*Hong Kong Teachers' Centre Journal*《香港教師中心學報》, *Vol. 11*


# *If wrong, fix it: Case studies of statistical misapplications in school-based projects*

# 如果錯了，改正它：校本研究統計錯誤的個案

**Kay Cheng SOH**

*Singapore*

## Abstract

Statistics have been used extensively in many school-based projects. Unfortunately, misconceptions have often been found in the statistical reports. In this paper, five case studies were used to illustrate some common conceptual and procedural errors found in reports and how these problems could be rectified.

## Keywords

educational statistics, effect size, experimental design, school-based project

## 摘要

許多校本研究採用統計分析。不幸地，統計數據的誤用和誤解並不少見。本文討論研究統計在概念上和程序上常見的一些誤用例子，並以五個個案說明如何在校本研究避免發生統計誤用。

## 關鍵詞

教育統計，效果强度，實驗設計，校本研究

> *Sometimes, statistical evidence that looks good on the surface nevertheless turns out to be flawed. Broadly, there are two ways in which an argument based on apparently persuasive evidence might lose its impact after further inspection. One possibility is that the data were mishandled or the statistical analysis was misapplied. The second possibility is the discovery of some artifacts in the research procedure, such that the substantive conclusions drawn by the investigator may not logically be warranted by the observational comparison made* (Abelson, 1995, p.78).

Abelson (1995) begins his chapter On Suspecting Fishiness with the above quote. It is an apt reminder that if a project reports statistics, it does not guarantee that the awesome figures have been derived correctly and interpreted validly. Errors of the first kind are procedural and technical which are easier to be noticed. Errors of the second kind are conceptual and interpretative and are therefore more difficult to detect. There may be cases of pure conceptual or procedural errors, but more often than not the errors are confounded as the examples below show. This paper deals primarily with the first kind of errors.

## Case study no. 1: Misplaced comparisons

### Group comparison

It is a very common statistical application in school-based projects where a group of students receiving intervention or treatment is compared with a group not receiving it. There is a need to ensure that the groups are equivalent before intervention. Then, a sizeable difference in favour of the treated group is expected after the project is completed. The box below shows the relevant information in the original report of a school-based project. For obviously reasons, the source of the case is concealed with no references made. This will be done for the other cases discussed later, too. In fact, it does not matter who made the errors; they are just what they are, that is, errors to be rectified.

> A quasi-experiment was conducted where pupils from one class formed the CG and pupils from a second class formed the EG. In this study, both classes were kept intact without randomization. A perception survey on self-esteem was conducted.
>
> As shown in Table 1, for the pre-survey, the EG scored a mean of 34.13 and in the

post-survey, a mean of 38.27. This may suggest that the intervention has made an improvement in terms of their self-esteem in the post-survey. The control group has scored a mean of 33.79 and in the post-survey, a mean of 34.58, indicating that the difference in the pre-survey and post-survey on self-esteem for the CG is very small. The Levene's test of 0.995 ($p>0.05$) was not significant, further showing that the two groups are homogeneous at pre-survey, before the intervention. The Cronbach's alpha value for the survey questions was calculated as 0.892, showing that it is an adequate reliability coefficient.

Table 1: Comparison of pre-survey and post-survey of EG and CG on pupils' self-esteem

| | Pre-survey | | | | Post-survey | | | |
|---|---|---|---|---|---|---|---|---|
| | N | Mean | SD | Levene's test | N | Mean | SD | Levene's test |
| Experimental | 30 | 34.13 | 6.704 | 0.995 | 30 | 38.27 | 5.527 | 0.169 |
| Control | 24 | 33.79 | 6.379 | | 24 | 34.58 | 7.040 | |

In this case, the experimental design is fine: there were two intact classes serving as the project and the comparison groups and, since the pupils were not randomly assigned, it is correctly described as a quasi-experiment. The teacher-researchers first mention the improvement in self-esteem means of the project group and then the very small change in the same measure of the comparison group. They further mention the non-significant difference in group homogeneity, citing the result of the Levene's Test (*Levene's Test, n.d.*). Finally, the Cronbach's alpha coefficient (*Cronbach's alpha, n.d.*) is reported.

**Correct way of comparing groups**

When reporting a project using an experimental design involving two groups, it is important to be clear about *what is to be compared with what for what purpose.* First, it is necessary to check group equivalence in the criterion measure (in this case, self-esteem). This needs to cover two aspects: (1) mean difference to see if the groups are comparable with regard to the criterion, and (2) difference in the standard deviations (SDs) to see if the groups are homogeneous in the criterion. If no differences are detected for both the means and the SDs, then the groups are taken to be equivalent. Such pre-project equivalence ensures that should a difference be found after project completion, the difference is not due to the initial difference (since there is none) but something else; and, the intervention is a strong candidate accounting for the post-project difference. In the present case, the pre-

project mean difference of 0.34 was not been formally tested, although it may be claimed that the difference is too small to need statistically testing. Instead, the result of the Levene's test which tested the difference in the two variances (the square of the standard deviations) was mentioned; obviously, the teacher-researchers used the Levene's test as if it is a test of mean difference, which should be tested with the Student's *t*-test.

To rectify, first compare the two groups' pre-survey means with the independent *t*-test and hope for no difference. (We leave the question of group homogeneity for the time being since the SDs 6.70 and 6.38 are close enough.) When group equivalence is assured, then, do the same to compare the groups on the post-survey means and hope for a statistically significant difference this time. If this is obtained, then the project groups can be said to have benefitted from the intervention. The effect size (Coe, 2002; Soh, 2008) used here for group comparisons is one version of the *standardized mean difference* (SMD), specifically the Glass's *delta* (Soh, 2008). The SMD was simply calculated by (Project mean – Comparison mean) / (SD of the comparison group). When this was done, Table 1 was re-structured as Table 2.

Table 2: Mean comparisons of pre-survey and post-survey

| Measure | Project (N=30) | | Comparison (N=30) | | Difference | SMD |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | | |
| Pre-survey | 34.1 | 6.70 | 33.8 | 6.38 | 0.3 | 0.05 |
| Post-survey | 38.3 | 5.53 | 34.6 | 7.04 | 3.7 | 0.53 |

Now, the pre-survey SMD of 0.05 in Table 2 shows that the two groups were equivalent in the criterion before project commenced, and the post-survey SMD of 0.53 shows a medium effect size in favour of the project group, thus the project was successful in producing a difference which cannot be ignored or dismissed.

**Why not the t-test?**

The teacher-researchers reported the results of the Levene's Test but not those of the *t*-tests. Why this is so is not known. The results of the *t*-test and the Levene's test appear together in the same run of the *t*-test in the *Statistical Package for Social Sciences (SPSS)* which could have been used by the teacher-researcher. Anyway, this is a blessing in disguise because the *t*-test should *not* have been run in the first place! Oftentimes, teacher-researchers routinely run this Null Hypothesis Significance Test (NHST) to compare group means (either doing this on their own accord or, perhaps more often than not, being misguided).

There are several reasons why the *t*-test should not have been run. First, what question do the teacher-researchers attempt to answer? Generally, this is about whether the project mean differs from the comparison mean, and if yes, what is the *magnitude* of the difference. For pre-test, small or no difference is hoped for because group comparability is desired. For post-test, medium or large difference is expected to show the intervention effect. The answers to such questions are found by using the SMD, and not the t-test, for the simple reason that the t-value does not answer the question on magnitude of difference.

What then does the *t*-test do? It tells the *probability* of an observed difference in the populations, and this is not the concern of the teacher-researcher doing a school-based project. For this, we need to quote Abelson again:

> *There is also a common confusion when using the significance level as an indication of the merit of the outcome. When the null hypothesis is rejected at, say, the .01 level, a correct way to state what has happened is as follows: "If it were true that there were no systematic difference between the means in the populations from which the samples came, then the probability that the observed means would have been different as they were, or more different, is less than one in a hundred. This being strong grounds for doubting the viability of the null hypothesis, the null hypothesis is rejected* (Abelson, 1995, p.40).

Note that the *t*-test it is *not* about the *magnitude* of group difference (which is of concern to the teacher-researchers) but about the *probability* of the observed group difference as an estimate of a similar difference in the *populations* (Fraley, 2003). Since when does a teacher-researcher become concerned with what may or may not happen to a very large group of other teachers' pupils who made up the populations? In practically all cases like the present one, there is hardly real sampling in school-based projects and, to call the groups of pupils *'samples'* is in fact a misnomer or misconception, or both. Since there is no real sampling (and hence no samples), inferential statistics like the *t*-test is irrelevant and therefore not applicable. Therefore, descriptive statistics such as the SMD is the only one to use with validity. On the misuse of the *t*-test and its like, Abelson (1995) has a strong view, thus,

> *The ethos of doing significance tests as the hallmark of an appropriately conservative style is now so deeply ingrained that tests are sometimes used even when they need not be. Indeed, there are several contexts in which it is really silly (Cohen, in press) to carry out a significance test, much less to present its result* (p.76).

There is yet another reason why the t-test cannot be trusted to compare group means, even if it is used for comparing samples which have really been randomly selected from their respective populations. The problem is the influence of sample sizes on the *p*-value corresponding to a *t*-value. Let's say a $t$=1.99 is obtained for comparing two groups which have together 42 pupils, the corresponding p-value is not significant ($p$>.05) as the required *t*-value is 2.02. But if the total number of pupils is 82, the same *t*-value (1.99) is significant ($p$<.05) because it is equal to the required $t$=1.99.

In a recent issue of a journal, a study reports almost all comparisons as non-significant and another almost all as significant. Of these studies (references cannot be given to safeguard the authors), one is too bad to be true while the other too good to be true. A careful look shows that the former study compared 10 pairs of respondents whereas the latter has a total respondent size of as many as 800! These are good contrasting examples of the influence of sample sizes on the results of the *t*-test. In the words of Sterne (n.d.), "Given a large sample size, even a small difference will be statistically significantly different from zero."

In the case study above, the teacher-researchers compared first the project group's pre-post-test means and then, likewise, the comparison group's pre-post-test. In other words, they did two separate within-group comparisons and then inferred from the results that there was a project effect.

This seems fine intuitively but doing so violates the logical of the experimental design used. On this, we have to listen to Abelson (1995) again:

> *But that would contradict the logic of including a control comparison in the first place.* Why is that so? *The point of running a control condition is to test the relative claim that the effect in the presence of the experimental factor exceeds the effect in its absence. The appropriate test seems to be a test of the* interaction *between the rows and the columns* (p.63).

Why do teacher-researchers make this kind of conceptual error? One possibility is that teachers typically are concerned with student's *improvement* which is always seen as a difference in performance *before and after* teaching the same students. This mode of thinking is consistent with commonsense exemplified by watching a plant or a child grows. It is a mode of thinking teachers developed over years which is difficult to change when change is necessary as they do school-based projects experimentally. Whatever the cause, teacher-researchers need to re-orientate and adopt a research mode of thinking when

analyzing and reporting school-based projects.

**The Levene's test**

When the *SPSS* is run to compare group means, the Levene's test is first done by default to check homogeneity in variability. The result shows whether the two groups have the similar or different degree of homogeneity. If there is non-significant difference (as was found for the present case), the "*equal variances assumed*" *t*-value is taken, otherwise, the "*equal variance not assumed*" *t*-value should be reported. Once the question of homogeneity is settled, the researcher will proceed to use the appropriate *t*-value and report the outcome of group comparison.

What does the Levene's test do? According to the Wikipedia (2010),

> *In statistics, Levene's test is an inferential statistic used to assess the equality of variances in different samples... It tests the null hypothesis that the population variances are equal. If the resulting p-value of Levene's test is less than some critical value (typically 0.05), the obtained differences in sample variances are unlikely to have occurred based on random sampling. Thus, the null hypothesis of equal variances is rejected and it is concluded that there is a difference between the variances in the population.* (Emphasis mine)

Note that the Levene's test (*Levene's Test*, n.d.; Wikipedia, 2010) is an *inferential* statistic for checking equivalence of variances of two or more *randomly* selected groups. Therefore its application in the present case is doubted, since the two groups are not random samples. By the way, *variance* is numerically the square of SD indicating the extent with which a set of scores spreading around its mean. When a group has a SD (and therefore a variance) much larger than another group has, its scores are spreading much wider, indicating there are more higher or lower scores or both. If this is the case, then the two groups are not equivalent on homogeneity, although they may have the same or similar means indicating the same or similar level of performance. Then something need be done to ensure group comparability before comparison is made on relevant measures (Soh, 2009).

In place of the Levene's test, a simple shortcut is to find the ratio of the two variances. This is done by (1) finding the variances by squaring the two groups' SDs, (2) dividing the larger variance by the smaller one, and (3) checking the ratio against the tabled value of the F-distribution which can be found in the appendix of any text on statistical analysis.

To teacher-researchers, the first two steps are no problems, but the third is a bit clumsy. As a rule of thumb, if there are about 30 or more pupils in each of the two groups, and if the variance ratio is less than 2, the groups can be taken to have the same or similar homogeneity.

**Cronbach's alpha coefficient**

A *Cronbach's alpha coefficient* (n.d.) of 0.892 is reported for the case study. This is far greater than the conventional expected minimum of 0.70 for research purposes (Siegle, 2002). While the teacher-researchers deserve to be congratulated for this, there is also the need for more information to understand what the coefficient means. The only relevant information in the report is "*A perception survey was conducted to ascertain pupils' self-esteem. The survey was designed such that questions of the same nature were repeated but they were phrased in different ways.*" It is not clear what different aspects of self-esteem were covered in the survey and the re-phrasing of the same items might have contributed to the unusually high alpha coefficient.

Also needed is the number of questions in the self-esteem survey. Number of item affects the alpha coefficient. Cronbach's alpha formula has two multiplicative components: (1) reliability component and (2) correction factor. The first is [1 – (Total item variance) / (Test variance)] which is the total test variance minus the unreliability portion. The second is k / (k-1) where k is the number of items; it 'corrects' the reliability component for number of items. If a test has a reliability components with a coefficient of 0.60 (which is quite a normal figure for affective measure like the self-esteem survey), and if there are 10 items, the Cronbach's alpha coefficient is 0.67, corrected up by 11%. If the test has only 5 items, it is 0.75, adjusted up by 25%. For three items, it is 0.90, adjusted up by 50%. And, if there are only two items, it is (2/1) * (.60) = 1.20 > 1.00, an alpha coefficient indicating that the test scores are more perfectly reliable than perfect reliability! Of course, this does not make good sense. Here, the paradox is that the shorter the test, the higher the score reliability appears to be, leading to over-confidence in short tests, contrary to the normal expectation that the longer the test, the more reliable the scores will be, given the same quality of items.

For the present case, an important question is for which set of data was the Cronbach's alpha coefficient obtained. There are six possibilities: (1) pooled pre-survey, (2) pooled post-survey, (3) project groups' pre-survey, (4) project group's post-survey, (5) comparison group's pre-survey, and (6) comparison group's post-survey. Alpha coefficients

calculated using these different sets of scores will yield different results and have different meanings. Considering the experimental design, pooled pre-survey scores are the best to use as they are not influenced by the intervention which may make the project and the comparison groups different in their self-esteem. It will be good if there is an indication of which sets of scores were used for estimating the internal consistency of the self-esteem survey.

## Case study no. 2: Missing standard deviations

For Case Study No. 2, the box below is an extract from another project report which deals with many aspect of student engagement in learning. The analyses done as reported by the teacher-researchers are the same for difference measures, only one (Vision of learning) is cited for illustration.

> The general methodology adopted was a two-group (experimental and control) pre- and post-programme quasi-experimental design… The Null Hypotheses were (1) FSP [the programme] did not increase the level of student engagement; and (2)…
>
> Table 1 (here, re-numbered as Table 3) shows the paired t-test analysis carried out on the means obtained by the two groups in the pre- and post-surveys. The experimental groups registered significant increase in all categories. The control group contained significant increase in the scores in five out of nine categories…
>
> Table 3: Comparison of mean scores between experimental and control groups

| Components | Experimental Group (EG) | | | Control Group (CG) | | |
|---|---|---|---|---|---|---|
| | N=74 Pre-E (a) | N=74 Post-E (b) | Difference (b) – (a) | N=73 Pre-C (c) | N=73 Post-C (d) | Difference (d) – (c) |
| Vision of learning | 15.59 | 16.68 | 1.09*** | 15.27 | 15.81 | 0.54 |

> … (the programme) showed a significant increase in 'Vision of learning'… There was a 1.1 point increase for the Experimental group, while the Control group mean increased by 0.5 point.

As is true of Case Study No. 1, this one also used within-in group comparison. As this is illogical (Abelson, 1995, cited above) as is the previous case, the data need be re-analyzed. However, the original table does not report the standard deviations (SDs) but only indicates the significance levels of differences using asterisks. To re-analyze, the information is re-organized for correct comparisons in Table 4, using *a, b, c*, and *d* to represent the missing SDs.

Table 4: Mean comparisons of pre-survey and post-survey

| Measure | Project (N=74) | | Comparison (N=73) | | Difference | SMD |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | | |
| Pre-survey | 15.59 | *a* | 15.27 | *b* | 0.32 | 0.32/*b* |
| Post-survey | 16.68 | *c* | 15.81 | *d* | 0.87 | 0.87/*d* |

It is a standard procedure that SDs are reported together with their respective means. But this was not done for this case. Had the SDs been available, the SMDs for the pre-survey and the post-survey can be obtained and will result in two Glass's *deltas*. Based on these, then, whether the groups differ in the pre-survey can be ascertained and the same can be done for the post-survey. As the needed information (*b* and *d* in Table 4) are missing, the SMDs cannot be calculated and there is no way we can make the comparisons. Is it, then, possible to do some guesstimate with the limited available data?

Fortunately, the report indicates that there are 15 items for measuring student engagement in four aspects, namely, Vision of learning, Tasks, Assessment, and Instruction mode. Looking at the patterns of the means in the original table for the various measures, it is possible that there are four items for Vision of learning. Since each item is a five-point scale, the lowest possible scale score is 4 and the highest 20. Armed with this information and assuming a normal distribution of the scores, the standard deviation can be estimated (*Estimating Standard Deviation,* n.d.), thus:

Estimated SD = (Largest possible score – Lowest possible score) / 6
= (20 – 4) / 6
= 16/6
= 2.7

If this is a correct guesstimate, then the estimated SMDs are 0.32/2.7=0.12 and 0.87/2.7=0.32 for the pre-survey and post-survey, respectively. Then, the conclusion is that the two groups were equivalent on the pre-survey and there was a small effect size

in favour of the project group on the post-survey. Therefore, the intervention was able to engage the project students slightly better than it did the comparison students. Of course, here again, although the conclusion is similar to that of the teacher-researchers, but the thinking process and logic are different: the teacher-researchers reached the correct conclusion but for a wrong reason!

## Case study no. 3: Missing initial comparison

An important condition of a two-group design is the initial group equivalence. This is necessary for a valid interpretation of the post-test difference, if any. In the box below for Case Study No. 3, an initial group difference was not taken into account when interpreting the post-test difference.

The subjects of this experiment are the students in two of the classes in Secondary Four… Both classes stayed intact without randomization… For 4E1, the students came in with an average T-scores of 213 in terms of their English language. For 4E2, the average *T-scores* is 194.

The pre-test was the English Language mid-year examination… The post-test was a test designed… to used a text-type… using Impact Analysis as the subject matter.

As shown in Table 5 (originally, Table 1) below, for the post-test, 4E1 has a mean of 20.46 (1.80) after the treatment compared to a mean of 19.28 (2.06) before the treatment. 4E2 shows a mean of 18.00 (2.51) after treatment compared to 17.26 (1.57) before the treatment. The paired *t*-test on the scores yielded a *p*-value of 0.002 and 0.015 for 4E1 and 4E2 respectively, indicating both classes showing increase in the scores which are significant. The increase in the mean scores was however greater in 4E1.

Table 5: Comparison on post-test

| | Pre-test | | Post-test | |
|---|---|---|---|---|
| | Mean | SD | Mean | SD |
| 4E1 | 19.28 | 2.06 | 20.46 | 1.80 |
| 4E2 | 17.26 | 1.57 | 18.00 | 2.51 |

As can be seen in the conclusion, the conceptual error of misplaced comparison appears in this case. However, there are two other errors which deserve rectification and discussion. By the way, this does not include the erroneous statement "*in terms of their English language*", because *T*-score is an aggregate for four subjects examined in Singapore's Primary School Leaving Examination. The fact is that only subject grades but not subject *T*-scores are available to the school. In this case, where did the *T*-score for English Language come from? Obviously, there is a mis-reporting.

In the case, the pre-test means of 19.28 and 17.26 could well be different enough for the two groups to be non-equivalent before project commenced. This is confirmed when the SMD of 1.29 was obtained. This is a very large SMD when checked against Cohen's criteria (Cohen, 1988; Soh, 2008). That the groups were initially non-equivalent could have been noticed by the teacher-researchers at the outset when they compared the average *T*-scores of the two groups.

Table 6: Mean comparisons of pre-test and post-test

| Measure | 4E1 (N=?) | | 4E2 (N=?) | | Difference | SMD |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | | |
| Pre-test | 19.28 | 2.06 | 17.26 | 1.57 | 2.02 | 1.29 |
| Post-test | 20.46 | 1.80 | 18.00 | 2.51 | 2.46 | 0.98 |

Comparisons should have been done to compare groups on the pre-test and then again on the post-test, instead of two separate paired *t*-tests for the pre-post-test difference within each group, for the reason expounded earlier.

When the two groups were compared on the post-test, the SMD of 0.98, which is large by Cohen's (1988) standard, shows a large effect size which the teacher-researchers hoped for. However, since the groups were non-equivalent to begin with, comparing them on the post-test without due consideration for the initial difference renders the conclusion suspect. In fact, while the project group gained by 1.18, the comparison group gained by 0.74; the difference in gain is 0.44 in favour of the project group. To avoid this conceptual problem, the groups could have been equalized first by using some of the methods such as winsorizing or caliper matching (Soh, 2009) to create equivalent groups for valid interpretation. Alternatively, a gain-score analysis could be employed to off-set the initial difference.

A third conceptual error lies with the tests used. As indicated in the report, the pre-

test and the post-test are two different tests. Had the groups been equalized on the *T*-score and then compared on the post-test between-group, the project would use the *equivalent group post-test only design* which in fact is a simpler and good design. Unfortunately, the paired *t*-tests compared the two sets of scores which are not of the same measure. The computer software is blind to the sources of numbers put in for processing; it does not know where the scores come from and does not need to know either. It just obediently churns out whatever statistics it is asked to calculate. It is the researchers who have to ensure meaningfulness of the statistics. Had the pre-test and post-test been the *same* measure, the problem of non-equivalence between groups can be solved by a gain-score analysis as suggested above. This case shows that experimental design, measurement, and statistical analysis of a project are not independent but related and they need be considered together.

## Case study no. 4: Over-simplification

Compared with the two previous cases, Case Study No. 3 is a more complex one. The project studied the effect of interdisciplinary project-work (independent variable) on students' perceptions of life-skills (dependent variable) and ascertained if there were differences attributable to course and gender (two moderating variables). A moderating variable is one which influences the relationship of the independent and dependent variables. As rightly stated by the teacher-researcher, there was no control group since the entire Secondary Two cohort was involved in project-work. Incidentally, this so-called whole-level approach is another issue in research design but the discussion of which is not within the score of this paper.

> The research question posed with regard to this investigation is: What is the impact of interdisciplinary project-work… on making learning meaningful?… The LSQ (Life-skills Questionnaire) administered as pre- and post-tests comprised statements to identify the perception aspects of life-skills. The questionnaire consisted of four components: (1) Confidence, (2) …
>
> Table 7 (originally, Table 2) reports the means and standard deviations for the respective courses and gender of Time 1 (pre-test) and Time 2 (post-test). Paired *t*-test was carried out to examine significant differences due to course and gender at Time 1 and Time 2.

Table 7: Results of paired *t*-tests if LSQ

| Course | Scale | Gender | Mean (SD) | | Mean difference | *t*-statistics |
|---|---|---|---|---|---|---|
| | | | Pre | Post | | |
| Normal Academic | Confidence | Female (68) | 3.43 (0.71) | 3.57 (0.51) | -0.14 | -1.56 |
| | | Male (53) | 3.28 (0.63) | 3.67 (0.60) | -0.39 | -3.94*** |
| Express | | Female (75) | 3.51 (0.44) | 3.89 (0.64) | -0.35 | -4.79*** |
| | | Male (79) | 3.61 (0.63) | 3.81 (0.63) | -0.20 | -2.66*** |

The results of the paired t-test showed that the perceived life skills measured by mean scores on the LSQ for Express students were higher than that of N(A) students. The course differences were still observed at Time 2, with the Express students displaying higher life skills development. At Time 1, male students collectively showed higher learning (sic) to learn life skills. Nonetheless, these effects were not observed at Time 2.

As shown in Table 7, there are in fact four independent analyses of single-group pre-and-post-test design experiments. The results of analyses as presented in the original table do not provide the needed information for the conclusion reached. For instance, when comparing between courses, the data of female and male students need be pooled. Likewise, when comparing by gender, data of the two courses need be pooled. The way it was done by the teacher-researchers is an over-simplification. To justify the conclusion, the data need be re-organized and analyzed. This is shown in Table 8A for comparing courses and Table 8B for comparing gender.

Table 8A: Mean comparisons by courses

| Measure | Express (N=128) | | Normal Academic (N=121) | | Difference | SMD |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | | |
| Pre-test | 3.56 | 0.55 | 3.36 | 0.68 | 0.20 | 0.29 |
| Post-test | 3.85 | 0.64 | 3.61 | 0.56 | 0.24 | 0.42 |

As can be seen in Table 8A, for pre-test, Express students scored higher on Confidence than did Normal (Academic) students with a small SMD of 0.29. For post-test, Express students also scored higher than did Normal (Academic) students with a greater SMD of 0.42. The conclusion is that the experience of doing interdisciplinary project-work was able to enhance the difference in Confidence between the two groups and in favour

of the Express students. This is consistent with the conclusion reached by the teacher-researcher, at least for Confidence. However, the initial difference (shown by SMD=0.29) cannot be ignored, though small.

As shown in Table 8B, there are no differences in both the pre-test and the post-test between male and female students, as shown by the SMDs of 0.02 in both comparisons. Thus, where Confidence is concerned, the conclusion is not the same as that reported.

Table 8B: Mean comparisons by gender

| Measure | Male (N=132) | | Female (N=143) | | Difference | SMD |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | | |
| Pre-test | 3.48 | 0.63 | 3.47 | 0.59 | 0.01 | 0.02 |
| Post-test | 3.75 | 0.62 | 3.74 | 0.59 | 0.01 | 0.02 |

Had the students been truly randomly sampled from their respective populations, the data could well be analyzed by a 2X2 repeated measure analysis of variance, since there are crossings of two genders with two courses and each student is repeatedly measured by the pre-test and post-test using the same test. Such an analysis allows the evaluation of the course main effect, gender main effect, and the course-gender interaction effect, plus testing occasion effect. This obviously will be a highly complex situation. However, since the four groups of students are not random samples, this analysis does not apply. For practically oriented school-based projects like this case, using SMD would suffice.

## Case study no. 5: Information overload

Information overload is as problematic as information insufficiency. Giving too little information makes thinking and conclusion vague. Giving too much information confuses people. When a simpler analysis is made more complex than it needs be, communication and thinking problems may arise. It is really an art to say what is necessary and stop there. Case Study No. 5 is a case in point.

> The subjects were 42 students from a secondary two normal academic class… Another class of secondary two academic students was assigned to be the control group… A pre-test was conducted using an instrument developed by the teachers. The format of the post-test was similarly designed.

Comparison between the two classes using ANOVA showed that the performance of the control group in the pre-test was similar to that of the treatment group, $F(1, 82) = 2.19$, $p>0.05$.

Table 9A: Statistics for pre-test (original Table 1)

| Group | Total number | Average | Variance |
|---|---|---|---|
| 2A1 (control) | 42 | 6.57 | 4.06 |
| 2A2 (Treatment) | 42 | 5.88 | 5.08 |

Table 9B: Statistics from single-factor ANOVA of pre-test results (original Table 2)

| Source of variance | SS | df | MS | $F$ | $p$ | $F_{crit}$ |
|---|---|---|---|---|---|---|
| Between group | 10.01 | 1 | 10.01 | 2.19 | 0.143 | 3.96 |
| Within group | 374.69 | 82 | 4.57 | | | |

(The same is done for the post-test in the report.)

This case has two good points. First, it compared *the project and the comparison groups* on two separate occasions, first for the pre-test and later for the post-test. As discussed above, it is the correct and logical way to make between-groups comparison in a two-group experiment. Second, since the pre-test and the post-test are two different tests, within-group comparison (like what is done in the previous case) will be erroneous.

The teacher-researchers use ANOVA (analysis of variance) instead of the conceptually simpler t-test, perhaps a preference. However, Table 9B is a correct standard way of presenting the result of an ANOVA but it contains many information which need not be shown for a school-based report, although it may be required in, say, a MEd thesis. The additional information is not meaningful to teachers and may make them wonder what they are for (and at the same time awed by mysterious numbers and labels). This is the problem of information overload: What do those labels across the top of Table 9A mean? Do readers need to know all these to understand the result? And, is there a simpler way to communicate the project outcome?

When the same information for the two tables is re-organized and analyzed, the result is shown in Table 10 below. Here, the SDs were calculated by taking the square-roots of the variances in Table 9A. In Table 10, the SMD of 0.39 indicates a small between-group difference and the conclusion is similar to that of using the more complex F-value obtained through the much complex ANOVA.

Table 10: Mean comparisons on pre-test

| Measure | Project (N=42) | | Comparison (N=42) | | Difference | SMD |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | | |
| Pre-test | 6.75 | 2.02 | 5.88 | 2.25 | 0.87 | 0.39 |

That the results of the two methods of analysis are similar is due to fact that, in a two-group experiment, $F = t^2$ or $t = \sqrt{(F)}$. However, in the re-analysis, SMD instead of *t*-value was obtained for reasons discussed earlier. The question is, since the *t*-test is more direct and simpler in concept and procedure, why go for the conceptually much more complex and procedurally much more cumbersome ANOVA and thus causing information overload with its ill-consequences?

## Discussion and conclusion

This paper illustrates five different kinds of statistical misapplications found in school-based project reports: (1) misplaced comparisons, (2) missing SD, (3) missing initial comparison, (4) over-simplification, and (5) information overload. Perhaps, with the exception of the last one (which may not be considered an error) the other fours are common errors.

The most common error is to report first on the pre-post-project mean difference of the project group, followed by the same of the pre-post-test mean difference of the comparison group, and then put the two results together and conclude that, since a difference is found for the project group but not the comparison group, the intervention benefits the project group and therefore the project works. This sounds logical but "*it is tempting to stop there, declare victory, and write it up for publication* (Abelson, 1995)." In short, it violates the logic of having a comparison group. This can be complicated by regression-to-the-mean threat if the two groups are non-equivalent initially.

Another common conceptual error is the use of the *t*-test when in fact it is not applicable and, worse, irrelevant. It is worth repeating that the *t*-test is an *inferential* statistics which can be used only when the data comes from groups randomly sampled from their respective populations. In the context of school-based projects, this condition is seldom, if occasionally, satisfied. The *t*-value and its corresponding *p*-value do not address the question of concern to teacher-researchers (and school administrators); these values,

however awesome they may look, are about the *probability* of observed difference and not about the *magnitude* of the observed difference. Moreover, the significance of a *t*-test result is also sensitive to group size. Again, why this conceptual error is so often made is unknown. Most probably, teacher-researchers are awed by the small decimal numbers, the word 'significance', and also probably misguided.

Statistics tell stories about projects and their effects, but the stories must be the correct ones that make statistical sense. The value of school-based projects does not depend on whether statistical techniques (especially the more complicated ones) are used, nor does it depend on the statistical *significance* - a word which is always mistaken to mean '*importance*' (Soh, 2011). As can be seen in many such reports, statistics seem to have been used for a cosmetic purpose because, after presenting one or more tables, the teacher-researchers go on presenting their views, instead of telling the story contained in their statistics.

Statistical misuses as exemplified by the five case studies here are not exclusive to school-based project or more generally educational research. It is also commonly found in other social research (Dodhia, 2007). And, Roehm (n.d.) gives ample examples from medical research. The question is not who make the most mistakes but how can mistakes be avoided and, if found, rectified. This calls for better training, more careful application, and more stringent editorial screening.

If we have to use statistics, use them correctly by referring to the right concepts, the right techniques, and the right language. A job worth doing deserves to be done well. Otherwise, we behave like a little boy who has just been given a hammer and finds everything needs knocking. Statistics may look like pure simple truth, but as Oscar Wilde once said, *"The pure and simple truth is rarely pure and never simple."*

## References

Abelson, Robert P. (1995). *Statistics as Principled Argument*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Coe, R. (2002). *It's the Effect Size, Stupid. What effect size is and why it is important?* Paper presented at the Annual Conference of the British Educational Research Association, University of Exeter, England, 12-14 September 2002. Retrieved December 5, 2011, from: http://www.leeds.ac.uk/educol/documents/00002182.htm

Cohen, J. (1988). *Statistical Power Analysis for the Behavioural Sciences, 2nd Edition.* Huillsdale, NJ: Earlbum.

*Cronbach's Alpha.* (n.d.). Retrieved December 5, 2011, from: http://www.experiment-resources.com/cronbachs-alpha.html

Dodhia, R. (2007). *Misuse of Statistics.* Retrieved August 8, 2012, from: http://statsconsult.com/Articles/Misuse%20of%20statistics.pdf

*Estimating Standard Deviation* (n.d). Retrieved December 5, 2011, from: http://www.statit.com/support/quality_practice_tips/estimating_std_dev.shtml

Fraley, R. C. (2003). *End of the Semester Thoughts on the Significance Testing Debate: A Review of the Problems with Significance Testing*. Retrieved December 5, 2011, from: http://www.uic.edu/classes/psych/psych548/fraley/NHSTsummary.htm

*Levene's Test.* (n.d.). Retrieved December 5, 2011, from: http://changingminds.org/explanations/research/analysis/levenes_test.htm

Roehm, E. (n.d.). *Improving Medical Statistics and the Interpretation of Medical Studies.* Retrieved August 8, 2012, from: http://www.improvingmedicalstatistics.com/index.html

Siegle, D. (2002). *Reliability*. Retrieved August 8, 2012, from: http://www.gifted.uconn.edu/siegle/research/Instrument%20Reliability%20and%20Validity/Reliability.htm

Soh, K. C. (2008). Effect size: What does it do for educational action researchers? *North Star, 1*(1), 63-70.

Soh, K. C. (2009). *Analyzing Data & Interpreting Outcomes: Statistical Toolbox for Teacher-Researchers*. Singapore: Cobee Publishing House.

Soh, K. C. (2011). Statistically speaking, correctly. *North Star*, *2*(2), 108-127.

Sterne, J. (n.d.). *The End of Statistical Significance?* Power-point presentation. Department of Social Medicine, University of Bristol UK.

Wikipedia. (2010). *Levene's Test*. Retrieved December 2, 2011, from: http://en.wikipedia.org/wiki/Levene's_test