, it represents the maximum risk of Type I error that the researcher is willing to tolerate. The second probability of error in hypotheses testing is *beta*, which is the probability of not rejecting the null hypotheses when the null is indeed false. As with alpha, beta can also be set at any level, and *independently* of alpha, and represents the maximum risk of Type II error that the researcher is willing to tolerate.

In addition to the concern for the probability of making a Type I or Type II error, we are also concerned with the *probability* of a *correct rejection*, also known as *power*. Because we would have to reject the null hypothesis in order to claim support for the alternate hypothesis, in which our theory or hunch is invested, we should be especially concerned with the probability of rejecting it *correctly*; that is, the *probability* of *rejecting* the null when the null is in fact false. Like the setting of alpha, the setting of power is a subjective decision on the part of the researcher, and can be set at any level. Further, since power is in fact, 1 - beta (see, for example, Brewer 1988, Glass and Hopkins 1984, Hays 1994), setting beta would automatically be setting power. Hence, if a researcher were to set beta very low, he or she would automatically be setting power very high. Because of the importance of power as it pertains to the probability of correct rejection, and because researchers do not have a lot of guidance when it comes to setting power, Cohen (1965) suggested that studies utilising hypothesis testing as a statistical inference technique should aim for a power of *at least* .80. However, it should be remembered that this is the *minimum* power that should be considered acceptable for any particular study, and should not be taken unthinkingly as a convention on which to base the level of power. Indeed, power should be set as close as possible to 1, if the researcher can afford it, since he or she would want as high a probability of correct rejection as possible (Brewer 1988, Hays 1994).

In planning their studies, researchers should thus give careful consideration to the two probabilities of error in hypotheses testing. In order that we may have comfort in our findings, we should set alpha and beta at the lowest level that we can possibly afford. Although we can never guarantee that our conclusions will be free from error, we want a very low *probability* of making erroneous decisions. That is, we want to aim for the highest probability of correct rejection (power) and probability of correct non-rejection (1 - alpha) that we can possibly afford. In hypothesis testing, it is possible to *guarantee* the level of power as well as the probability of making a Type I error to the levels that we have set (Brewer 1988, Cohen 1988, Davies 1961).

However, this can be done *if* and *only if* we fulfil the minimum sample size requirements necessary to justify these values.

Factors affecting adequate sample size

To control for the probability of making a Type I error and for the probability of correct rejection of the null hypothesis (power), a researcher has to obtain the required minimum sample size. Indeed, it has been shown (Cohen 1988, Dixon and Massey 1957, Davies 1961, Patnaik 1949) that once the null and alternate hypotheses are established, the minimum sample size required for hypothesis testing is a mathematical function of three factors: alpha, power, and *a priori* effect size. For most statistical tests these can be expressed by fairly simple formulae. The third determinant of sample size, known as effect size, which has not been discussed up to this point, can be understood as the size of the 'treatment effect' that the researcher wishes to detect with preset power (Brewer 1978, Cohen 1988). Researchers should carefully consider effect size when planning their studies, as it addresses the question 'How large an effect (such as a true difference in means) in the population do I expect *actually* exists, or want to be able to detect?' (Cohen 1962).

Consider the following situation. Suppose the researcher believes that the null hypothesis, 'The true (population) average mathematics scores of students who undergo instruction under the traditional approach will *not differ* from those who undergo instruction under the integrated approach' were *really false*. Suppose he or she believes there really *is* a difference in the population between the average mathematics scores of students who have undergone instruction under the integrated approach and those who have undergone instruction under the traditional approach, and that students who have undergone instruction under the integrated approach have *higher* average mathematics scores than those who have undergone instruction under the traditional approach. The researcher's judgement as to how large this difference must be in the population before he or she considers it to be important or meaningful is called *a priori* (before data) effect size. In other words, in considering effect size, the researcher is saying that the integrated approach to mathematics instruction will not be considered to be any more effective than the traditional approach unless the averages differ by *at least* a certain amount. This difference in the population means will be *estimated* by the difference in sample means (this estimate is called *post hoc* effect size, i.e., after data). Referring again to the previously stated null and alternate hypotheses, what would the researcher consider to be a meaningful difference in the true mean scores of the two populations, a difference that he or she could attribute to the 'effect' of the two different treatments? Would she consider a difference of 1 point as being meaningful? Or would she consider only a difference exceeding 5 points or 8 points or 10 points to be meaningful? Thus, *a* *priori* effect size, in effect, is the researcher's subjective *judgement*, or *criterion*, or *standard* for what is to be considered a meaningful population difference. When the researcher has collected her data and has found an *actual difference* in the mean scores between the two sample groups, this difference, the *post hoc* effect size, is then compared to the *a priori* effect size that was set before collecting the data. If the difference in the observed mean scores meets the researcher's expectation of what would constitute an important true difference, the observed or sample difference is then considered important, or meaningful. (This is not to be confused with statistical significance, which will be discussed later.) Hence, if a researcher feels that there should be a difference of at least 10 points between the two population means before he or she would consider it large enough to be important, then this difference of 10 points is the *a priori* effect size. Any difference in the observed sample means (*post hoc* effect size) below 10 points could be considered by the researcher as small or 'trivial'.

Despite its importance, however, effect size has long been neglected by researchers (see, for example, Bakan 1967, Brewer 1972, Chase and Tucker 1975, Cohen 1962, 1990, Crookes 1991, Shaver 1993). A perusal of the journals in the behavioural sciences indicates that researchers have often failed to discuss what they consider to be a meaningful difference (*a priori* effect size), and thus whether the findings meet the expectation of what this 'meaningful difference' should be is also seldom discussed. Researchers have also often confused population (*a priori*) effect size with sample (*post hoc*) effect size (Sedlmeier and Gigerenzer 1989).

One of the reasons why many researchers fail to consider effect size carefully may be their preoccupation with obtaining significant results. Indeed, concern with significance has often misled researchers into equating *significance* with *importance* (see Bakan 1967, Bracey 1991, Brewer 1988, Carver 1978, Cohen 1990, Gold 1969, Harcum 1989, Rosnow and Rosenthal 1989, Shaver 1993). Researchers should not confuse significance (which only means that the null hypothesis has been rejected) with substantive importance (which is when the difference in, for example, observed mean scores meets the researcher's expectations of what a meaningful true difference should be). A statistically significant finding can be obtained on very small, or trivial, differences. Would a researcher consider differences of one point or even $^1/_2$ a point to be meaningful, as judged by his or her *a priori* effect size? Would the sample results, then, be able to suggest anything about the 'effect' of one treatment over the other? Thus, it should be obvious that *a priori* and *post hoc* effect size should be one of the main concerns of researchers who are conducting hypothesis testing, and the question, 'How large an effect in the population would I like to detect?' should be one of the first questions they should ask when planning their studies (Brewer 1978, Cohen 1962, 1990, Shaver 1993).

Determining adequate sample size

Once a researcher has determined *a priori* effect size, alpha and power, he or she can determine the required sample size by referring to the sample size formula (if available) for the particular statistical test that he or she has chosen (see Brewer 1988, Dixon and Massey 1957, Glass and Hopkins 1984, Hays 1994, Patnaik 1949). A somewhat easier alternative, however, would be to refer to Cohen's *Statistical Power Analysis for the Behavioral Sciences* (1988), in which he has provided tables for a large range of values for sample size, alpha, power and effect size for most standard statistical tests. Also included with the tables are discussions and interpretations of the inter-relationships between alpha, power, effect size and sample size.

Suppose a researcher wishes to test the null hypothesis 'The true average mathematics scores of students who undergo instruction under the traditional approach will not differ from those who undergo instruction under the integrated approach'. Suppose she feels comfortable that she could meet the assumptions of the *t*-test, and then decides to use the *t*-test to test the null against a directional alternative hypothesis at an alpha of .05, power of .80 (hence a beta of .20, since power = 1 - beta), and an *a priori* medium effect size. Then, referring to Cohen's (1988) sample size tables for the *t*-test (see Table 1 for an abbreviated example of Cohen's table), and using the values that the researcher had set for alpha, power and medium effect size (using Cohen's suggested difference of .50 standard deviation as being the size of a medium effect: see Cohen 1988, p. 26) she will find that the sample size she would need would be 50 per group, thus bringing the total sample size to 100. On the other hand, if she wanted to test the hypothesis at a more stringent alpha of .01, with power and effect size remaining at the same levels (i.e. at .80 and medium respectively), the sample size needed would be much larger, that is, 82 per group, or a total sample size of 164. A higher level of power, with effect size at medium and alpha at the level that was originally set at .05, would also require a larger sample size. For example, if the researcher wanted power to be increased to .95, she would need a sample size of 87 per group, or a total sample size of 174. Thus, if the researcher could afford the increase in sample size, it would be in the best interest of the research if she set power as high as the sample size allows, for a hypothesis test with a power of .95 would certainly give the reader more assurance of a correct rejection than a test with a power of .80 or lower.

It should be obvious, at this point, that the sample size needed for a particular hypothesis test depends on the values of alpha, power and effect size set by the researcher. Often, the maximum level of power as well as the risk of Type I error that he can set for his study will be influenced by practical considerations such as finding an adequate sample. This is where knowledge of the determinants of sample size will allow him to adjust the values of alpha and power to the level that he considers appropriate, given the sample size. To illustrate, Table 1 provides the minimum sample size needed for testing hypotheses with the two independent samples *t*-test at different levels of alpha, power and effect size using Cohen's tables.

	alpha = .01 (directional)				**alpha = .05 (directional)**		
	Effect size				**Effect size**		
Power	small	medium	large	Power	small	medium	large
.25	138	24	10	.25	48	8	4
.50	272	45	18	.50	136	22	9
.80	503	82	33	.80	310	50	20
.90	652	105	42	.90	429	69	27
.95	790	128	51	.95	542	87	35
.99	1084	175	69	.99	789	127	50

* Where n is the sample size per group. Hence, total sample size needed will be 2n.

Table 1: n to detect effect size by independent samples *t*-test*

Relationship between alpha, power, effect size and sample size

As Table 1 indicates, for the same effect size, the lower the alpha and beta (which is 1 - power) set by the researcher, the larger the sample size needed. If, on the other hand, alpha and effect size are kept constant, the higher the power, the larger the sample size needed. In essence, with all else held the same, there is a price that the researcher has to pay to enjoy lower probabilities of Type I and Type II errors, namely a larger sample. This should be somewhat intuitive, since by collecting more data the researcher gains more assurance of the correctness of any decision made on the basis of those data. It is important to emphasise once again that alpha and beta can be set at *any level* desired by the researcher, the most desirable of which is to set *both* alpha and beta at the *lowest level* that the researcher can afford. Unfortunately, many researchers have the misconception that alpha and beta are *inversely related*, and believe that if they set alpha at a low level, beta would automatically become high. The following statement by an author of an applied linguistics textbook reflects this misconception: 'But in attempting to control for Type I error by selecting a low alpha level, we increase the likelihood of Type II error. We might, therefore, consider raising the .05 [alpha] level to .10 to ensure that we will be able to detect a false null hypothesis' (Lazaraton 1991, p. 760).

Lazaraton's statement is not only erroneous, but it would lead researchers to believe that they have no choice but to accept the trade-off of getting a high beta for setting alpha low, and vice versa. And this would deprive them of the privilege of 'enjoying' tolerable error rates, which they could have obtained by determining the sample size on the basis of *both* low alpha and beta values plus the desired effect size set prior to collecting the data.

A perusal of journal articles and books written on quantitative methodology reveals that the misconception of the inverse relationship between alpha and beta is widely held by researchers. However, in all fairness, perhaps researchers should not be blamed entirely for having this misconception, as it is perpetuated, regretfully, by many statistics textbooks themselves, where the authors assume that researchers are starting with a *fixed* sample size and effect size (Brewer 1985). Even textbooks by reputable authors such as Glass and Hopkins (1984) and Hays (1994), although very clear on other conceptual matters, make this tacit assumption. The inverse relationship of alpha and beta is true *only* if sample size and effect size are fixed, but not when the researcher sets alpha, beta and effect size at the planning stages of his or her study and *then* obtains the required sample size to justify those values. The relationship between alpha and beta should not be discussed without considering sample size and effect size, as they are also part of the mathematical relationship. Doing so could cause researchers to have the misconception that the inverse relationship between alpha and beta is always true – without the caveat, 'if sample size and effect size are fixed'. The values of alpha and beta can be set independently of one another, and are *not related*, except when the researcher uses a readily available sample of a fixed size.

Using a sample size of convenience

When a researcher obtains a readily available sample of a fixed size, in other words, a sample size of convenience, he or she is not at liberty to set all the values of alpha, beta and effect size at the level that he or she wishes. Given a fixed sample size, she would be able to set only two values at most, and the third will be determined by their mathematical relationship. Consider the following example. Suppose a researcher, using the independent samples *t*-test to test the hypothesis that methods A and B do not differ, had obtained a readily available sample of size 60, which she split into two groups of 30 subjects per group. Suppose she then decides to adopt the conventional alpha of .05 and a medium effect size. She will then find that she will no longer be able to set beta, and hence power, at any level. Using Cohen's sample size tables for the *t*-test (Table 1), power would automatically come out to be .60. This means that she will also not be able to control beta, which would come out to be .40, since beta is 1 - power. However, if she had set alpha, beta and effect size independently and *then* based her sample size on those values, she would find that she could set both alpha and beta at any level. For example, if she wanted both her alpha and beta at the .05 level, and effect size at medium, what she would have to do is to obtain the minimum sample size required, which, referring to Cohen's sample size tables (see Table 1), will turn out to be 87 per group, or a total sample size of 174.

Thus, it is important that researchers carefully consider the values of alpha, beta and effect size *prior* to data collection and then determine sample size based on these values, for this is the only way that they can set the values independently of one another. Conducting an inferential study based on a readily available sample of a fixed size would be considered a very weak research procedure. Not only do samples of convenience often fail to meet the requirement of random selection, but they also do not allow the researcher the privilege of setting alpha and beta at tolerably low levels (and thus power at a sufficiently high level) to make them justifiable from a research point of view.

Implications of not meeting sample size requirements

When doing hypothesis testing, it is thus very important that the researcher obtain an adequate sample size, in order that the levels of alpha, power and effect size that the researcher sets may be justified. Failure to meet the minimum sample size required means that he cannot claim the levels of alpha, power and effect size that he set *a priori*. For instance, suppose a researcher wishes to use the independent samples *t*-test to test his null hypothesis against a directional alternative hypothesis at an alpha of .05, power of .90, and small effect size. He needs to obtain a total minimum sample size of 858 for the two groups (see Table 1). Instead he obtains a total sample size of 272 because of unavailability of subjects. Then, assuming that he wanted to maintain the significance level at the previous .05 and the same small effect size, he would not have any control over the level of beta, and hence power. Beta would automatically be raised from the previous .10 to .50, and consequently power lowered from .90 to .50. In other words, whether the researcher realises it or not, he would not be able claim a power of .90 because of the drastic decrease in sample size. For researchers who are aware of what a decrease in sample size would do to the rates of alpha and power that they had previously set, the most justifiable thing to do would be to readjust the levels of alpha, power or both to levels justified by the new reduced sample size and report them as such. However, if they cannot tolerate the adjusted levels of alpha, power or both, that is, if alpha is too high or power too low to be justifiable from a research point of view, then they should consider postponing the study until they can get a larger sample.

There is ample evidence in the literature to indicate that most researchers are unaware of the implications of not having an adequate sample size. In a meta-analysis of the studies published in the *Journal of Abnormal and Social Psychology*, *Journal of Educational Research*, *Journal of Research in Science Teaching*, *Research Quarterly*, *Journal of Educational Measurement*, *Journal of Communication* and *Journal of Applied Psychology* by Cohen (1962), Brewer (1972), Jones and Brewer (1972), Penick and Brewer (1972), Brewer and Owen (1973), Katzer and Sodt (1973), and Chase and Chase (1976), respectively, the researchers found a dismal state of affairs with regard to the power of the tests conducted, as indicated in Table 2.

		Effect size		
Journal studied	**Researcher(s)**	**small**	**medium**	**large**
Journal of Abnormal and Social Psychology	Cohen (1962)	.18	.48	.83
Journal of Educational Research	Brewer (1972)	.14	.58	.78
Journal of Research in Science Teaching	Penick and Brewer (1972)	.22	.71	.87
Research Quarterly	Jones and Brewer (1972)	.14	.52	.80
Journal of Educational Measurement	Brewer and Owen (1973)	.21	.72	.96
Journal of Communication	Katzer and Sodt (1973)	.23	.56	.79
Journal of Applied Psychology	Chase and Chase (1976)	.25	.67	.86

Table 2: Results of studies on the power of hypotheses tests

As Table 2 indicates, the tests surveyed in all seven journals had extremely low power to detect a small effect size and not much better for detecting a medium effect size. For example, Cohen (1962), in a seminal study on the analysis of power, found that the tests he surveyed had, on the average, a power of only .18 for detecting small effect sizes. In other words, if the researchers had wanted to detect a small treatment effect, they had a probability of *correctly rejecting* the null hypothesis of only .18, which is hardly encouraging from a research point of view. The rest of the tests published in the other six journals did not fare much better in terms of power to detect a small effect, with the highest having a combined average power of only .25. The fact that the calculation of power was done on tests that had reached significance, that is, in which the null hypotheses had been rejected, in no way changes the probability of correct rejection (which remains at 1 - beta) and the probabilities of error (which remain at alpha and beta) since these values are probabilities of occurrence *in the long run*, with repeated random sampling, given a sample size n (Brewer 1988, Shaver 1993). The null hypotheses of those studies may have been rejected, but we have no way of knowing if they had been *correctly rejected* or if the researchers had made an error, since the true state of affairs regarding the null is never known, as one does not have the population scores at hand. However, the extremely low power of the tests does raise concerns as to the long run correctness of the rejections, and hence the appropriateness of the conclusions made.

The power of the tests to detect medium effect sizes could also be considered meagre. With the exception of the studies analysed by Penick and Brewer (1972) and Brewer and Owen (1973), in which the mean power of the tests was computed at .71 and .72 respectively, the rest of the tests in the journals surveyed had a more or less 50–50 chance of detecting a medium effect. This implies that the researchers who conducted hypothesis testing at such low levels of power would have saved a lot of time and energy had they just tossed a coin in deciding whether or not to reject since the probability of a correct rejection was approximately $^1/_2$. The only instance when the tests could be considered to have sufficiently high power is when, for example, the researchers had wanted to detect a large effect size.

However, it is not likely that the researchers would have wanted to detect large effects, as large differences 'would be so obvious as to virtually render a statistical test superfluous' (Cohen 1962, p. 147). An example of a difference this large would be the difference between the English language proficiency scores of non-native students of English who had scored 450 on the TOEFL and that of honours students who are native speakers of English. An additional point to note regarding the studies published in the seven journals is that the researchers had not based their sample size on an analysis of the levels of alpha, power and effect size. Hence, the sample size they used was too small to yield reasonable levels of power to detect a small or even a medium effect size, as indicated by the lack of justification for the sample sizes used. Indeed, the levels of alpha, power and effect size were hardly mentioned in these studies. Had the researchers been aware of the importance of setting alpha, power and size of effect, and consequently based their sample size on those values, then the power of their tests might not have been as dismal as has been reported.

In the preceding discussion, we used the *t*-test to illustrate the relationship between alpha, power and effect size to sample size, and the importance of obtaining a sample of an adequate size, as it relates to the likelihood of *correctly* rejecting the tested hypotheses, given a particular alpha level and effect size. It should be pointed out that these considerations apply to the use of *all* statistical tests. It is beyond the scope of this paper, however, to discuss sample size requirements for different statistical tests, but the reader is encouraged to refer to Cohen (1988), in which he has provided tables for a large range of values for determining adequate sample size for most statistical tests, given certain pre-set values of alpha, power and effect size, as well as clear and detailed explanations on how to determine the required sample size. The reader may also refer to Bradley, Russell and Reeve (1996) for a discussion of power in complex experimental designs, and Snyder and Lawson (1993), Kirk (1996) or Friedman (1968), (all cited in Thompson 1997), on the kinds of effect size statistics that can be reported, as well the formulas to compute *post hoc* effect sizes, which can be easily implemented with a computer spreadsheet or calculator. Computer software is now also available to calculate power and sample size, and an excellent review of statistical power analysis software is given by Thomas and Krebs (1997), who rate the packages in terms of the scope of their tests, their accuracy, ease of use and ease of learning. However, as is the case with the use of any software, it is not only important for the researcher to understand how the procedure works, that is, how the statistics are computed, but also its underlying assumptions and limitations.

It is noteworthy to mention that as a result of the ongoing debate on the use of hypothesis testing in psychology journals, and following the publication of Cohen's 1994 article, the American Psychological Association formed a Task Force on Statistical Inference to explicate, among other things, issues surrounding the use of statistics, including the use of significance testing (Wilkinson and Task Force on Statistical Inference 1999). This has led to the revision of the research section of the fifth edition of the *Publication Manual of the American Psychological Association* to include the requirement that effect size estimates be provided: 'For the reader to fully understand the importance of your findings, it is almost always necessary to include some index of effect size or strength of relationship in your results section' (American Psychological Association 2001, p. 25). Indeed, the first chapter of the fifth edition includes the failure to report effect sizes as among the defects in the design and reporting of research. This represents a shift in editorial policy from that of the fourth edition that merely 'encouraged' (American Psychological Association 1994, p. 18) effect size reporting. The fifth edition also admonishes researchers to 'Take seriously the statistical power considerations associated with your tests of hypotheses' (American Psychological Association 2001, p. 24). Whether or not these requirements will be heeded by researchers remains to be seen, but it represents a key step towards establishing the importance of power analysis and the reporting of effect size among those conducting significance testing.

Finally, we wish to reiterate that we have focused our discussion on obtaining an adequate sample size for hypothesis testing. The calculation of an adequate sample size for confidence intervals, the other statistical inference procedure, deserves a separate treatment.

Summary

Hypothesis testing typically aims to reject the null hypothesis, that is, the hypothesis that there is no difference in the treatments in question, in order that the researcher may gain support for the alternate hypothesis, that the treatments do, in fact, make a difference. Given the fact that hypothesis testing involves making inferences from a random sample of observed scores to the entire population of unobserved scores, any inferences made will be subject to error. There is the probability of falsely rejecting the null hypothesis, that is, the probability of making a Type I error, and there is also the probability of not rejecting the null hypothesis when it is indeed false, that is, the probability of making a Type II error. Because of the probabilistic nature of hypothesis testing, what the researcher has to do, in advance of data collection, is to specify the probability of making these two kinds of errors and, through beta, setting the probability of correctly rejecting the null hypothesis (power).

Obtaining a sample of an adequate size is very important in hypothesis testing, as it allows the researcher to control for the rates of error as well as the probability of correct rejection and correct non-rejection. Thus, prior to conducting their studies, researchers should plan on collecting a sample of an adequate size based on a careful consideration of the values of alpha, power and *a priori* effect size. Failure to obtain an adequate sample size could mean that they end up with tests with high probabilities of error and hence very low power. If rejection of the null hypothesis is a major concern of researchers, then they should surely be concerned about power and, in fact, should try to set power as high as they can afford in terms of sample size, as power addresses the issue of the probability of *correct* rejection. Conducting a study where the probabilities of error are high and the power of the test very low would raise concerns about the correctness of the conclusions. In such cases, even if our data were to result in a rejection of the null hypothesis, there would be doubts as to whether it had been correctly rejected. If results of inferential studies are to be considered meaningful, then researchers should make a concerted effort to meet the assumptions and requirements of statistical inference, among which is the obtaining of a sample of an adequate size.

References

American Psychological Association (1994) *Publication Manual of the American Psychological Association*, 4th ed., American Psychological Association, Washington, DC.

American Psychological Association (2001) *Publication Manual of the American Psychological Association*, 5th ed., American Psychological Association, Washington, DC.

Bakan, D. (1967) *On Method: Toward a Reconstruction of Psychological Investigation*, Jossey-Bass, San Francisco.

Bracey, G. W. (1991) Sense, non-sense, and statistics, *Phi Delta Kappan*, vol. 73, no. 4, p. 335.

Bradley, D. R., R. L. Russell and C. P. Reeve (1996) Statistical power in complex experimental designs, *Behaviour Research Methods, Instruments, and Computers*, vol. 28, pp. 319–26.

Brewer, J. K. (1972) On the power of statistical tests in the *American Educational Research Journal*, *American Educational Research Journal*, vol. 9, pp. 391–401.

Brewer, J. K. (1978) Effect size: the most troublesome of the hypothesis testing considerations, *Center on Evaluation, Development, and Research Quarterly*, vol. 11, no. 4, pp. 7–10.

Brewer, J. K. (1985) Behavioral statistics textbooks: source of myths and misconceptions?, *Journal of Educational Statistics*, vol. 10, no. 3, pp. 252–6.

Brewer, J. K. (1988) *Introductory statistics for researchers*, 5th ed., Burgess International Group, Minnesota.

Brewer, J. K. and P. W. Owen (1973) A note on the power of statistical tests in the *Journal of Educational Measurement*, *Journal of Educational Measurement*, vol. 10, no. 1, pp. 72–4.

Brewer, J. K. and P. T. Sindelar (1987) Adequate sample size: a priori and post hoc considerations, *Journal of Special Education*, vol. 21, no. 4, pp. 74–84.

Carver, R. P. (1978) The case against statistical significance testing, *Harvard Educational Review*, vol. 48, no. 3, pp. 378–99.

Chase, L. J. and R. B. Chase (1976) A statistical power analysis of applied psychological research, *Journal of Applied Psychology*, vol. 61, pp. 234–7.

Chase, L. J. and R. K. Tucker (1975) A power-analytic examination of contemporary communication research, *Speech Monographs*, vol. 42, pp. 29–41.

Cohen, J. (1962) The statistical power of abnormal-social psychological research: a review, *Journal of Abnormal and Social Psychology*, vol. 65, pp. 145–53.

Cohen, J. (1965) Some statistical issues in psychological research, in B. B. Wolman, ed., *Handbook of Clinical Psychology*, McGraw-Hill, New York.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed., Lawrence Erlbaum, Hillsdale, NJ.

Cohen, J. (1990) Things I have learned (so far), *American Psychologist*, vol. 45, pp. 1304–12.

Cohen, J. (1994) The earth is round (p < .05), *American Psychologist*, vol. 49, pp. 997–1003.

Crookes, G. (1991) The forum: research issues, *TESOL Quarterly*, vol. 25, no. 4, pp. 762–5.

Davies, O. L. (1961) *Statistical Methods in Research and Production*, Hafner Publishing Co, New York.

Dixon, W. J. and F. J. Massey, Jr. (1957) *Introduction to Statistical Analysis*, 2nd ed., McGraw-Hill, New York.

Friedman, H. (1968) Magnitude of experimental effect and a table for its rapid estimation, *Psychological Bulletin*, vol. 70, pp. 245–51.

Glass, G. V. and K. D. Hopkins (1984) *Statistical Methods in Education and Psychology*, 2nd ed., Prentice Hall, Englewood Cliffs, NJ.

Gold, D. (1969) Statistical tests and substantive significance, *American Sociologist*, vol. 4, no. 1, pp. 42–6.

Harcum, E. R. (1989) The highly inappropriate calibrations of statistical significance, *American Psychologist*, vol. 44, no. 6, p. 964.

Hays, W. L. (1994) *Statistics for the Social Sciences*, 5th ed., Holt, Rinehart and Winston, New York.

Jones, B. and J. K. Brewer (1972) An analysis of the power of statistical tests reported in the *Research Quarterly*, *Research Quarterly*, vol. 43, pp. 23–30.

Katzer, J. and J. Sodt (1973) An analysis of the use of statistical testing in communication research, *Journal of Communication*, vol. 23, pp. 251–65.

Kirk, R. (1996) Practical significance: A concept whose time has come, *Educational and Psychological Measurement*, vol. 56, no. 5, pp. 746–59.

Kraemer, H. C. and S. Thiemann (1987) *How Many Subjects: Statistical Power Analysis in Research*, Sage, Beverly Hills, CA.

Labovitz, S. (1968) Criteria for selecting a significance level: a note on the sacredness of .05, *American Sociologist*, vol. 3, no. 3, pp. 200–22.

Lazaraton, A. (1991) The forum: research issues, *TESOL Quarterly*, vol. 25, no. 4, pp. 759–62.

Olejnik, S. F. (1984) Planning educational research: determining the necessary sample size, *Journal of Experimental Education*, vol. 53, no. 1, pp. 40–8.

Patnaik, P. B. (1949) The non-central χ^2 and F-distributions and their applications, *Biometrika*, vol. 36, pp. 202–32.

Penick J. E. and J. K. Brewer (1972) The power of statistical tests in science teaching research, *Journal of Research in Science Teaching*, vol. 9, no. 4, pp. 377–81.

Rosnow, R. L. and R. Rosenthal (1989) Statistical procedures and the justification of knowledge in psychological science, *American Psychologist*, vol. 44, no. 10, pp. 1276–84.

Sedlmeier, P. and G. Gigerenzer (1989) Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, vol. 105, pp. 309–16.

Shaver, J. P. (1993) What statistical significance testing is, and what it is not, *Journal of Experimental Education*, vol. 61, no. 4, pp. 361–77.

Snyder, P. and S. Lawson (1993) Evaluating results using corrected and uncorrected effect size estimates, *Journal of Experimental Education*, vol. 61, no. 4, pp. 334–49.

Thomas, L. and C. J. Krebs (1997) A review of statistical power analysis software, *Bulletin of the Ecological Society of America*, vol. 78, no. 2, pp. 126–39, viewed 28 February 2003, <http://www.zoology.ubc.ca/~krebs/power.html>

Thompson, B. (1997) Computing effect sizes, viewed 15 January 2003, <http://www.coe.tamu.edu/~bthompson/effect.html>

Wilkinson, L. and Task Force on Statistical Inference (1999) Statistical methods in psychology journals: guidelines and explanations, *American Psychologist*, vol. 54, pp. 594–604, viewed 28 February 2003,
<http://www.apa.org/journals/amp/amp548594.html>