Research in Science Education (2005) 35: 3–22
DOI: 10.1007/s11165-004-3430-5
© Springer 2005

Issues Arising from the Long-Term Evaluation of Cognitive Acceleration Programs

Philip Adey
King's College London

Abstract

Since the early 1980s we have been engaged in the design, implementation, and evaluation of "cognitive acceleration" programs, which claim to enhance students' general intellectual processing skills. We have published extensive data showing long-term effects of these programs on students' cognitive development and academic achievement. In this paper the theory base of cognitive acceleration will be briefly outlined and some typical results re-presented. The main body of the paper will be devoted to a discussion of a number of issues which arise from the analysis of long-term effects. Although the examples will be drawn from the cognitive acceleration programs, the issues are general ones which might apply to any long-term study.

Key Words: intelligence, value added

"Cognitive Acceleration" refers to a teaching method which is grounded in theory drawn from Piaget and Vygotsky and which aims to promote the development of students' general ability to process information. That bald statement will raise immediate questions about the validity of the underlying theory, the existence of any *general* processing ability, and the possibility of influencing any such general ability through education. All of these questions have been addressed at length elsewhere (see, for example, Adey, 1997, 2003; Adey & Shayer, 1994) and will not be explored further here. For the purpose of the exploration of longitudinal data which is the subject of this paper, it is necessary to note only that the claim for the effects of cognitive acceleration will be sought in long-term effects on students' academic achievement. The argument is that if an intervention with, say, 12–14 year olds does indeed have a permanent positive effect on their general intelligence, and assuming that general intelligence is at least one important factor contributing to academic success, then one would expect these students to show gains in academic achievement which (a) can be observed years after the intervention, and (b) appears across a wide range of subject-matter contexts.

Cognitive acceleration methods are often described as resting on three main "pillars": *cognitive conflict* is the provision of problems which challenge a student's current way of thinking; *social construction* is the process of making knowledge together, as a group; and *metacognition* is the process of reflection on one's own, and the group's, thinking. These three, together with two subsidiary pillars (*concrete preparation* – introducing the topic; and *bridging* – showing how the same sort of

thinking can be used elsewhere) drive the design of sets of activities which are published as teachers' guides and materials for students. It is, however, recognised that no pedagogy can be delivered by print or software materials alone, and all cognitive acceleration programs have extensive professional development courses associated with them.

The first cognitive acceleration (CA) work was developed, from 1982, at Chelsea College London University – soon to be merged with King's College London. The initial work focused on junior secondary science students (12–14 year olds), and CASE (Cognitive Acceleration through Science Education) is still probably the best known CA program. Other members of the growing family include CAME (in mathematics for junior secondary), PCAME (mathematics for Years 5 and 6, ages 9–11 years), *Let's Think!* (science/general reasoning for Year 1–5–6 year olds), *Let's Think through Science!* (for Years 3 and 4, 7–9 years) – all developed at King's – and CATE (technology), and ARTS (junior secondary music, drama, and visual arts) – developed by colleagues closely associated with King's College. Each has published its own curriculum materials and each has an associated professional development course, but all rest on the same five pillars and all aim to promote general intellectual development. All of them are "intervention" programs in two senses: they aim to intervene in the "normal" (a word which begs all sorts of questions we will not go into here) cognitive development of students, and the activities are interventions in the normal curriculum. CASE provides one special activity to be used every two weeks instead of a regular science lesson over two years, *Let's Think!* provides one special half-hour activity to be fitted into Year 1 (5-year olds) once a week for one year, and *Let's Think through Science!* offers one activity every two weeks for one year for Year 3 children (7-year olds). For the purpose of this paper, and in recognition of the nature of the journal, it is the studies on these three cognitive acceleration in science programs that will be discussed. The published curriculum materials associated with each project are, respectively, *Thinking Science* (Adey, Shayer, & Yates, 2001), *Let's Think!* (Adey, Robertson, & Venville, 2001), and *Let's Think through Science!* (Adey, Nagy, Robertson, Serret, & Wadsworth, 2003).

Evaluation Methods Used and Results

In this section the quantitative evaluation designs and results will be summarised. As described above, the main evaluation effort on cognitive acceleration programs has been the search for evidence of long-term far-transfer effects which can be attributed to the CA intervention. In this paper we will discuss only quantitative data since this has been the predominant approach taken for CA evaluation. Readers interested in such long-term qualitative data on CA as exists are referred to Chapters 7, 8, and 9 of (Adey, Hewitt, Hewitt, & Landau, 2004).

In the development phase of each CA program a classical quasi-experimental design (Cohen & Manion, 1994) has been employed, with pretests, posttests, and delayed posttests of various types being applied to whole CA intervention classes

and to matched non-CA control classes. Tables 1, 2 and 3 summarise the time lines and tests used for the original CASE work and the more recent *Let's Think!* and *Let's Think through Science!* studies. Note that the CASE program is a two-year intervention, while *Let's Think!* and *Let's Think through Science!* are (currently) one year interventions. In each, long-term effects of the interventions have been or will be sought on both measures of cognitive development and on publicly accepted measures of academic achievement. At the secondary level the two "public" measures are the tests administered nationally at the end of Year 9, known as the "Key Stage 3 National Curriculum tests," and the "General Certificate of Secondary Education" (GCSE) examinations taken by all students at the end of Year 11, when they are 16 or 17 years old. In primary schools, there are national tests for 7 year olds ("Key Stage 1

Table 1
Experimental Design for Original CASE Research.

Group	Sept. '85 Y7 or 8; 11 or 12 yrs	. . . 2 yrs . . .	June '87 End of Y8 or 9; 13 or 14 yrs	June '88 End of Y9 or 10; 14 or 15 yrs	June '89 or '90* End of Y11; 16 yrs
CASE	2 Pretests of levels of cognitive development	*Thinking Science* CA intervention.	2 Posttests of cognitive development; science test	1 Delayed post-test of cognitive development; science test	GCSE** grades in science, mathematics, English
Control		Normal science			

*Some classes in each condition started in Year 7 (11+ years), some in Year 8 (12+ years), so reached GCSE in different years.
**GCSE: General Certificate of Secondary Education.

Table 2
Experimental Design for Let's Think!

Group	Sept. '99 Y1; 5 yr olds	. . . 1 yr . . .	June '00 End of Y1	June '01 End of Y2
CA	Pretests of levels of cognitive development	*Let's Think!* CA intervention.	Posttests of levels of cognitive development	Delayed test of inductive reasoning test & KS1 levels*
Control		Normal science		

*There are national tests of literacy and numeracy for all children at the end of Year 2.

Table 3
Experimental Design for Let's Think through Science!

Group	Sept '01 Y3; 7 yr olds	. . . 1 yr . . .	June '02 End of Y3	June '02 End of Y3
CA	Pretest of inductive	*LTTS!* CA intervention.	Posttests of inductive	Posttests of science
Control	reasoning	Normal science	reasoning	understanding

national curriculum tests") and for 11 year olds ("Key Stage 2 National Curriculum tests").

These are quasi-experiments since for a true experiment individual students would be randomly assigned to the intervention and control groups, rather than the different treatments being applied to intact existing class units. There does, however, exist a true experiment of the effect of CASE and CAME (the mathematics equivalent), conducted by Hautamäki, Kuusela, and Wikström (2002) in one town in Finland. They randomly assigned every student in the town who was in the school year in which they turned 12 years old, to one of three conditions: CASE intervention, CAME intervention, or no intervention. Then, depending on the group to which they had been assigned, children were bussed around the town to schools where one of the researchers would provide the appropriate intervention or a regular teacher would provide the control condition. I believe this to be a unique example of a true educational experiment. The results, which were quite surprising, will be summarised later in this section.

The original CASE work on which the first quasi-experiment was run started in 1985 and so more than enough time has elapsed for long-term delayed effects to be measured. It will be seen that in each case we looked for gains in the general intellectual factor (cognitive development, inductive reasoning) which was the subject of the pretest, but we also looked for effects on tests of achievement which have common currency in the educational world, such as the General Certificate of Secondary Education (GCSE) or National Curriculum "Key Stage" tests. Since these achievement tests relate to a factor distinct from that of the general intellectual tests, and are scored on a different scale, simple gain scores could not be used. For these tests, when inspection suggested that there was a possibility of a difference between experimental and control group, we calculated the "residualised gain scores" (Cronbach & Furby, 1970), by using the regression equation for those tests on pre-intervention cognitive measures for the control group to predict expected grades for the experimental groups if they had been no different. Comparing expected with actual gives the measure of gain – the residualised gain score of each pupil. This process depends, of course, on there being some correlation between the pretest of reasoning and the achievement tests taken some years later. If the pretests are

Table 4
Effect Sizes from the Original CASE Experiment.

		Immediate post '87		Delayed	GCSE, '89 or '90		
		Cog. dev.	Science	post '88 Science	Science	Maths	English
Year 7	Girls	–	–	0.60	0.67	0.72	0.69
start	Boys	–	–	–	–	–	–
Year 8	Girls	–	–	–	–	–	0.44
start	Boys	0.75	–	0.72	0.96	0.50	0.32

Only statistically significant ($p < 0.01$) effects sizes are shown. All effects are of CASE > non-CASE.

not moderate predictors of achievement, no meaningful regression equation can be obtained.

Results from the original CASE work have been discussed extensively (e.g., Adey & Shayer, 1993, 1994). The main effects are summarised in Table 4. Effect sizes are pre- to posttest gains, or residualised gains, expressed in units of standard deviation of the control gains. Effect sizes of 0.3 are normally considered respectable, 0.5 is large, and 1.0 exceptional.

When all CASE students are compared with all non-CASE students, the immediate and delayed effects on cognitive gain are statistically significant but when we look separately at boys and girls and at those who started aged about 11+ and those who started aged about 12+ years, the significant effects are found to be concentrated in the 12+ boys. Although they may not reach statistical significance, in every other group CASE scores were greater than non-CASE scores. Numbers are in the order of 30–50 students in each condition/age/gender subgroup and with quite large variances, hence large absolute effects are needed to reach statistical significance. Likewise, the overall effects on GCSE are significant when all CASE students are compared with all non-CASE students, but deeper analysis shows the effect to be concentrated in the girls who started CASE younger, and the boys who started CASE older. The two main features of these results which we have emphasised are:

1. that there is a very long-term effect; a two-year intervention with 12–14 year olds produces an effect on their achievement when they are 16+; and
2. there is a far-transfer effect; an intervention set in a science context and delivered by science teachers has an effect on mathematics and on English achievement.

It is on the basis of these features that claims for the effect of cognitive acceleration on *general* intellectual processing ability ("intelligence") are justified.

The true experiment in Finland (Hautamäki et al., 2002) had a population of 276 students with 92 in each condition (CASE, CAME, or control). There was a sophisticated testing program including immediate posttests, two occasions of delayed

tests, and tests of motivation as well as of cognitive gain. Here only the barest outline of the effects will be summarised. The immediate cognitive gain, residualised with respect to the control group, was 0.46 SD for the CASE group and 0.09 SD for CAME, and the first delayed effects were 0.79 SD and 0.59 SD, respectively. However, three years after the intervention, these differences between the control and experimental students had evaporated, but this was not because those who had had the intervention regressed to the norm. Rather it was because those who had been control students had accelerated unexpectedly to reach the same high levels of cognitive development as the experimental students – this is confirmed by comparing all of the students with Finnish national norms. A possible cause of this effect is that on average two-thirds of the students in any one class will have been in a CASE or CAME group, and this majority would have influenced the general level of thinking and intellectual interaction in the class in such a way as to "raise the game" of the control students.

The results from the first *Let's Think!* experiment have also been reported (Adey, Robertson, & Venville, 2002; Shayer & Adey, 2002). The quantitative effects are summarised in Table 5, all effect sizes are for CA students over non-CA students.

Here we have an immediate effect on cognitive development (the effect size for boys on conservation was 0.36, but this did not reach statistical significance) with evidence of transfer since no conservation activities are included in the intervention. However, there is apparently no long-term effect either on general (inductive) reasoning tested by Raven's matrices, nor on academic achievement in numeracy and literacy tested by the National Curriculum tests. Possible interpretations of this will be discussed in the next section.

For the Year 3 *Let's Think through Science!* (*LTTS*) intervention, we were able to demonstrate effects on neither Raven's matrices nor on science achievement tests taken immediately after the intervention. At the time of writing, no delayed-test data is yet available.

Before turning to the issues that arise for longitudinal studies in general from these results, an alternative approach to evaluating effects of an intervention program should be considered. With a new intervention for which one has no evidence

Table 5
Effect Sizes of the Let's Think! *Intervention on Cognitive Gain and Achievement.*

	Immediate posttests 2000		Delayed posttests 2001	
	Drawing	Conservation	Ravens Matrices	Nat. Curric. test
Boys	0.35	–	–	–
Girls	0.59	0.55	–	–

of any effect, it is appropriate to use a design which involves a control group. But once one has something for which there is evidence of a positive effect on students, consciously using control groups means denying a treatment to a group who would benefit from it, simply for the purpose of research. This is ethically problematic, and so an alternative approach must be taken, comparing treatment groups with others which have chosen not (or not yet) to have started to use the intervention although it is available (a selection process which brings other problems in train, which will need to be addressed). The alternative method we have used (Shayer, 1999) is an "added value system," looking at whole school units rather than individual children or classes (although class-level analysis was used on occasion). There is a reliable general relationship between the mean cognitive level of a school's Year 7 intake expressed as a percentile and the school's mean GCSE grade in various subjects achieved five years later, at the end of Year 11. It is possible to use this relationship to predict from a school's known mean intake abilities a likely mean grade at GCSE. If for all schools which implement CASE in Years 7 and 8 it is found that their actual mean GCSE grades exceed that predicted on the basis of national norms, we can infer that CASE has a general effect of raising academic achievement (Shayer, 1999). Figures 1 and 2 illustrate this effect, for science (the context in which CASE activities are set) and in English (indicating far transfer.) This has been the general approach to longitudinal evaluation of CASE since the 1989 and 1990 GCSE results first provided evidence for the long-term effect of CASE on academic achievement.

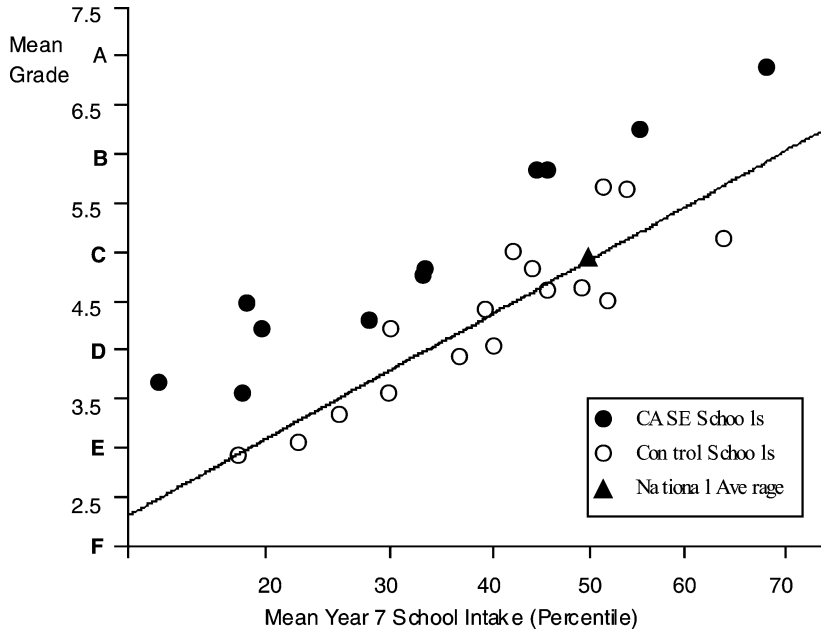

Figure 1: Value added effects of CASE on GCSE science.

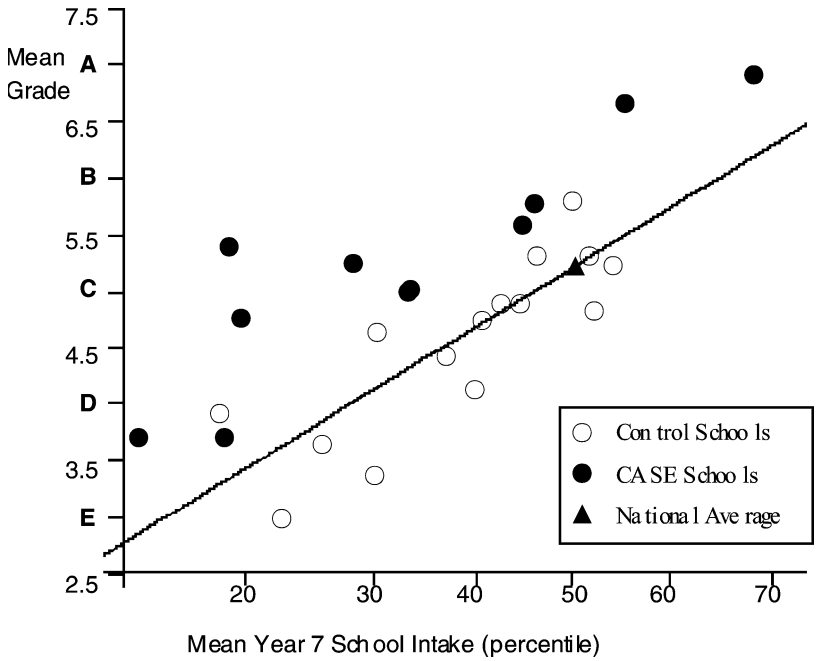

Figure 2: Value-added effect of CASE on GCSE English.

Issues with the CASE Longitudinal Studies

In this section a number of issues which have arisen in considering the validity of the results reported from the CASE projects will be raised. However, all of these issues have a general applicability to many longitudinal studies; CASE simply provides the exemplifications for this paper. There are nine such issues, which will be grouped under the headings: Interpretation, Design Issues, Measurement, and Analysis.

Interpretation Issues

1. Attributing the effect

The task for an evaluator is to attribute an observed effect unequivocally to a specific cause, or intervention. The problem is that schools are incredibly complex systems with millions of variables – the vast majority of which are uncontrollable. It is therefore always open to the sceptic to ask "How can you be certain that the observed gains in GCSE grades (or Key Stage National Tests) were caused specifically by the Cognitive Acceleration program?" The first response is to ensure that the sample size is large enough so that the extraneous uncontrollable variables are

effectively randomised. Showing that one class using CASE gets better examination results than another class not using CASE tells us nothing, since the difference could be due to different students or to generally different teaching which owed nothing to the specifics of CASE. But if 100 CASE classes show consistent gains in achievement compared with 100 control classes (either selected as controls or the virtual controls of national norms), then we are justified in claiming that CASE is by far the most likely cause of the effect. That is, unless our sceptic can point to a systematic difference (apart from the CA treatment) between the experimental and control classes, then we are justified in assuming that all of those unseen and uncontrollable variations are randomly distributed amongst the control and experimental classes.

To some extent this is an issue which arises with any quantitative evaluation of an intervention, but it is exacerbated in longitudinal studies which offer more opportunities for the effects of extraneous factors. That is, if one is following up a cohort of students for some years after an intervention, the chance increases that one or more of the classes or schools containing experimental students will become the subject of another intervention with broadly the same intended outcomes (gains in cognitive ability or academic gains).

A more insidious version of the idea of alternative attributions is the "Hawthorne Effect," much beloved by questioners at almost every lay audience to which I have presented our results. Generally, the questioner has no knowledge of the original Hawthorne paper (Mayo, 1945) which reported short-term gains in output from office workers who were told that a particular colour paint on the office walls would make them work more effectively. The "insidious" aspect of this is that the alternative explanation, the thing that is really supposed to bring about the effect, is inevitably brought in train with the intervention. The argument is that simply telling people (in our case, 12-year olds) that they will become more intelligent is enough to make it happen. Leaving aside the fact that telling people that they will work harder (a largely volitional variable) is somewhat different from telling them that they will become cleverer (a variable largely beyond the immediate conscious control of the individual), the whole point of the Mayo work was that it is short term. The effect lasted only a few weeks, and so did not represent a valid substantial effect. With CASE we have, not just a long-term effect, but one of a handful of studies reported in education with follow-up evaluations years after the intervention.

While the reliable attribution of an effect to one cause is difficult, it is not impossible within given probability limits. In contrast, attributing the "cause" for a non-effect is, in principle, impossible without further experimentation. Table 5 shows that no long-term effect was observed from the *Let's Think!* (*LT*) intervention, and with the *Let's Think through Science!* (*LTTS*) program we did not even get immediate effects. One may offer hypotheses to explain away such non-effects: for *LT* such a hypothesis is offered under "controls" below, and for *LTTS* the current best guess is that 15 hours over one year is simply too slight an intervention to have a significant effect on general (inductive) thinking, especially since the funding situation meant that the trial and implementation period had to be combined. Plausible though this may sound, it cannot be offered as an established conclusion until it can be shown

that a longer intervention (say two years) introduced to teachers with a dedicated PD program unconstrained by the need to collect feedback for development purposes does, in fact, produce a significant and lasting effect on cognition.

2. Characterising the intervention

Another version of the Hawthorne question is "wouldn't any special attention paid to pupils over one or two years have the same effect," to which the answer is "please describe your special attention." CA has a substantial theory base and a set of procedures and activities based on this theory, trialled, and made explicit for teachers' use. It certainly does give "special attention" to students, but it is special attention of a very specific kind which can be observed and evaluated. It is open to anyone else to devise alternative intervention strategies based on different theories and to trial them against CASE or against "normal" education, and they might well prove to be more effective and more efficient. The theoretical possibility of alternative intervention strategies does not invalidate the evidence that CASE is effective.

Nevertheless, there is a valid question underlying the rather woolly one presented above, and that is, what exactly is it in CASE which is doing the work? CA interventions consist of a complex of things, including:

1. Promotion of a theoretical model of what should cause cognitive stimulation, based on Piagetian ideas of cognitive conflict and schema theory, Vygotskyan ideas of social construction, and catholic ideas of metacognition (reflective abstraction from Piaget, language as a mediator from Vygotsky).
2. A set of materials (printed teachers' guides and student notesheets, and apparatus) which (a) spell out this theory and relate it to the pedagogic practice which supposedly follows, and (b) provide a set of exemplar activities designed to provide contexts for cognitive conflict, social construction, and metacognition.
3. An extensive professional development program including a number of inservice days and coaching visits over one or two years.

So what do the experiments test: the theory base, the materials, or the professional development? This question is not just difficult to answer experimentally (because of the practical difficulty of separating the components), but may be impossible to answer in principle. One needs to go back to Stenhouse's (1975) ideas of what constitutes a curriculum: essentially everything that happens within a school, including those actions which are purposeful towards particular learning and development objectives, and those which are incidentally provided by the environment. An intervention into this curriculum necessarily impinges on many of the interactions, and especially on the quality of those interactions in classrooms between teachers and students, and students and students. The change in these interactions will be a function of the materials, theory, and PD which characterise the intervention, but these elements will act differently and with different weights in different environments. For some teachers, the materials may be enough to bring about change; for others, the PD program will shift perceptions of teaching and learning which will lead to changes in all of their lessons, with less emphasis on the special lessons.

The importance of this issue depends on one's perspective. For the teacher whose main goal is improved student achievement, it really does not matter much how that is attained. If the package works, don't take it apart. But a cognitive psychologist interested in building models of the mind which have predictive ability might be unhappy with such a global view, and want to gather far more detail of classroom interaction – perhaps by some form of discourse analysis linked to cognitive gain tests. A government officer interested in the biggest effect for a given sum of money might take an Occam's Razor approach – can we get away without the coaching visits? In every case, the situation is again complicated in longitudinal studies, especially where there seems to be some sort of delayed effect. Table 4 shows that there was no immediate effect of CASE on science achievement at the end of the intervention, but one year later, with no further intervention, science scores of those who had experienced CASE rose significantly relative to controls. This implies that any of these further analytical tests sought by psychologists or by funders will need as long a time to work out as did the original experiment.

Design Issues

3. Sampling: Representativeness, decay

As well as sample size, discussed above, it is of course important to choose a sample for the experimental design which is representative of the population for which claims are being made. In the original experiment, CASE classes were chosen (1) from secondary comprehensive schools recommended by science advisers as either being in need of help, or as likely to follow the program as offered; and (2) within the chosen schools by the head of science as representing middle of the range ability. Control classes were also identified in the same schools as being as similar as possible in mean ability to the selected CASE classes. This was a crude attempt at choosing a "middle-of-the-road" sample broadly representative of the great majority of students in secondary schools in the UK. In this case, wider generalisation of the effects to both gifted children and to those with learning difficulties has, in practice, been a matter of collecting anecdotal evidence. Even after a decade, this does not add up to anything which would pass the review process of an academic journal, but the matter is not pressing and until an enthusiastic PhD student chooses it as a question to address, it will probably remain uninvestigated.

For the *Let's Think!* work in Year 1, the experimental CA classes were pre-ordained by the source of funding, a government "single Regeneration Budget" targeted at a specific disadvantaged area of an inner city borough. The project had to include all of the Year 1 classes in schools in that area, a total of 14 classes in 10 schools. For controls we identified 8 classes in 5 schools in an immediately adjoining area which was not quite as economically disadvantaged, but which had a similar ethnic, free school meal, and language profile. Pretests showed that the CA and control samples were not significantly different in starting levels of cognitive development. However,

in extrapolating the results obtained from this sample nationally, there is a legitimate fear that a cognitive stimulation program developed for, and showing effects on, a rather especially socially-disadvantaged population may not have similar effects on more favoured populations. It may be supposed, for example, that the gains made by the experimental group reflected compensation for an intellectual deficit in these children's lives, and that in areas of greater affluence children would already have received their "quota" of intellectual stimulation such that not much more gain could be made. To report that this hypothesis has been ridiculed by every teacher in a middle class area to whom I have floated it is reassuring, but it does not constitute research evidence. We are currently replicating the original experiment in a number of regions of Britain, including some with markedly different social profiles from the original one. Only when we have data from these wider trials can we be confident that the effect is general. Again, replicating longitudinal studies presents particular problems in terms of time frames and costs.

There is another issue related to sampling which is peculiar to longitudinal studies, and that is attrition of the sample over the years. Data from a total of 10 tests over 4 or 5 years were accumulated for the sample in the original 1984–87 CASE experiment. In some cases, pretests were administered to the students at middle school level and those students had to be tracked to whichever high school they had moved in order to administer the delayed tests or collect their GCSE results. There is an inevitable general process of attrition of the sample as students' parents move and take their children to different schools, added to which is loss of data due to particular tests being missed either by individuals through illness, etc., or occasionally by whole classes whose teachers do not "find time" to administer the tests. Obtaining pupil GCSE data from schools can be an extremely time-consuming process. The outcome is that the matrix of (students x test scores) contains many "holes." The strictest approach to this problem is to enter into processing only those students who have a complete set of (in this case 10) test scores. However where, as here, two parallel tests of the same construct on the same occasion have been used to increase reliability, if one is missing the other may be used alone. For reporting some of the intermediate effects (for example, the first delayed posttest), it is not strictly necessary to have the long delayed (GCSE) data in order to report effects thus far. What is essential is that one has pretest data. Students who for one reason or another miss the pretests are excluded from any further analysis. In cases where the GCSE data but not delayed posttest data was available, we processed those with and those without delayed-test data separately to ensure that there was no systematic difference in the mean effects that emerged. Having established this, one could report the results including those with missing delayed tests with confidence.

At this remove, nearly 20 years on, I cannot recover the original data to determine percentage losses suffered in this experiment, but memory suggests that those for whom we finished with a complete set of data (across five years from pretest to GCSE) comprised at least 80% of those for whom pretest data was obtained. Bear in mind that this was essentially a "middle England" population. The population for the *Let's Think!* experiment was very different. In line with the nature of the funding

for the project, the children came from a wide variety of backgrounds: local working class, travellers, third, second, and first generation immigrants, refugees and asylum seekers. Of 302 children for whom pretest drawing task data was obtained, only 237 (78.5%) were also present for the delayed posttest taken two years later. Here it was important to establish that, in terms of pretest mean scores, those for whom delayed posttest data were obtained were not significantly different from the whole original sample. Otherwise, it might be claimed that any effect observed only applied to the less mobile portion of the sample.

4. Nature and behaviour of controls

An important point about control groups is that the experimenter actually has little or no control over them. There is no obvious benefit to a headteacher to permit her or his students to be pretested, posttested, and to have their public examination scores released to a researcher whose purposes may not be totally transparent. The researcher has to use a combination of cajolery – perhaps with the help of the Local Education Authority – promises of confidentiality, and mild bribes such as offering the test data on the control children as a valuable diagnostic tool, or agreeing to provide the school with the intervention in the following year. But even then controls can misbehave and damage one's carefully designed experiment, since they have little vested interest in maintaining "normal" teaching (whatever that may be) when there are exciting innovations to be had. We have had two examples of this in the CA work.

In the original experiment, one CA class was selected in a London school which happened to have a rather charismatic head of science who also had strong theoretical interests. CASE had not been running in the experimental class for the full two years before he decided that CASE was too good a thing to restrict to one class only, and introduced CASE and CASE-like activities across all science classes in all years. Initial disappointment at finding no experimental-control differences in this school led to curiosity when it was seen that the control groups had made much greater gains than should have been expected. When questioned, the head of science cheerfully admitted his "misdemeanour." (He has since become one of the main CASE trainers in the UK!)

A less benign example is provided by one of the control schools in the *Let's Think!* experiment. This was a school dedicated to examination success and to attracting the children of more ambitious parents. Everything in the curriculum and the management of the school was devoted to maximising the levels obtained by children in the national curriculum tests. When we looked at the mean residualised gain scores of the Key Stage 1 national tests (the achievement part of the delayed-test battery – see Table 2) for each class in the experiment, it was noticeable that the two control classes from this school had gain scores far higher than the other controls. The effect, of course, is to raise the overall control mean considerably, and since all residualised gains are related to the control mean, to depress the residualised gain scores of all of the experimental classes. To an academic referee (including this author were it to be

presented to him blind by a journal editor), this may look like special pleading, or a post-hoc explanation for the absence of a desired result. Indeed, it may be argued that if the effect of improvement in national curriculum test scores as measures of achievement can be obtained by behaviourist or mechanical methods which are cheap, why engage in the deeper level and far more expensive process of professional development required to implement a model rooted in cognitive psychology? That question will be left open as a bit of mental activity for the reader to engage in.

The point of this section is that, especially with longitudinal studies, controls can mess you up. The simple answer is to increase one's sample size in the hopes of diluting or randomising the effect of maverick control groups, but such a solution is expensive. A more complex answer is to add a detailed qualitative analysis of the transactions in experimental and control classes in an attempt to relate features of the theoretical model (in this case, cognitive conflict, social construction, and metacognition) directly to individual class (or even student) gains, but that is even more expensive.

5. Delayed effects

It is not putting it too strongly to say that one of the great joys of longitudinal work is to find, years after an initial intervention, that there is an important positive effect of the intervention on students' lives which was not immediately apparent at posttest. American programs for disadvantaged pre-school children, such as Head Start and High Scope (Sylva & Ilsley, 1992), have been able to claim lower rates of school drop-out, criminal convictions, and teenage pregnancies in young people who participated in the programs 10 or more years earlier, as well as enhanced achievement (Lazar & Darlington, 1982; Zigler & Burman, 1983, quoted in Sternberg, 2003), although there had been no obvious immediate effects of the interventions. Interventions may have effects which are too small to detect immediately, but which have a cumulative effect, such that as the years go by differences between treatment and control groups become magnified. Better learning leads to better learning, through either or both of cognitive and affective mechanisms ("nothing succeeds like success"). In the original CASE experiment, there was no immediate effect after two years of intervention on student's ability in science, but one year after the end of the program, a clear difference emerged between students previously in experimental and control groups, in spite of the fact that many of them had been re-distributed in new schools and were no longer in intact groups with the same teachers.

Too much research, unfortunately, is dominated by the duration of a Ph.D. course or by the length of time a funder is prepared to commit to a project. In three years one can only design an intervention, trial and modify it, and then give it one run-through, at the most. The most important effects, those which appear some years later, must often be missed.

Measurement Issues

6. Baseline tests

All longitudinal quantitative studies rely on baseline data to which all subsequent test data can be referred. It goes without saying that the baseline measures must be reliable and valid with respect to the underlying construct chosen by the researcher as of interest and importance As noted previously, if they are to act as useful baselines to which subsequent test results can be referred, pretests must be moderately good predictors of success in the criterion measures. In the business of accelerating general cognitive development, and in looking for evidence of effects both in measures of general intelligence and in academic achievement in school subjects, the task of identifying appropriate pretests is, in principle, relatively easy since there are many measures of intelligence and of inductive reasoning (which is often claimed as the core of general intelligence (Carroll, 1993; Vernon, 1971)). We have generally used pretests based on Piagetian notions of intelligence, because the reasoning which they tap into is transparently related to the several operations described by the Genevan school as central schemata of intelligent behaviour. We have also used more general intelligence tests, such as Raven's Coloured Progressive Matrices (Raven, 1998). Typically, the correlation of our Piagetian tests (Shayer, Wylam, Küchemann, & Adey, 1978) taken at the beginning of Year 7, with GCSE examinations taken at the end of Year 11, are in the order of .35 to .45, and with Key Stage 1 national curriculum tests taken two years later are .41 with language and .51 with numeracy, for the control group.

For the *Let's Think!* work there was an alternative baseline test available: an assessment of every child in the year before they entered Year 1 conducted by their class teachers according to a specific criterion-referenced system (Birmingham City Council, 1997). In fact, these also had comparable correlations with the Key Stage 1 national tests (for example, .49 for mathematics on mathematics, and .53 for language on language) but at the time there was concern about their reliability since the scheme was new in the borough. It had been suggested that teachers had been inadequately monitored in the assessment procedure. It was also made clear to us by at least one teacher that it was accepted practice to tend towards undervaluing the baseline scores in order to promote the "value-added" measures from baseline to Key Stage 1, a measure on which schools are judged in a high-stakes inter-school competitive atmosphere created originally by the Thatcher government and enthusiastically maintained by the Blair government.

It is not only in England that teachers learn to manipulate high-stakes testing. CASE was introduced to the nine high schools of one school district in Arizona in 1992, as a one-year program completely replacing the regular grade 9 science curriculum (Forsman, Adey, & Barber, 1993). Pretests and posttests were administered for two years in a row, firstly before the implementation of CASE, and then in the first implementation year. Initially the results appeared gratifying, since gains in the CASE year were significantly greater than gains in the previous year. Closer

inspection revealed, however, that there was no significant difference between the post scores year on year. What had happened in the CASE year was that the pretest scores had dropped sharply. There had been no demographic change which could explain such a radical effect across all nine schools, and the most likely explanation seems to be that teachers accustomed to being judged (for purposes of promotion and pay) by gain scores had learned to depress baseline measures.

All of this highlights the fact that much of the validity of claims made for the impact of an intervention from a comparison of intervention and control groups hangs on the validity of the data from baseline tests as administered. It follows that careful attention to the selection, administration, and scoring of pretests is essential if years of work are not to be put in jeopardy.

7. Outcome measures: theory driven versus common currency

A general principle of the evaluation of the effect of an intervention is that the outcome measures (post- and delayed posttests) should be related to the intention of the intervention, that is, to the hypothesised effect. Thus an intervention designed to improve the attitude of students towards science as a school subject would need to measure, as the main outcome, changes in attitude towards science as a school subject. For the effect to be convincing, the outcome measures would have to demonstrate construct validity – that they really do provide a measure of the construct of "science as a school subject." In many (if not most) cases, however, the immediate outcome measure is seen as only one step towards a broader, secondary effect. While the primary outcome is typically a psychological construct for which valid and reliable measures can be developed, secondary outcomes, especially in educational research, are more likely to be rather broad measures of "general good." Attitude to school science is a construct which lends itself to definition and to measurement, but the broader impact of improving this attitude – the secondary aim of the intervention – might be to increase the numbers of students choosing science in higher education and as a career, and/or to increase the general public's understanding of, and attitudes towards, science in society. While primary outcomes are measurable, they may be seen by the lay public (and by some potential funders of research) as rather academic and as promoting educational research for its own ends. The secondary outcomes are those which are of more obvious value to the society as a whole, but they are also the ones which are more difficult to define clearly or to measure reliably. Secondary outcomes, moreover, are inevitably longer-term and likely to be the subject of longitudinal studies.

In the case of cognitive acceleration, the primary outcome is gains in intelligence, conceived in terms of the cognitive operations described by Piaget and his colleagues as lying on a progressive scale of increasing complexity and sophistication. Were it only measures of cognitive gain that were reported, then the communication of results would be limited to psychology and education journals, and would be considered as somewhat arcane by the wider educational public. On the other hand,

reporting the effects also on national tests, which are common currency in the educational (and wider) world, raises a number of issues. On the one hand, it does place the research work clearly in the public domain so that teachers can see directly the value of adopting the methods of the intervention into their practice. On the other, it involves an element of risk for the researcher, since the connection between primary and secondary outcomes may be uncertain. If, as is most common, one reports the primary outcome and then concludes "this effect should lead to greater ... (secondary outcome) effects," one is playing safe but limiting the audience to academics. It may be plausible that improved attitudes lead to greater uptake of science, but the users of research (typically teachers) would not be unreasonable in replying "OK, but I want to see evidence of this effect before I adopt this intervention which will be effortful and will disrupt my normal practice." On the other hand, if one looks for the secondary effect and does not find it – at least initially, then one cannot even make a case for the possibility of a secondary effect – it has already been shown not to occur.

Within our cognitive acceleration work we have chosen to take that risk, and to take any non-effect on secondary outcomes as a challenge to improve the power or implementation quality of the intervention. If even this eventually fails to produce a secondary effect, we may draw conclusions about, for example, critical ages at which cognitive acceleration techniques can have a permanent transfer effect.

8. Gain scores: ceiling and floor effects

Prieler (2001) and others have argued forcefully against the validity of comparing score gains of different groups of subjects. Their argument is based on the fact that the relationship between an underlying construct and the score on a test is not linear, so that at one point in the scale a small gain in the value of the construct may cause a large score gain, and at others (typically near the "floor" and the "ceiling") large underlying gains are required to raise the test score only moderately. Thus if the two groups being compared are not comparable in initial mean levels and distribution, the raw gain of one cannot be meaningfully compared with the raw gain of the other. We believe that we have met the conditions that our experimental and control groups have very similar initial profiles, neither near the floor or ceiling of the tests being used. However, this is an issue which can become important in longitudinal studies with repeated use of the same measures, where a ceiling might be approached by a significant proportion of the students.

Analysis Issues

9. Levels of analysis

Although not strictly a particular problem for longitudinal studies, any experiment which compares students in classes, in schools, needs to pay attention to the unit

of analysis for which claims of significance are made. In quasi experiments such as have been described here, where whole classes are assigned to experimental or control conditions, the *N* in calculations of significance is the number of classes. To use the number of students is to inflate measures of significance and to ignore the fact that the "treatment" is a treatment directly of the teachers (through materials and professional development), and only indirectly of the students (through the effect on their teachers). In a genuine experiment such as that conducted by our Finnish colleagues, it is legitimate to treat the individual student as the unit of analysis. Having spelled out the gospel according to statisticians, it must now be admitted that we have in fact calculated residualised gain scores for individuals (as one must) and then, as a first pass on the data, reported differences and significances of differences between whole experimental and whole control groups, taking the number of students and the standard deviation of their scores as the parameters for estimating significances. This is on the grounds that in the original experiment, with 10 experimental classes, by the time they were divided by starting year we have *N* (class) values of 6 and 4, far too small from which to expect significances without massive absolute differences in gain scores.

We did also look at the mean residualised gain scores of each class individually in the CA and control schools. This is sensitive data as it points directly at individual teachers, and must therefore be treated with considerable caution. In particular, the numbers are often small, allowing for much random variation around any true mean. Even if the impact on a particular class can be attributed immediately to the way that the teacher delivers the intervention, her delivery will be the result of a complex of variables, including teacher variables (comprehension of the intervention, beliefs about teaching and learning, classroom management competence, etc.) and school variables (resources, senior management support, timetabling issues, etc.).

Conclusion

Longitudinal work is rarely fully planned for in advance, since research funding is not commonly available for more than three years at a time, and so it depends on the stability of employment of some key players in the original research and their ability to find more resources for continuing data collection and analysis. It depends also on the willingness of headteachers to permit repeated testing or the release of individual data from public examinations. If these conditions can be met, however, longitudinal work offers by far the most powerful method of really testing the effect of an intervention. In this paper I have raised a number of issues which can threaten the validity of conclusions drawn and which are peculiar to longitudinal studies. While to an extent they may reinforce the view that longitudinal research is practically too difficult to conduct, I hope that the methodological successes that have been achieved within evaluations of cognitive acceleration programs will encourage others to start their own long roads, or to revisit shorter lanes they thought that had finished with, to see whether follow up studies might still yield useful, and possibly

surprising, results. Maybe it is not too much to hope that a continual drip of reports on the benefits of longitudinal research will slowly re-shape the attitude of research councils, charitable trusts, government departments, and others who fund research but who want to see immediate effects. Important changes in people do not happen overnight.

Correspondence: Philip Adey, Department of Education and Professional Studies, King's College London, Franklin Wilkins Building, Stamford Street, London, SE1 9NN, UK
E-mail: philip.adey@kcl.ac.uk

References

Adey, P. (1997). It all depends on the context, doesn't it? Searching for general, educable, dragons. *Studies in Science Education, 29*, 45–92.

Adey, P. (2003). Changing minds. *Educational and child psychology, 20*(2), 19–30.

Adey, P., Hewitt, G., Hewitt, J., & Landau, N. (2004). *The professional development of teachers: Practice and theory*. Dordrecht, The Netherlands: Kluwer.

Adey, P., Nagy, F., Robertson, A., Serret, N., & Wadsworth, P. (2003). *Let's think through science!* London: NFER-Nelson.

Adey, P., Robertson, A., & Venville, G. (2001). *Let's think!* Slough, UK: NFER-Nelson.

Adey, P., Robertson, A., & Venville, G. (2002). Effects of a cognitive stimulation programme on Year 1 pupils. *British Journal of Educational Psychology, 72*, 1–25.

Adey, P., & Shayer, M. (1993). An exploration of long-term far-transfer effects following an extended intervention programme in the high school science curriculum. *Cognition and Instruction, 11*(1), 1–29.

Adey, P., & Shayer, M. (1994). *Really raising standards: Cognitive intervention and academic achievement*. London: Routledge.

Adey, P., Shayer, M., & Yates, C. (2001). *Thinking science: The curriculum materials of the CASE project* (3rd ed.). London: Nelson Thornes.

Birmingham City Council (1997). *Signposts*. Windsor, UK: NFER-Nelson.

Carroll, J. B. (1993). *Human cognitive abilities*. Cambridge, UK: Cambridge University Press.

Cohen, L., & Manion, L. (1994). *Research methods in education* (4th ed.). London: Routledge.

Cronbach, L., & Furby, L. (1970). How should we measure change, or should we? *Psychological Bulletin, 74*, 68–80.

Forsman, J., Adey, P., & Barber, J. (1993, April). *A thinking science curriculum.* Paper presented at the American Association for the Advancement of Science, Boston, MA.

Hautamäki, J., Kuusela, J., & Wikström, J. (2002, July). *CASE and CAME in Finland: "The second wave."* Paper presented at the 10th International Conference on Thinking, Harrogate, UK.

Lazar, I., & Darlington, R. (1982). Lasting effects of early education: A report from the consortium for longitudinal studies. *Monographs of the Society for Research in Child Development, 47*(2/3).

Mayo, E. (1945). *The social problems of an industrial civilization*. Boston, MA: Harward University.

Prieler, J. (2001, April). *The value of derivative scores.* Paper presented at the annual meeting of British Psychological Society, Glasgow, UK.

Raven, J. (1998). *Coloured progressive matrices*. Oxford, UK: Oxford Psychologist's Press.

Shayer, M. (1999). Cognitive acceleration through science education II: Its effect and scope. *International Journal of Science Education, 21*(8), 883–902.

Shayer, M., & Adey, P. (Eds.). (2002). *Learning intelligence: Cognitive acceleration across the curriculum from 5 to 15 years*. Milton Keynes, UK: Open University Press.

Shayer, M., Wylam, H., Küchemann, D., & Adey, P. (1978). *Science reasoning tasks*. Slough: National Foundation for Educational Research.

Stenhouse, L. (1975). *An introduction to curriculum research and development*. London: Heinemann Educational Books.

Sternberg, R. J. (2003). *Cognitive psychology*. Belmont, CA: Wadsworth.

Sylva, K., & Ilsley, J. (1992). The high scope approach to working with young children. *Early Education*, Spring, 5–7.

Vernon, P. E. (1971). *The structure of human abilities*. London: Methuen.

Zigler, E., & Burman, W. (1983). Discerning the future of early childhood intervention. *American Psychologist, 38*, 894–906.