ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 2, No. 4, pp. 784-792, April 2012

doi:10.4304/tpls.2.4.784-792

Rethinking Convenience Sampling: Defining Quality Criteria

Farahman Farrokhi
English Department, Faculty of Persian Literature and Foreign Languages, University of Tabriz, Tabriz, Iran
Email: ffarrokhi20@yahoo.co.uk

Asgar Mahmoudi-Hamidabad
PhD candidate at University of Tabriz, Iran
Email: mahmoudi301@gmail.com

***Abstract*—Convenience sampling is one of the most commonly used sampling procedures in second language acquisition studies, but this non-random sampling procedure suffers from a lot of problems including the inability of controlling for initial differences between experimental and control groups. The present study tries to introduce conditions and criteria which enable researchers to account for these drawbacks and at the same time make validity claims. Individual scores and group statistics are compared with regard to a group of essential factors known to be important for the purpose of the study. The overall value calculated for essential factors is then used to make judgments about the groups' comparability. The contribution of this method to the current procedures of sampling arises from its factual accuracy which is supposed to enhance the validity of findings obtained from studies employing non-probability sampling procedures.**

***Index Terms*—convenience sampling, comparability, conditions, criteria, outlier**

I. Introduction

Convenience or opportunity sampling is the most common type of sampling in L2 studies where the only criterion according to Dörnyei (2007) is the convenience of the researcher. But, in convenience sampling or in working with intact groups, there is a high probability that the two selected groups will constitute "low-set" and "high-set" groups, to borrow a few words from Muijs (2004), or will be different in other ways not allowing us to conceive of them as comparable. Groups chosen by convenience sampling are conducive to self-selection, administrative decision, time of the class, number of the years of exposure and many other polluting influences. Sometimes, the difference in the composition of the two groups is so grave that undermines the overall validity of the research. Yet, in most cases these shortcomings either remain unnoticed by researchers or are simply winked at as if something natural has happened.

Another issue related to sampling is the presence of outliers. This problem is particularity relevant to the non-probability sampling, of which convenience sampling is a notorious one. This is because of the high self-selection possibility in non-random sampling. Outliers are cases which would make our findings subject to suspicion. The statistics computed from samples are often used to draw inferences about population parameters. Outliers adversely affect sample statistics and decrease the precision of estimates about population. Therefore, to approach the true parameters of population as closely as possible or, in other words, to make the best possible estimates, we have to control for the effect(s) of outliers. If reliability and validity are important and if generalizations are to be made, it seems reasonable to think of a way for accounting for the presence of outliers and defining similarity in precise terms.

This article addresses these two issues and tries to identify conditions and criteria that can be applied to convenience sampling to make its precarious status a bit more stable. The conditions and criteria are mainly related to imposing limits on the individual scores and group statistics. The comparability of groups and the degree of their similarity will be determined by reference to these individual scores and the relevant statistics such as means and standard deviations calculated for these groups. The suggestions are not made to approach the principles of laboratory studies or exert stricter control upon the sampling procedures but they are made for the sake of transparency and factual accuracy. Comparability is an elusive concept which can be interpreted in different ways. There seems to be an urgent need for this term to be clarified adequately so that the interpretations converge among the consumers of research results.

When we compare two groups and make claims on the basis of our findings, it should be clear what our groups' compositions look like. Though, two truly comparable groups should be exactly the same concerning all aspects to be able to make strong claims, this state of affairs is only an idealization far removed from reality in educational studies. Variegation is a norm in studies of second language acquisition. Accepting the impossibility of uniformity, however, researchers should be alert to the aspects and degrees of dissimilarities. Measures and standards should be defined to let the researchers decide confidently if the two available groups, with all the apparent limitations, are comparable or not.

The conditions and criteria introduced below provide us with a clearer picture of the groups' make up and the degree of faith we should put on the findings. Conditions are related to individual scores and determine whether they should be

included in the study or not. Criteria are related to group characteristics and qualify whether two groups are comparable based on the limits set on their statistics or not. The overall value calculated in each study is a function of the number of essential factors taken to be important for the essence of the study. The difference between two groups with regard to each essential factor is quantified based on a 10 point scale. If this value exceeds the number of essential factors multiplied by ten the two groups will be regarded as incomparable. If this value falls below this limit the groups will be regarded as comparable, but the comparability increases as the value approaches 0.

II. Review of the Related Literature

There are a lot of people who may well fit the design of our research, but it is a hard fact that, except for case studies and nationwide headcounts, we can never examine all the people fitting our research design. Mangal (2002) points out that it is quite impractical and inessential to approach every person fitting our research design. The convenient as well as practical solution, according to Mangal, lies in estimating the population parameters from sample statistics. Brown (1995) states that few researchers in applied linguistics are in a position to study an entire population. That is why they are usually forced to work with samples drawn from population. But, we often hope to be able to generalize our research findings with samples to broader contexts from which they are drawn. Otherwise, our findings will not be of much practical value.

A sample is defined by Tailor (2005) as a subset of a population or universe. However, a word of caution is in order here as population is often taken by many to refer to people only. Population, as Walliman (2011) puts it, does not necessarily mean a number of people. It can also refer to total quantity of the things or cases which are the subject of our study. Robson (1993) also talks about non-people-related sampling like where and when interviews take place. However, he stresses the point that particular attention needs to be given to the selection of 'people sample'. But what guarantees the applicability of our findings? One way that Hatch and Lazaraton (1991) suggest is to obtain a representative sample through random selection.

Sir Ronald Fisher is credited with the design of experiments with random sampling. The justification for randomizing is that it maximizes the internal and external validity of the findings by giving an equal chance to every single subject to be assigned to either experimental or control group and thereby leveling group differences off. Another strategy adopted by some is to enlarge the size of the sample in addition to randomizing. The assumption underlying both augmenting sample size and randomizing is that, these two techniques reduce the effects of extreme scores and extraneous factors to a negligible amount. Of these two techniques randomizing is, of course, more important because without randomizing even with big groups of subjects there is a high probability of having very different groups. The problem of subject variation is hard to surmount with small groups even with randomizing. Whereas, Hatch and Farhady (1981) maintain that randomizing allows the researcher to have two truly comparable groups prior to the start of the experiment, Robson (1993) asserts that random selection does not guarantee that the two groups, even if they are big, will in fact be equivalent. According to him, there are also practical and ethical problems when randomizing is applied to people.

From what was said it is clear that, randomizing is a remedial action, which even along with increasing sample size is not capable of completely removing uncertainties regarding the comparability of the two groups designated as experimental and control groups. But, even if we accept the idea that randomizing assures the soundness of subject selection, as it is the case with Fisherian hypothesis, in most experimental studies in educational and similar areas like psychology and social sciences, which are mainly dealing with human beings, researchers cannot afford satisfying the requirement of randomizing. Furthermore, this is an impossible line of action to be taken in qualitative and descriptive studies which do not lend themselves to manipulation and artificial categorization. The worse that happens, therefore, is the use of intact groups or convenience sampling without employing any clear criteria suggesting the similarity and therefore comparability of groups. Neither is any criterion available as to the degree of similarity.

The purpose of presenting this background to sampling is not to review the history of different types of sampling and their merits and demerits. Rather, the purpose of the study is to grapple with one major but problematic category of sampling which frequently figures in the literature, i.e. convenience sampling. The problems with convenience sampling are so acute that Robson (1993) calls it as a cheap and dirty way of doing research. Who gets sampled, according to Robson, is determined by all kinds of unspecifiable biases and influences introduced to the sampling procedure. This does mean that most experimental findings' validity is subject to attack.

Convenience sampling is a kind of non-probability or nonrandom sampling in which members of the target population, as Dörnyei (2007) mentions, are selected for the purpose of the study if they meet certain practical criteria, such as geographical proximity, availability at a certain time, easy accessibility, or the willingness to volunteer. Dörnyei further explains that, "captive audiences such as students in the researchers' own institution are prime examples of convenience sampling." Mackey and Gass (2005) point out that the obvious disadvantage of convenience sampling is that it is likely to be biased. They advise researchers that the convenience sampling should not be taken to be representative of the population.

Still, there is another problem of great concern related to intact groups or convenience sampling, i.e. the problem of outliers. Because of the high self-selection possibility in non-probability sampling, the effect of outliers can be more devastating in this kind of subject selection. Outliers are cases whom Hatch and Lazaraton (1991) consider as not belonging to the data. Larson-Hall (2010) regards outliers as points which stand out as being different from the bulk of

the data and believes that they are very problematic for "classic statistics" in which normality of the distribution is an obvious prerequisite. According to Larson-Hall, "classic statistics" advises using graphic summaries to identify the outliers and then remove them by running the analysis both with and without them. However, she identifies that there are several problems with manually deleting outliers. The first is that, the deletion process is arbitrary and up to the discretion of the individual researcher. Moreover, eliminating one outlier may make another subject stand out as outlier. Best and Kahn (2006) consider outliers a problem for correlational studies. They believe that outliers can cause spuriously high or low correlations. When this occurs, these authors believe that, the researcher needs to decide whether to remove the individual's pair of scores from the data analyzed or not.

Outliers are, therefore, a constant threat to the homogeneity of our experimental and control groups on the one hand and to the reliability and validity of our findings on the other if they are not accounted for in a systematic manner. Of course, the presence of outliers does not impose as big a challenge on qualitative researchers as they do on quantitative researchers since these researchers are quite often interested in exceptional cases. For instance, a case study, according to McKay (2006) can be a single instance of some bound system although it can also encompass an entire community.

III. Logic of the Study

It is clear from many studies that there are some essential factors that affect any kind of learning. Of course, some of these factors are related to group characteristics and others to the characteristics of individuals participating in the study. In the case of language learning, studies suggest that age, gender, level of language proficiency, years of schooling, attitude, motivation, etc. affect the rate and amount of learning. It is also certain that people with different abilities reflect on linguistic problems and situations differently and approach them adopting different strategies. So, it is surprising that no criteria as yet have been proposed for controlling these variables in sampling, especially in non-random subject selection. The problem, therefore, is that while logic has brought the need for reevaluation of subject selection to the fore, there is not a reliable way to this day to make sure that non-random groups or subjects assigned to experimental and control groups are acceptably similar.

Of course, the issue of subjective or extraneous variables has long been identified by researchers. For example, Seliger and Shohamy (1989) suggest that steps should be taken before the experiment to choose subjects who meet some pre-established criteria. They also point out that, the increase in the comparability of groups depends on the specifics of the study. Dörnyei (2007, p. 117) similarly stresses that, "in order to be able to make causal claims based on a quasi-experimental study, the effects of initial group differences need to be taken into account." Unfortunately, however, Seliger and Shohamy do not propose any practical way for controlling these variables or factors. The procedure introduced by Dörnyei too is difficult to apply since it depends on a case-by-case matching of subjects in experimental and control groups. For example, he states that we should identify participants in the two comparison groups with very similar parameters, say, a girl with an IQ around 102 in both groups. The analysis of covariance (ANCOVA) too is not applicable at this stage. This statistical procedure can be applied to studies after they are completed to remove the confounding effects of variables. Also, the analysis of covariance removes the effects of single variables, but does not give us any measure of groups' overall initial similarity or comparability.

Neither is any clear know-how suggested in the literature as to how to deal with outliers after they are identified. The application of some statistical procedures mentioned above or trimming the groups too are directly dependent on individual investigator's intuition and the decisions made are arbitrary in most cases. It is important to note again that, outliers are not always the black sheep of research designs. In qualitative research, for example, researchers are largely interested in values which are inconsistent with the rest of the dataset and it is thought that their analysis would help clarify other more important issues. In fact, extreme case analysis is a procedure exclusively devoted to the identification and study of subjects with deviant scores. However, in experimental studies every attempt is made to choose groups which are as closely similar as possible.

For these reasons, it seems necessary to think of a way to control the essential factors affecting subjects' learning in any kind of research which uses non-probability procedures of sampling prior to the beginning of the study. In any study, according to Seliger and Shohamy (1981), decisions must be made as to whether and how to set limits on the scope or focus of the investigation which may confuse the interpretation of results. To put it more simply, we cannot select any two groups at our disposal as our experimental and control groups, as the different kinds of differences have differential effects on the rate and amount of learning. The initial difference between groups prior to their involvement in the study is an issue that researchers should take account of. These differences if big can undermine the internal validity of our study and because of this, great deviations in groups' attributes should be ruled out before the study begins. Otherwise, the users of research results should be cautioned about the reliability and validity of the outcomes.

The issues related to the reliability and validity raised here are largely relevant to the preparatory stage of a research project. It is clear that there are some other factors that may affect these two important quality criteria of any research during the study. For example, the inadequacy or unsuitability of data collection method, research environment, time spent on the treatment, teacher variable, and the irrelevant immediacy or undue delay in administering post tests all can have deteriorating effects on the value of findings. But, most of the polluting factors affecting the study during its operational stage can be kept under control if they are identified in time by the researchers. For example, teacher variable can be controlled if the same teacher teaches both experimental and control groups. In the same way, the time

element can be accounted for if both classes are taught at about the same time of the day. Needless to say that, like initial differences, researchers should be on their guard against these kinds of threats to the reliability and validity of their study as well.

As to the presence of outliers, we cannot arbitrarily include subjects in or exclude them from the study. For example, if the differences in language proficiency and age are more than what are considered to be acceptable levels, the findings of the study would be the result of these factors at least to some extent rather than the teaching method employed. To control the essential factors and to identify outliers at the preparatory stage, then, we have to define legitimate boundaries for the differences within which individuals are qualified for the groups and the groups are considered as similar and therefore comparable. It is falling within the defined limits for similarity which in fact, qualifies two groups to be considered as comparable. The narrower the ranges defined for similarity the more similar the groups will be regarded to be. The broader the ranges, on the other hand, the more variety in subjects' attributes and, therefore, the less the reliability and validity of the findings will be. At the same time, the boundaries identified for individual characteristics will tell us which students should be included in the study or excluded from it.

So, any research of this kind should first and foremost employ groups whose scores, for example on a proficiency test, are restricted to the predetermined limits in order to eliminate outliers from the study. Concurrent with this restriction, the mean scores for affecting factors should also fall within the defined ranges for mean scores. Therefore, more than anything else, the main concern of researchers should be starting inquiry with groups that are more similar with regard to the essential factors which are believed to be affecting the performance of groups.

From what was said it is clear that, the purpose of the following heuristic study is to explore the issue of convenience sampling and the use of intact groups in some detail and introduce a set of conditions and criteria which can be applied to most research work in the field. These conditions and criteria will determine the acceptable ranges for subjects' scores and groups' essential factors and the way the groups' similarity is measured. The procedure introduced here is claimed to be applicable to educational field in the first instance. Its extension to other fields is a matter which prospective researchers should decide on with regard to their particular area of interest.

IV. Conditions and Criteria

In this procedure, depending on the study, a number of conditions and criteria are taken into account. Conditions are mainly related to the range of scores within which the individual subjects' scores should fall. Criteria are, however, related to group statistics. For example, in the educational field the mean and standard deviation of the groups obtained from a test of language proficiency, age, language background, years of schooling, and gender are considered to be essential for most studies. The number of essential factors, of course, can be more or less than this depending on the description of groups by the researcher or how rigorous the researcher wishes to be. The smaller the range of individual scores and the bigger the number of essential factors the more controlled our sampling procedure will be. However, it is obvious that controlling all factors is neither possible in practice nor wise. The kind and number of essential factors, therefore, should be supported by logic and the criterion of practicality.

If we accept that the above six factors are essential in the context of language learning and should be kept in check, the first step will be deciding on the acceptable ranges for scores and relevant statistics where these are available. For example, for proficiency test we can decide the range for subjects' scores to be between 30 and 70 and the range for mean scores between 40 and 60. The first range will be used to either include students into or exclude them from the study. The range defined for mean scores is but a factor which will be used in comparing groups. The overall score calculated for similarity criterion, however, will be obtained only by putting together the difference values calculated for all essential factors. Therefore, using a 10 point scale for each factor, we will consider the groups to be similar or comparable if their difference regarding each factor is not more than 10 and the total difference score does not exceed 60 (since we have decided on controlling for six essential factors).

It is also possible, to assign weightings to factors and categorize them from the most to the least important or at least tell our readers which factor or factors have been more important to us. It should not be forgotten that, the ranges defined for more important factors should be narrower than the ranges defined for less important ones.

Suppose we want to start convenience sampling or select two intact groups for a study which is designed to investigate an issue related to SLA. Let us assume that we have decided on the above six elements, i.e. mean and standard deviation of language proficiency test, age, language background, years of schooling, and gender as essential factors. To identify our groups, first we should decide on the degree of acceptable variation in scores. The second step will be applying the same 10 point scale to the differences in mean scores obtained from comparing differences in essential factors between groups. The final judgment as to whether the groups qualify as similar and therefore comparable or about the degree of their similarity, therefore, will be made based on four types of information:

1. whether the scores of all subjects in experimental and control groups fall within the ranges defined for subjects' scores,
2. whether the mean values of essential factors fall within the limits defined for mean scores,
3. whether the differences calculated for nominal variables like gender and language background fall within the defined ranges, and
4. whether the overall value calculated for similarity criterion exceeds the maximum number possible or not.

If our groups are substantially big, we can comfortably exclude subjects whose scores are not within the range to homogenize our groups. Excluding a few subjects from big groups will not do much damage to the reliability and validity of our study. So, the regulation that the scores of all subjects for essential factors should be within the defined ranges does not mean that we should relinquish studying if there are a few outliers. With relatively small groups reporting the values for ranges will help the readers of our research results know the circumstances within which the results were obtained. A detailed report of values obtained for each essential factor will give our readers some idea of the aspects of groups' makeup and the extent of their similarity.

It is clear that, every researcher has an idea about the level of language proficiency and the amount of variation in the groups he/she wants to study before leaping into the inquiry. Therefore, not every pair of mean scores, even if they are the same, can be useful. The mean scores may be too high or too low, though equal, for the purposes of our study. Clearly, it is possible to define wider ranges for the scores if we want to include more students in our study or have a wider standard deviation. If so, however the similarity criterion will suffer to some degree and the validity of our findings will diminish since we are not comparing two closely similar groups. The more the range of subjects' scores approaches the number of points on the scale, the more rigorous our study will be, but in actual experiments this will be a very hard-line approach to adopt.

A thorough example would clarify this. If the range of our accepted mean range is 45 – 55 in a 100 point proficiency test, knowing that all students' scores are within the range defined for scores, for example 30 – 70, we can assign 1 negative point to each difference point in the mean values of our experimental and control groups. So, two groups can be considered as similar if the subjects' scores are within the range, i.e. between 30 and 70, and the groups' mean values are not smaller than 45 or bigger than 55. The ideal situation is where both groups have the same mean score within the expected range. But if the mean scores differ, 1 negative point will be given per each point of difference to the groups' similarity criterion originating from the essential factor of mean scores. If the mean scores have decimal points, the numbers can be rounded off to the closest aggregate numbers and then the values be assigned.

For the standard deviation, we follow a similar path. We may decide that the acceptable standard deviation for our groups, to be considered as similar, is within +/- 1 SD. Again, groups whose standard deviation fall out of this range cannot be considered as similar since the dispersion of scores in the groups is not analogous for them to be considered as similar. The degree of similarity will be determined, as with mean scores, by assigning negative points to the differences. We divide the distance between -1 and +1 standard deviations to 10 points and compare the groups accordingly. Therefore, if, for example, our experimental group has a SD of -0.80 and our control group a SD of +0.80 the negative score given to their similarity criterion will be 8.

The age factor is another important factor which researchers largely wish to control in educational research. The same procedure will be applied to the essential factor of age. For example, if our expected age range is between 18 and 30, first, all the students should fall within this range and the outliers be excluded from the study. The second step would be checking if the mean values calculated for age in experimental and control groups are within the range, say, 22 to 27. A negative point, then, will be given to the similarity criterion for each half a year of difference between the average age numbers of experimental and control groups.

Language background may be taken to be another defining or essential factor. On some occasions the mother tongue of all subjects, in both experimental and control groups, is the same. Or, the compositions of groups are so that each group has a similar number of subjects belonging to particular linguistic backgrounds. On most occasions, however, subjects would come from different linguistic backgrounds. The former situation is the ideal state which enables the researcher even to ignore language background as an essential factor. The latter, nevertheless, should be subjected to the above procedure of assigning negative values against the criterion of similarity. That is, per each linguistic background in the experimental or control group with no parallel in the corresponding control or experimental group one negative point should be given to the similarity criterion. Thus, if our subjects in the experimental group, for example, come from three linguistic backgrounds, but our subjects in the control group are from seven linguistic backgrounds five of which are different from the linguistic background of our subjects in the experimental group, 5 negative points will be given to the similarity criterion. It is clear that two classes with completely different linguistic backgrounds are not comparable in this sense unless the purpose of study is to investigate only this aspect of difference and its possible effect(s) on the performance of subjects.

Years of schooling is another important factor that according to many may affect the subjects' linguistic gain. Subjects, especially adult second language learners, at higher educational levels, who have spent longer years in educational environments in their home countries, are generally faster and more adept in learning a second language. Therefore, it is necessary to take this defining factor into account. The number of years, as in other defining factors introduced above, is a matter of taste and depends on the purpose of the study, but our 10-point scale will be applied to all essential factors uniformly. Therefore, if we limit the mean years of schooling to 12.5-17.5 for example, we will assign 1 negative point to each half a year of difference in the groups' mean years of schooling. It is obvious that the years of schooling for all students in both groups should be restricted to the defined number of years of schooling, as with the range of scores on proficiency test or subjects' age, and the outliers be excluded from the study. So, a group will make a comparable group with another if its mean years of schooling is not more or less than the upper and lower limits defined for mean years of schooling and none of the students' years of schooling falls out of the predetermined

range for subjects' educational experience. If these two conditions are met and a group's mean years of schooling is not equal to the other group's mean years of schooling, a negative point will be assigned to the criterion of similarity for each half a year difference in mean years of schooling. The fractions, if there are any, as before, can be rounded off to the closest aggregate numbers.

Deciding on a scale for gender is a somewhat thorny issue since it is a dichotomous variable. However, in the current study, we are not dealing with the gender factor as a dependent or an independent variable. Like linguistic background, we are mainly concerned with the composition of our groups at this stage and simply take into account the number of male or female students in each group. The ideal situation is where the groups are all-male, all-female, or mixed groups with equal numbers of male and female subjects. In these cases the difference number or the value assigned against the similarity criterion will be 0. As with language background, this state will convince the researcher to overlook this factor as an essential factor and look for other important factors that may bias the results. But, the hard fact is that, seldom equal combination of males and females is available. This is while, the equal proportion of male or female subjects is essential for the design of many studies in educational field. That is why, most researchers, to exclude the moderator factor of gender, limit their subject selection to one gender only. With this procedure, however, it is possible to address the issue of gender by again assigning negative points to any difference in the number of males and females between our experimental and control groups. Then, there is no need for restricting our groups to one gender only. If our groups are constituted of 15 subjects each and, say, the experimental group has 13 males and 2 females, but the control group has 10 males and 5 females the negative points to be given to the similarity criterion will be 3 regardless of whether we count males or females.

But, on many occasions the number of subjects in the experimental and control groups are not the same, i.e. the groups are not balanced. What shall we do in cases like this? One solution may be to define ranges for the differences in subjects' gender. Another way out of the dilemma would be to find the percentages of male and female subjects in each group and assign negative points to the differences in percentages instead of the simple frequency of students. The first step, then, would be deciding on the acceptable range of difference in percentages. If we happen to decide on this range to be 20%, for example, we can give 1 negative point per each two percent of difference in the composition of the groups. The second step is to compare the actual groups and assign relevant negative point(s) to the similarity criterion. Let us adopt the second procedure here, not to make the number of essential factors unwieldy.

For example, if we have 15 subjects of 10 males and 5 females in our control group and 19 subjects of 11 males and 8 females in our experimental group, it will be easy to determine the percentage of males and females in each group and then compare the percentage values to assign necessary negative point(s) to the similarity criterion. If we work through the control group the percentages for the males and females in this group will be obtained as follows:

100 : 15 = 6.6
6.6 x 10 = 66 male out of one hundred subject
66 : 100 = 66% percentage male
100% - 66% = 34% percentage female

The percentage values for the experimental group will be:

100 : 19 = 5.3
5.3 x 11 = 58 male out of one hundred
58 : 100 = 58% percentage male
100% - 58% = 42% percentage female

Whether we look at the percentages of males or females in the two groups makes no difference; in any case the difference value calculated will be the same. This difference value will tell us which negative value to assign to the similarity criterion if, of course, this value happens to be within the range. For the above percentages the negative value to be assigned to the similarity criterion will be 4, since the difference adds up to 8%.

Violating each of these ranges would mean our groups are dissimilar and therefore not comparable. However, this does not mean that researchers should desist studying simply because the numbers they have calculated for essential factors contradict predetermined values. If the numbers of subjects in both control and experimental groups are big enough, which is the favorable situation, as it was said above, we can simply exclude students who do not meet the requirements of this procedure from the study. If we have a limited number of subjects, on the other hand, reporting the values would suffice for our readers not to go astray in their interpretations. The conclusions that we draw based on our research findings, on the other hand, should be very cautious and not claimed to be obtained from comparing similar groups. These kinds of findings would only suggest tentative possibilities which need to be approved by studying more similar groups.

Another point which comes to mind is that, replication is a relative term. The more similar our experimental and control groups are in repeated measures of the same dependent variable or variables with the same instrument the more the possibility of obtaining similar results and the higher the reliability and internal validity of the findings will be. It is obvious that, in humanities exact replications are impossible. But, with less similar groups with outcomes similar to the initial study our external validity will increase. So, if like similarity we think of replication as a continuum, we can say that close approximation of groups' attributes is what is desired particularly in situations where the findings are controversial. With less controversial issues we can feel relatively relaxed using less similar groups. In any replication,

therefore, the study can be repeated with different groups or in different settings, but depending on the purpose of study we can approach the reliability and internal validity end of the continuum with manipulating very similar groups or the external validity end with less similar groups. Findings obtained from comparing both more and less similar groups, which correspond each other closely, will be highly reliable and will have internal and external validity both at the same time. In both cases, but, the essential factors affecting the performance of subjects should be within the defined ranges and the values calculated for essential factors and similarity criterion be reported to project a clear picture of circumstances in which the study was conducted.

It seems reasonable, therefore, to recommend researchers firstly to choose as big groups of samples as possible for their research purposes of which outliers can be excluded without doing much damage to the research design and secondly, replicate studies with maximally and minimally similar groups to be able to generalize their findings. This does not apply, however, to those researchers whose research objectives necessitate working with smaller groups or even individuals as in case studies.

Accepting the differentiations made above, we can say that we will have six scales of ten for educational studies with a total difference number getting at 60 and we will accept groups to qualify for convenience sampling if they fall within the ranges defined for these factors. So, even if two groups gather an overall number which is smaller than 60 but violate one or more ranges defined for essential factors we will not be able to consider them to be comparable, unless the students falling above or below the limits are excluded from the study. Looking at the overall number obtained for the similarity of our experimental and control groups, we can say to what extent they are similar. Actually, we can look at the similarity criterion too as a continuum of the most to the least similar. The smaller the overall number calculated, the more similar the groups will be. The bigger the number, on the other hand, the less the groups will resemble each other. This is a number which will give us a rough idea about the reliability and validity of our study as well, since the more similar the two groups are, the more confident the researchers will be in the claims they make. The diagrammatic representation of the similarity criterion can be shown as below:

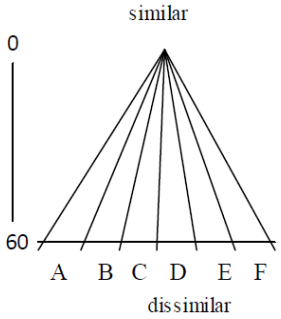

Figure 1. Similarity Continuum

In Figure 1, the more we go up toward 0 the more similar the groups will be. And, the further we approach the base of the diagram, the more our groups will diverge to the extent that after 60, our groups cannot be considered as similar. Drooling for the dreamy value of 0 might be an unrealistic expectation. Nonetheless, our attempt as researchers in the educational field should be approaching this value to the extent possible if we want to have reliable and valid outcomes at the same time.

Another point we understand from Figure 1 is that the essential factors are all of the same importance or weight to the researcher. If, as it may be the case, a researcher desires to attribute differential importance or weightings to the essential factors, the lengths of A, B, C, D, E, and F, which represent essential factors, will change depending on the importance assigned to them. It is clear by now that, shorter distances will be indicative of higher importance of the factors. Even, we can assign hierarchical importance to our essential factors and arrange them from the most to the least important. Obviously, our scale should not change, even if the lengths differ. We will divide each distance between two limits of a factor by ten points despite the limitation put on the range to be able to uniformly assign negative values against the similarity criterion. Figure 2 shows a situation in which the researcher has assigned hierarchical importance to different essential factors.

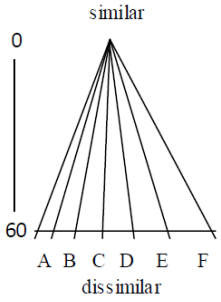

Figure 2. Differential hierarchical importance assigned to essential factors

At first glance, the application of this method to convenience sampling may seem daunting to most researchers for the limitations it imposes upon their work. But, applying this procedure in practice does not appear to be very difficult. On the other hand, the resulting confidence build up regarding the findings makes it worthy of serious consideration. Two issues of prime importance in working with this procedure are:

1. Applying logic to the process of determining the number of essential factors and their ranges, and
2. Starting studies with big enough groups wherever possible.

The application of logic to the process of determining ranges is a matter which is directly related to the researchers' degree of meticulousness. The deliberation should not however, surpass the issue of practicality. In any case doing something for the sake of knowing is better than doing nothing under the excuse of inability in meeting the requirements. Starting with big enough groups will guarantee the success of the study to a great extent as it gives us the leeway to exclude from our study subjects whose presence would damage the homogeneity of our groups and the validity of our findings.

Deciding on the ranges at the beginning of the study, we can specify conditions and make tables like Table 1 below which will direct us all through the sampling process. This table will represent the degree of rigor that the researcher wishes to exert and will tell him or her which subjects to include in or exclude from the study. Additionally, readers will be able to decide on the degree of faith they should keep with the findings. Deciding on the conditions and criteria is the first step to be taken and precedes the construction of Table 1. For the above essential factors we can specify the following conditions and criteria where As stand for conditions and Bs for criteria.

TABLE 1.
RANGES OF ACCEPTABLE INDIVIDUAL AND GROUP SCORES

A. Range of individual scores on language proficiency: 30–70 B. Range of scores' mean on language proficiency: 40–60 ⟶ each two scores of difference = 1 negative point
B. Range of standard deviation: +/- 1 SD ⟶ each SD difference of 0.2 = 1 negative point
A. Age range: 18–30 B. Range of age mean: 22–27 ⟶ each half a year of difference = 1 negative point
B. Acceptable between-group difference in gender: 20% ⟶ each two percent of difference = 1 negative point
B. Language background range: 1–5 ⟶ each difference in language background = 2 negative points
A. Years of schooling range: 10–20 B. Range of years of schooling mean: 12.5–17.5 ⟶ each half a year of difference = 1 negative point

It is clear that for our groups to qualify as similar the first requisite is that conditions A are fulfilled. If these conditions are met for both experimental and control groups, then the fulfillment of criteria B will convince us of having two similar groups.

Table 2 below is made to obtain the overall degree of similarity (or dissimilarity) between two imaginary groups for which the above conditions and criteria apply. The degree of similarity is, of course, determined by the difference value which is presented at the rightmost end of the table bottom. Smaller difference values will be indicative of more similarity while a big value will be a sign of less similarity.

TABLE 2.
DIFFERENCES IN MEAN VALUES OF ESSENTIAL FACTORS

	mean language proficiency	SD	mean age	language background	mean years of schooling	gender	
experimental	52	+0.80	22	2	14	65% male	
control	48	-0.80	23	3	11	55%male	
Difference score	4	8	1	1	3	5	Total = 22

Since, all subjects in both experimental and control groups have gained or given scores which are within the defined ranges for essential factors, i.e. they have met conditions A, and the differences in the mean values for essential factors do not exceed the defined ranges, i.e. criteria B are fulfilled, and since the overall difference value obtained for these two groups is smaller than 60, we can confidently assert that these two groups are similar enough to be compared as intact groups.

Of course, some of these factors would not be considered as essential in other contexts. For example, in many EFL situations students generally come from the same linguistic background. It is clear that in situations like this linguistic background can be forgone, as mentioned above, and other important factors like aptitude be included in the list. In any case, deciding on the kind of essential factors is dependent on the individual researcher and the context in which the research is carried out. But, the number of essential factors is a matter that researchers can agree on. Our suggestion is that in any research using convenience sampling at least five essential factors be taken into account and the values calculated for them for both experimental and control groups reported.

Returning back to our discussion of randomizing and the claim that it minimizes the effects of extraneous and subjective factors, we want to say that this procedure, if employed properly, is even more capable of eliminating the effects of irrelevant factors from the study. Another advantage of this procedure is that, by reporting the values for conditions, criteria, and the overall difference the readers will be aware of the compositions of groups and the context in which the study was conducted and, therefore, the comparisons made between the research results will be theoretically sound.

V. Conclusion

Doing research in humanities has its own problems. Unlike physical sciences, in humanities and particularly in educational field we are not dealing with static materials or consistent states. Humans' abilities, states of mind, attitudes and motives are in constant change making the prediction of their behavior extremely hard. Applying the notion of subjectivity to testing, Bachman (1995) quotes Pilliner (1968) that language tests are subjective in nearly all aspects. The same is true about researching second language acquisition which uses tests as an instrument for data collection purposes. However, subjectivity in research does not originate from the subjective nature of tests only. The major problem in carrying out any research in educational field is the compromise that should be reached at and the balance that should be created between what happens in real-life and what would happen under controlled conditions. This kind of compromise is inevitable if we are not going to sacrifice internal validity for external validity or the other way round. However, an important point when it comes to research is what Maxwell (1992) calls descriptive validity. Although, Maxwell discusses the issue of validity in the context of qualitative research, leaving an audit trail, as Dörnyei (2007) puts it, should be an integral part of any kind of research. This audit trail is perfectly applicable to the sampling stage of quantitative research as well, especially where non-random groups are used for research purposes. Describing the conditions under which the investigation was carried out removes a lot of misinterpretations of or overreadings from the research results. The purpose of this article is not to support imposing uncalculated control on research projects. Its purpose, on the contrary, is to encourage reluctant researchers to conduct inquiries even though they feel that they do not have access to comparable groups. The only thing required, however, is to precisely report the circumstances in which the research was conducted. We believe that the guidelines suggested here are flexible enough to encourage interested researchers to give their ideas a good try. But, the choice of essential factors and the ranges within which the researcher will operate should be logically sound and defensible.

References

[1] Bachman, L. (1995). Fundamental considerations in language testing. New York: Oxford University Press
[2] Best, J. W. & Kahn, J. V. (2006). Research in education. Boston: Pearson Education Inc.
[3] Brown, J. D. (1995). Understanding research in second language learning. New York: Cambridge University Press.
[4] Dörnyei, Z. (2007). Research methods in applied linguistics. New York: Oxford University Press.
[5] Hatch, E. & Farhady, H. (1981). Research design and statistics for applied linguistics. Newbury House Publishers, Inc.
[6] Hatch, E. & Lazaraton, A. (1991). The research manual: Design and statistics for applied linguistics. New York: Newbury House Publishers.
[7] Larson-Hall, J. (2010). A guide to doing statistics in second language research using SPSS. New York: Routledge.
[8] Mackey, A. & Gass, S. (2005). Second language research: Methodology and design. New Jersey: Lawrence Erlbaum Associates, Inc.
[9] Mangal, S. K. (2002). Statistics in psychology and education. New Delhi: PHI Learning Private Limited.
[10] Maxwell, J. A. (1992). Understanding the validity in qualitative research. *Harward Educational Review. 62(3),* 279–300.
[11] McKay, S. L. (2006). Researching second language classrooms. New Jersey: Lawrence Erlbaum Associates, Inc.
[12] Muijs, D. (2004). Doing quantitative research in education. London: Sage Publications.
[13] Robson, C. (1993). Real world research. Massachusetts: Blackwell Publishers Ltd.
[14] Seliger, H. W. & Shohamy, E. (1989). Second language research methods. New York: Oxford University Press.
[15] Tailor, G. R. (Ed.). (2005). Integrating quantitative and qualitative methods in research. Maryland: University Press of America Inc.
[16] Walliman, N. (2011). Research methods: The basics. New York: Routledge

Farahman Farrokhi received his PhD in English Language Teaching from Leeds University, England. Currently, he is an associate professor at the University of Tabriz. His research interests include classroom discourse analysis, feedback types, negative and positive evidence in EFL classroom context and research methodology.

Asgar Mahmoudi-Hamidabad is currently a PhD candidate of ELT at the University of Tabriz. This article is extracted from his PhD dissertation to be presented in Tabriz University. His research interests include oral fluency and naturalness, theories of language learning and teaching, and research methodology.