Hong Kong Teachers' Centre Journal《香港教師中心學報》, *Vol. 10*

What's wrong when it goes wrong?

行動研究若不成功，原因何在？

Kay Cheng SOH

Singapore

Abstract

Action research, school-based curriculum innovations, and school improvement projects shared the common element of action taken by teachers to improve the students' learning and the schools' performance. Even with very careful planning and implementation, success cannot be guaranteed. When a project does not work out as expected, something has gone wrong. But, what is it? This paper discusses three main types of failures causing nil or negative results: theory failure, design failures, and treatment failures. Examples are given to illustrate these. While expecting success, teacher-researchers also need to be psychologically prepared for non-success and learn from honest failures.

Keywords

school-based curriculum innovation, school improvement, theory failure, design failure, treatment failure

摘要

行動研究、校本課程改革及學校改善計劃的共同特點是，教師採取措施以改進學生的學習和學校的績效，即使謹慎設計和認真推行，也不能保證一定成功。研究計劃如果未能得到預期效果，在哪些方面有差錯？本文探討無效果或相反效果的三個可能原因，即理論失當、設計失當和處理失當，並舉例說明。教師進行行動研究，當然預期成功，但也必須有心理準備，去面對不成功的計劃，並且從誠實的失誤中有所學得。

關鍵詞

校本課程改革，學校改善計劃，理論失當，設計失當，處理失當

The call for teachers to be engaged in classroom-based action research for professional development has always been sounded. It is most convincing when it comes from a practising teacher. Of late, Bijal Damani, an 11th and 12th grade teacher in Rajkot (India) who has received numerous honours including the 2009 ASCD Outstanding Young Educator Award, shares the thought (Damani, 2011) about teacher research, thus:

> *As teachers, we are always thinking about what we can do to reach out and engage students in our classes. And haven't we been experimenting by changing the way we give homework, grouping students differently, changing the classroom layout, or introducing some game or technology to see its effect on students learning? We have been doing these things for years – informally, maybe, but this is still a type of research.*

Like Damani, conscientious teachers are in a continuous process of trying to improve their teaching with the view to improve students' learning in terms of achievement, behaviour, and attitudes. Such efforts are generally referred to as classroom-based action research (AR). It has variedly been called school-based curriculum innovations (SCI) when the projects try out alternative instruction in a subject, and school improvement project (SI) when the aim is a school-wide improvement in academic performance and beyond. In this paper, the three terms (AR, SCI, and SI) are used interchangeably, since the common element is action taken by teachers for improvement in the students individually and the school as a whole. For such efforts, we naturally expect positive results since there are usually careful planning and implementation. However, as Murphy's Law suggests, "Anything that may go wrong will go wrong," nil and negative results may unexpectedly occur.

As an aside, in scientific and medical research, nil and negative results far out-number successful ones. It took 277 failed trials to successfully clone Dolly (*Failed Experiments*, n.d.), and the wonder drug penicillin (Bellis, n.d.) was discovered because experiments with certain fungus did not work. We are oblivious to failed experiments mainly for two reasons. First, successful projects get publicized a lot because of their implications for our sociological, psychological, and physical well-being. Secondly, "failed" projects do not get published in learned journals; for every published successful project, there are numerous well-designed "failed" ones. This creates the file drawer problem or publication bias (Sridharan & Freenland, 2009) which has only been recognized in the recent years as detrimental to proper understanding of the phenomenon researched on. This has led to the publication of learned journals, trying to rectify the situation, such as the *Journal of Failed Experiments, Journal of Articles in Support of Null Hypothesis, Journal of Negative Results in Biomedicine, Journal of Failed Crystallization Experiments, Journal of Pharmaceutical Negative Results, Journal of Failed Studies, Journal of Negative Results in Speech and Audio Sciences, etc*. These are not

meant to encourage complacency but to develop in the relevant research community intellectual integrity and to prepare the mind to learn from honest failures. For instance, the *Journal of Negative Results* (2010) explicates the purpose of its publication, thus,

> *The primary intention of Journal of Negative Results is to provide an online-medium for the publication of peer-reviewed, sound scientific work in ecology and evolutionary biology that may otherwise remain unknown. In recent years, the trend has been to publish only studies with "significant" results and to ignore studies that seem uneventful. This may lead to a biased, perhaps untrue, representation of what exists in nature. By counter-balancing such selective reporting, JNR aims to expand the capacity for formulating generalizations (p.1).*

In AR/SCI/SI projects, nil and negative results may come in two forms. First, the expected improvement fails to show up: the project students do not score higher than the comparison students do; even if there is a difference in favour of the project students, the effect size may be too small that it can be considered as trivial or null (Cohen, 1988; Soh, 2008). Secondly (and worse), the comparison students score higher than the project students do, contrary to expectation. It is, therefore, wise to evaluate the possibility of nil and negative results by looking up meta-analysis before embarking on a AR project (Hattie, 2009; Soh, 2010) This is shown by a surprising negative mean difference and a corresponding negative effect size. This reversal is termed Type III Error. By the way, there is another definition of Type III Error of getting the right answer for the wrong question (Wuensch, 2005).

Nil and negative results appear basically for two main reasons: (1) theory failure, and (2) implementation failures (which are sub-divided as design failures and treatment failures in ensuring discussion). These "failure" concepts help us take a critical and honest look at projects which do not work. Because any research entails a long process of many related actions, foreseen problems might have been prevented early enough, but those unforeseen can only be recognized when after the event *post hoc*. Hence, we are always wiser after the event and need be aware of Murphy's Law *"Anything that may go wrong will go wrong!"*

1. Theory Failure

We may begin our AR projects with some popular theories (e.g., Experiential Learning, Habits of Mind, Multiple Intelligences, Philosophy for Children, Problem-Based Learning, Inquiry-Based Learning, Socrates Questioning, Understanding by Design, Whole-Brain Learning, etc.) These theories may guide designing and planning and enable forecasting probable outcomes. Everything looks so proper (rationally) and rosy (emotionally) before the project starts. But, the end may be a different story. So, what's wrong when it goes wrong?

An education theory (often borrows, adapts, or applies psychological or sociological theories) integrates a set of variables and explicates their inter-connectedness in a

generalizable pattern. It enables understanding, guides instruction, and allows predictions. For example, constructivism (or more accurately, constructionism), attributed to the Swiss developmental psychologist Jean Piaget (Gray, n.d.), posits that students acquire knowledge and meaning from interaction between their new experiences and existing ideas, in contrast with the conventional view that people learn knowledge and meaning in a pre-digested form from their teachers. When we subscribe to this theory, we will arrange the learning environments for our students with the belief and hope that learning takes place *by itself*.

We may also begin with just some simpler ideas (e.g., individualized instruction, reduced class size, integrated curriculum, peer tutoring, etc.) An educational idea is a mini-theory functioning just like a grand theory but on a much circumscribed scale with less variables and simpler inter-connectedness, for example, the conventional wisdom that *practice makes perfect*. When we believe in this, we will emphasize in our teaching a lot of routine drills and practice with the expectation that more practice leads to better test performance.

There is no doubt of the usefulness of grand theories and pet ideas. They encapsulate variables in a compact form, maybe drawn as a diagram or stated in a few sentences as a mental model. They facilitate thinking that guides instruction. However, their very nature of abstraction can become a cause of problems for AR projects, for the simple reason that a theory applies to all relevant situations but may not fit tightly anyone of them.

A theory or idea may include the critical variables but surely not all relevant variables that may modify its prediction; this is where they go wrong. Thus, *constructivist teaching* (there is an obvious contradiction to put the two words together!) may not deliver what it promises because of uncontrolled variables not considered when planning or implementing an AR project. For instance, if the learning tasks are far above the students' current abilities or if the students lack the relevant background, that is, beyond their current zones of proximal development (Coffey, n.d.), or their concepts of learning is "to be told" and teaching is "to tell", then they will not be able to benefit from constructivist teaching as expected. For another instance, practice may not make perfect because too much drills tire the students out and learning becomes a chore so boring to them that they do not pay intention and hence do not learn.

If it is true that one size does not fit all, theories and ideas related to education and instruction definitely do not. The reason is simple: such theories and ideas, as alluded to earlier, just do not take into account all relevant variables which modify the relationships among the variables. Such moderating variables can cause a project to go wrong. There is in fact a very large corpus of aptitude-treatment interaction (ATI) studies showing the effect of a third variable impinging on the relationship between two variables being investigated. For example, McInerney, McInerney, & Marsh (1997) compared the effects of self-questioning as a meta-cognitive strategy on students in cooperative learning versus teaching groups when learning computer competencies. They

found the outcomes varying with the students' initial computer competency and anxiety level. For another example, a teacher may have initially a pet idea of *modeling* in mathematics problem-solving and only later finds it effective for a certain kind of students but not another kind in terms of left- or right-dominance of the brain. Here, the learning effectiveness depends on the interaction between the problem-solving strategy and the students' aptitudes (i.e., brain dominance).

It cannot be over-emphasized that the purpose of AR projects is *not* to verify the validity of some theories or ideas to prove them right or show them wrong; that is the purview of academic researchers (i.e. Master's degree and PhD candidates, and post-doctoral scholars) and not of the practice-oriented teacher-researchers. It may be useful to based AR projects on some relevant theories or ideas, using them as a short-cut to avoid blind trial-and-error but doing this is not to test the theory or idea, much less to lend the projects awe of *significance*.

Thus, when a project goes wrong, it could be that the theory or idea does not apply; the theory is wrong or irrelevant. To prevent this from happening, careful reading of the theory is necessary. Quoting big names and citing complex models do not lend a project its importance. Admiration and enthusiasm need be consciously controlled when considering the adaptation of grand theories. Likewise, pet ideas need be critically reviewed before they are used as the conceptual base of AR projects. When negative results occur unexpectedly, accept the results and learn about them. Review the theory or revise the idea and try again with due modifications. Take this as a process of professional growth and institutional learning.

2. Design Failures

These are one types of implementation failure. They have to do with how the AR projects were designed in terms of the number of students involved, the kind of students involved, and whether the groups were equivalent to begin with.

2.1 Small Group Sizes

AR projects usually involved intact classes and hence have limited group sizes due to practical constraints in the school context. Small group sizes mean low statistical power – the ability of a statistical test to detect a group difference when it exists. For instance, a project class of 36 students and a comparison class of 40 give a total of 76. In this case, with an expected effect size of 0.5 and a p-value of 0.05, the power is only 0.7. This is short of the conventional 0.8. Thus, if a nil result is obtained, it could well be due to the small group sizes and not that the alternative did not work; a Type II error. To rectify, or better still to prevent this from happening, increase the group sizes to a total of around 100. With the same expected effect size and the same *p*-value, this group size gives the statistical test a power of 0.8 and the design will be sufficiently powerful to detect a difference, if the alternative is really more effective. There are many power calculators on the Internet to assist teacher-researchers to decide on group sizes for their AR projects, for instance the one by Daniel Soper (2004-2011).

When a project returns with unexpected

outcome showing the comparison group outperforms the project group, the negative results might have been caused by teacher differences in, say, ability, teaching style, or teacher-student rapport (ruling out the possibility of a *John Henry Effect* to be discussed later). Even if the same teacher teaches both the project and the comparison groups, there is no guarantee that she will teach the two classes exactly as planned; this is just humanly impossible. Thus, for AR projects, we just have to live with this inevitable confounder and see the project outcome in its proper perspective by taking into consideration the teacher factor. This may sound pessimistic and somewhat unscientific but, as is true of many things in real life, we have to take the rough with the smooth.

2.2 Learner Aptitudes

The term *aptitude* here does not mean *special ability* (or *talent, propensity*) but just student characteristics that interact with learning to produce differential outcomes, in the sense as used in *aptitude-treatment interaction* (ATI). In other words, a particular student characteristic may bias him toward a certain kind of learning to produce a certain kind of outcome. For example, Mills (1993) reported a study that compared academically talented students and a group of same-age peers of mixed ability and found them differed on four Myers-Briggs Type Indicators dimensions, namely, introversion-extraversion, sensing-intuition, thinking-feeling, and judging-perceiving). It stands to reasons that such personality differences will influence the ways the two groups of students learn.

In AR projects, nil and negative results may occur when student aptitudes are not taken into account, because pooling the scores of students with different aptitudes masks the differential effects. Worse, when there is over-balance of one aptitude than the other in the project design, nil and negative results may obtain. Such reversals of the expected outcomes are examples of the *Simpson's Paradox* often found in educational and social research. An interesting and educative example from medicine is cited by Julious & Mullee (1994). They cited a study of the outcomes of two different kidney stones operation procedures. When only the procedures (analogous to two approaches of instruction) were compared, one was found to be more success than the other. But, when the data was analyzed separately for patients with different stone diameters (analogous to student aptitude), the direction of effectiveness was reversed. For AR projects, the differences in performance due to different student aptitudes can be uncovered by analyzing the test scores separately for different aptitude groups. And, doing this is obviously a good practice as a routine in data analysis.

A more subtle phenomenon which is more difficult to discern is the use of extreme groups. For instance, a group of *low* ability students learned through games and outdoor activities while another group of similar ability learned through teacher-centred lessons. Contrary to expectation, the comparison group scored higher on a posttest, although the project students found the lessons more interesting. The project students actively involved in games and outdoor activities might have been distracted from the learning tasks and this

mode of learning was foreign to them, whereas the comparison students learned in a more controlled environment using their habitual way of learning with much repeated examples and practice of the concepts to be learned. In this case, learner aptitude had an influence on the project outcome, unexpectedly.

2.3 Non-equivalent Groups

In the school context, randomization of students to form the project and the comparison groups is normally not practised as doing so will cause inconvenience or even discipline problems. If the school groups students by tracking or streaming, the intact classes involved in an AR projects are likely to be non-equivalent in relevant ability at the beginning. When the project and the comparison groups are non-equivalent and when the pretest is also used as the posttest, the data is usually analyzed by gain-score analysis. Assume that the project group was a weaker group (on pretest) and was given the alternative teaching to help them, while the stronger comparison group continued with the usual or regular teaching. The surprise may be a negative gain (Figure 1).

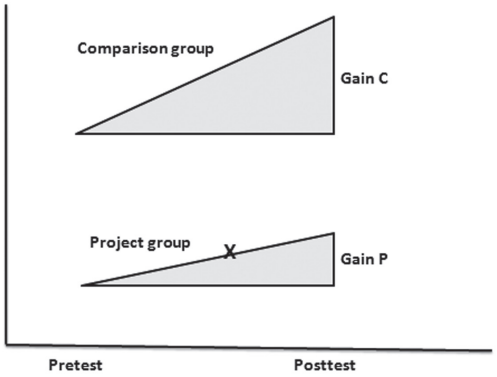

Figure 1. Negative Gain for non-equivalent groups

As shown in Figure 1, the negative gain (and therefore the negative effect size) simply indicates that the project group has gained (Gain P) *less than* has the comparison group (Gain C). This suggests that the alternative teaching is less effectiveness for the weaker project groups than the regular teaching is for the comparison group. To generalize, this cautions the teacher-researchers that not all methods (alternatives) are equally effective for all kind of students (i.e., the ATI problem). Simply put, one size does not fit all!

3. Treatment Failures

These are the second sets of possible reasons why a project fails and have more to do with implementation than project design, though related. They include low intensity of the alternative, insufficient time for the treatment to become effective, improper control of the comparison group, in adequate measuring tools, and teachers as a confounding factor.

3.1 Insufficient Dosage, Short Duration

In as sense, these are two sides of a coin. Insufficient dosage refers to the low intensity of treatment, just like that in medicine. An alternative teaching may not be sufficiently strong and distinct from the current teaching to produce the desired project effect. For example, two cloze-like worksheets were used to improve Secondary Four students' writing of qualitative analysis repots. Even if the idea is theoretically viable, two short exercises are most unlikely to change the level of report-writing skills in Chemistry which requires specific expression and register. Likewise, one field-trip is not

likely to change the students' ways of thinking and learning in science, geography, or history although it may influence their interests due perhaps to novelty effect. In short, when students are introduced to alternative teaching, it has to be intensive enough to take effect as the habits of learning the students have may just go against the alternative.

Short duration is related to the dosage problem, especially when AR projects tend to have short time frames. Certain things need only a short time to change, but others need more, again just like that in medicine. Changing language habits not only requires sufficient dosage (exposure and practice) but also sufficiently long time because the alternative or new language habits need to be reinforced and consolidated to counter-act against negative influences of past habits. Such student behaviours as punctuality, politeness, self-regulation, etc. are in the same vein. Too short a duration may lead to nil results (and Type II error).

Unfortunately, it is difficult to advise teacher-researchers on this problem, because there are so many different types of learning and different influencing factors which require different dosages. However, Bloom's taxonomies may be helpful here (Overbaugh, n.d.). Generally, the more specific learning is, the weaker dosage is required. For instance, learning simple factual knowledge need little time and less repetition, but learning to think critical requires lots of examples and practice. Thus, all other things being equal (but, they never are), higher-order thinking such as synthesis, analysis, and evaluation need stronger dosage and longer time to learn effectively.

3.2 Contamination

This refers to lack of proper control of the comparison condition. AR projects are usually implemented with both the project and the comparison group in the same school. Teachers teaching the comparison classes are supposed to teach them in the usual manner. But, there is no way these teachers can be kept in the dark and they are fully aware of the project intents. It is natural for them to unintentionally use some of the alternative activities or materials meant for the project groups. Worse still, control teachers may become worried since their classes are supposed to show up poorer at the end of the projects. This puts them in a defensive position and they may feel that their students are unfairly short-changed. This may motivate them to use the alternative teaching or even try harder to make the comparison students look good. This is the well-known *John Henry Effect* where by comparison railroad workers worked extra hard to out-perform the experimental group to maintain their egos and keep their jobs (Father Goof, 2008). The same may happen in the school. Even if the comparison students are "borrowed" from collaborating schools, it may be contaminated by their own projects which used different approaches but have similar goals.

3.3 Inadequate Measurement

Whether project effects are detected depends very much on the measurement of the criterion. If tests (broadly speaking to include attitude scales and observation schedules) are

not sensitive enough to detect group differences, project effects are under-estimated, leading to nil results. For example, a mathematics test is so easy that the project and the comparison groups both obtain high means showing little or no difference. The ceiling effect prevents the group difference to be detected. Or, the project aims to enhance high-order thinking but the test is heavily loaded with low-level items measuring recall without tapping on the thinking abilities. As AR projects tend to be short in duration covering only limited scope of content (and behaviour or feelings), nil results may occur because of inadequate measuring tools being used. Validity is "*the extent to which a test measures what it claims to measure. It is vital for a test to be valid in order for the results to be accurately applied and interpreted*" (Cherry, n.d.). It is obvious then that a test comprises mainly items testing recall of knowledge does not measure the student's ability in higher-order thinking; the test scores just do not represent what they are supposed to show. Thus, it is critical for teacher-researchers to ensure the validity of the test scores if the project outcomes are to be trustworthy.

3.4 Teacher Confounder

It is a truism that teachers are the most critical factor in AR projects because they are the very people who translate theories and ideas into actions that may influence student learning. It is also a truism that teachers have their personalities and abilities that determine their teaching styles when interacting with their students. Therefore, it is doubtful whether there are indeed teacher-free approaches and methods. This being the case, in the AR context, it is almost impossible to keep teacher factors under control. There simply are not sufficient teachers to be assigned to a large number of classes so that teacher effect can be evaluated as an independent variable. In fact, in the long history of educational research, there is hardly any large-scale project which had teachers randomized to rule out teachers as a confounding factor; *Project Star* (Tennessee's K-3 Class Size Study, 2009) is a rare exception.

Closing Note

We have never planned to get negative results but they do happen for various reasons. When they happen, the most rational thing to do is not to hide them, but to accept them and try to figure out why. This is not only a question of intellectual honesty but because negative results have lessons to learn. It is with this spirit that many new "failures" journals listed earlier were initiated to publish well-design research which produces nil or negative results.

This is not to encourage a culture of failure but to learn from honest failures. Such efforts are to cope with the file drawer problem or publication bias which arises from the common practice of publishing only studies with positive results and ignore those with nil or negative ones. Perhaps, we in education also need a *Journal of Projects with Negative Results* so that we can learn from both successes and failures in our effort to improve the students and the schools through action research.

References

Bellis, M. (n.d.). The history of penicillin. *Inventors*. Retrieved from http://inventors.about.com/od/pstartinventions/a/Penicillin.htm

Cherry, K. (n.d.). What is validity? *About.com Guide*, *Psychology*. Retrieved from http://psychology.about.com/od/researchmethods/f/validity.htm

Coffey, H. (n.d.). Zone of proximal development. *Learn NC*. Retrieved from http://www.learnnc.org/lp/pages/5075

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

Damani, B. (2011). Action research: A self-directed approach to professional development. *Educational Leadership*, *53*(7).

Failed Experiments (n.d.). Oracle think quest. Retrieved from http://library.thinkquest.org/03oct/01880/failed_experiments.htm

Father Goof (2008). John Henry vs. the bicycle. *For Father Only*. Retrieved from http://forfathersonly.blogspot.com/2008/07/in-one-of-many-statistics-courses-ive.html

Gray, A. (n.d.). Constructivist teaching and learning. SSTA Research Centre Report #97-07. *Journal of Negative Results* (2010), *7*(1), 1.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London, England: Routledge.

Julious, S. A., & Mullee, M. A. (1994). Confounding and Simpson's paradox. *British Medical Journal*, *309*, 1480.

McInerney, V., McInerney, D. M., & Marsh, H. W. (1997). Effects of metacognitive strategy training within a cooperative group learning context on computer achievement and anxiety: An aptitude–treatment interaction study. *Journal of Educational Psychology*, *9*(4), 686-695. doi: 10.1037/0022-0663.89.4.686

Mills, C. J. (1993). Personality, learning style and cognitive style profiles of mathematically talented students. *European Journal for High Ability*, *4*, 70-85.

Overbaugh, R. C. (n.d.). *Bloom's taxonomy*. Retrieved from http://www.odu.edu/educ/roverbau/Bloom/blooms_taxonomy.htm

Project Star: Tennessee's K-3 Class Size Study (2009). HEROS, Inc. Retrieved from http://www.heros-inc.org/star.htm

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

Soh, K. C. (2010). What are the chances of success for my project? And, what if it was already done? Using meta-analyzed effect sizes to inform project decision-making. *Educational Research Journal*, *25*(1), 13-25.

Soper, D. (2004-2011). *Statistics Calculators*, *Version 2.0*. Retrieved from http://www.danielsoper.com/statcalc/

Sridharan, L., & Greenland, P. (2009). Editorial policies and publication bias: The important of negative studies. *Archives of Internal Medicine*, *169*(11), 1022-1023.

Wuensch, K. L. (2005). *Controlling for type III errors*. Retrieved from http://core.ecu.edu/psyc/wuenschk/stathelp/Type_III.htm