Herding cats: Reflections on conducting randomised control trials in schools

Kevin Wheldall

When I was an undergraduate student of psychology in the late '60s, carrying out true experiments seemed relatively straightforward, according to our lectures and textbooks. One allocated the rats, for example, to different treatments and it was relatively simple to keep all other environmental variables constant. Even when we carried out experiments ourselves with human subjects, randomly allocating undergraduate students (the most studied group in all of psychology) to different treatment conditions was not a major issue and they too shared quite similar environments on the whole. What we could not directly control (for example, by ensuring that equal numbers of males and females were present in each group), we relied on careful randomisation of participants to conditions/treatments to avoid, or at least minimise, other potential influences, by ensuring that any possible differences among participants were just as likely to be present in one group as in the other.

When I graduated, I soon learned from bitter experience that research with human subjects was rarely as straightforward as it might initially have appeared. Nevertheless, I remained (and, indeed, still remain) as committed to truly experimental research as the gold standard. In the very first issue of the journal, *Educational Psychology: An international journal of experimental educational psychology* in 1981, Richard Riding and I (as the two newly appointed, founding joint-editors) proclaimed proudly in our editorial article our determination to promote a truly experimental approach to research in educational psychology. That this was easier said than done rapidly became increasingly apparent: not much truly experimental research was being conducted, it was largely correlational. While there has been some progress in this regard, it remains broadly as true today as it was then. In his 2009 book, *Visible Learning: A Synthesis of Over 800 Meta-analyses Relating to Achievement*, John Hattie commented:

"Some have argued that the only legitimate support for causal claims can come from randomized control trials (RCTs, i.e. trials in which subjects are allocated to an experimental or a control group according to a strictly random procedure). There are few such studies among the many outlined in this book ..."

And the reason, of course, was that so few true experiments were being carried out in educational contexts. This is because such research is very difficult to carry out in schools. By way of illustration, I offer two 'war stories', highlighting some of the problems and difficulties typically encountered when attempting to carry out randomised control trials (RCTs) in schools.

The disappearing control group

In the early 1990s, I jointly led a research team commissioned by the NSW Department of Education to evaluate the efficacy of Reading Recovery. After considerable debate, I was able to convince both the Department and our research team that a truly experimental evaluation was essential. To this end, a study was designed in which Year 1 students who were struggling to learn to read were randomly allocated to one of three conditions/groups from schools where Reading Recovery was operating. Equal numbers of young students

were allocated to one of three groups: an experimental group of young struggling readers who received Reading Recovery for 15 weeks; and a first and a second control group of struggling readers who received 'business as usual'; whatever remedial help was typically available in the school other than Reading Recovery. (A comparable 'comparison' group from different schools in which Reading Recovery was not operating was also recruited.)

We had thought ourselves smart, if not prescient, to include two control groups because we knew that once students had completed Reading Recovery, they would need to be replaced with fresh students for instruction by the Reading Recovery teachers. To this end, teachers were asked to recruit replacement students from the second control group but to leave the first control group strictly untouched. After 15 weeks, students in the experimental and the first control group were assessed and their performance on a battery of measures compared. All was well; the groups remained intact and fair comparisons could be made.

The idea was also to test for maintenance of gains and to retest students after a further 15 weeks of regular instruction following their exit from Reading Recovery. It was at this point that we realised that we had not been nearly as clever as we had thought. The teachers had recruited fresh students from the second control group, as requested, but once this source had been exhausted they then went on to recruit further fresh students from the first (real) control group. Consequently, our control group at 30 weeks was sorely depleted; not only this but it appeared that it was the weakest students from the control group who had been taken into Reading Recovery. This meant that the control group was not only smaller than desired but also far less representative than it had been initially. This made comparisons difficult and our findings at 30 weeks were thus subject to caveats. Fortunately, the comparison group comprising students from schools not receiving Reading Recovery was shown to be very similar to the experimental group at pre-test and hence comparisons between this group's performance at 30 weeks and that of the experimental (Reading Recovery) group could be made. But this evidence was far weaker than that from a true experimental comparison, as had originally been planned.

Were these teachers evil? Determined to wreck our research? No, not at all. They were simply doing their job which was to help as many struggling Year 1 students as possible.

The reluctant recruit

In a more recent study with which I am familiar, an independent research team was contracted to evaluate the efficacy of another remedial reading program with a strong emphasis on phonics. Schools were invited to take part but the decision to accept was taken by principals and not by the individual teachers who would be involved. Teachers who were to provide the novel program were carefully trained in exactly how to deliver the program. In order to ensure that this training had been effective and that the teachers were delivering the program as designed, all teachers were subsequently observed and their performance rated according to their compliance with the various key aspects of program delivery. This is known as treatment fidelity (sometimes called treatment integrity). Clearly, if measured treatment fidelity is low, then any evaluation of the program's efficacy will be invalid. If the program is not being taught properly, it is unlikely to be effective.

Treatment fidelity is typically expressed as a percentage of the number of critical components being reliably implemented by the teacher. In this study, one teacher was observed to have a treatment fidelity rating of 5-10%; she was not following the requirement of the program for over 90% of the time! To state the obvious, it is simply not possible to tell whether the program is effective or not if it is not being delivered properly most of the time. This teacher had been heard to observe that she simply could not bring herself to ask a child to "sound it out".

Was this teacher evil? Was she determined to wreck the research? No, not at all. She was simply doing her job which was to teach reading as well as she knew how. Unfortunately, her inclusion in an 'intention to treat' analysis has the potential to seriously compromise the findings of the study unless appropriate steps are taken to mitigate the effects.

Emeritus Professor Kevin Wheldall AM
Joint Editor