The Suitability of External Control-Groups for Empirical Control Purposes: a Cautionary Story in Science Education Research

Franz-Josef Scharfenberg
University of Bayreuth

Franz X. Bogner
University of Bayreuth

Siegfried Klautke
University of Bayreuth

Abstract

This article deals with a specific effect in one external control group incorporated to account for any pretest bias in a more comprehensive cognitive achievement study in a gene technology lab (as part of a modified Solomon's four-group plan). We monitored 12th graders (N = 117) in two external groups without any intervention: a one-test group (n = 55) and a three-test group (n = 62). Both samples participated in identical tests which quantified the relevant knowledge of the lesson unit applied in the main study. The three-test group yielded an unexpected increase in achievement scores. Subsequent analysis revealed two subsamples: one with no changes, the other with an increase (although without an intervention took place). A likely reason for the latter situation may lie in the role of the teacher(s) involved who might have wish to avoid potential negative results in his/her class. Consequently, we recommend the application of a modified Solomon's four group plan in science education research in order to prevent the influence of teacher intervention in future empirical analyses.

Correspondence should be addressed to Franz-Josef Scharfenberg, Centre of Math & Science Education, University of Bayreuth, Institute of Biology Didactics, Universitätsstr. 30, D-95445 Bayreuth, Germany. Phone: ++49-921-55-2590; Fax: ++49-921-55-2696. Email: franz-josef.scharfenberg@uni-bayreuth.de.

Introduction

A typical aspect of empirical analyses within science education is the desire to control as many variables as possible (Keeves, 1998). In particular, potential confounding variables, (e.g., maturation), possible external influences, or merely test repetition can threaten the internal validity of a study (Campbell, 1963; Campbell & Stanley, 1963). Almost 60 years ago, Solomon introduced a form of experimental design which is today typically referred as the Solomon four-group plan (Solomon, 1949). It is not the only four-group design (e.g., Huck & Chuang 1977; Marlatt, Demming & Reid, 1973) nor is it as commonly used as some other four-group plans (Rosenthal & Rosnow, 1997). However, this design is the only one known to assess adequately the confounding effect of pretesting with regard to the independent variables of interest (Walton Braver & Braver, 1988). 'In this case, the process of measurement may change which has to be measured or repeating the measurement may enable subjects to perform more well' (Michel & Haight, 1996 [p. 367]). This effect is usually termed the 'pretest effect' (Bortz & Döring, 2001

[p. 505]), 'test reactivity' or 'pretest sensitization' (Huck & Chuang, 1977 [p. 409]), 'premeasurement sensitization' (Michel & Haight, 1996 [p. 367]), or 'memory carry-over' (McNemar, 1963 [p. 149]). The specific key element for the Solomon four-group design is that two groups (one experimental and one control group) perform a pre- and a posttest while the other two groups (again one experimental and one control group) receive no pretesting (Solomon, 1949). The comparison of the two control groups may unveil a potential pretest effect. Thus, such a design increases the degree of internal validity.

Although a number of studies report pretest effects, others do not (Rosenthal & Rosnow, 1997). Especially in the area of cognitive outcomes, Willson and Putnam (1982) described more (and additionally higher) pretest effects as given in studies of the affective domain. With regard to the delay between pre- and posttest larger effects were found for several days to two weeks. Additionally, these effects appeared to be larger for control groups than for treatment groups. Nevertheless, despite its advantages the Solomon four-group plan is underused (Walton Braver & Braver, 1988). Repeated calls for its application (e.g., Michel & Haight, 1996; Morgan 1997), especially in educational research (Cohen & Manion, 1994) have widely been ignored (e.g., Blanchard & Spence 1999). Walton Braver and Braver (1988) give four reasons for this: (i) the necessity of a higher number of subjects compared to simpler designs; (ii) the researcher's belief that pretest effects may not exist in his/her research arena; (iii) the greater difficulty of drawing conclusions due to the complexity of the design; and (iv) problems with regard to the statistical treatment of the results (e.g., Michel & Haight, 1996).

However, often in science education research as well as in our main study, it is impossible for investigators to perform experimental designs, because randomization is quite impracticable. Students in intact course groups allow only quasi-experimental designs (Cook & Campbell, 1979), and corresponding modifications of Solomons four-group plans have been described (e.g., Davies & Gould, 2000). With regard to the objective of our main study - monitoring the effectiveness of out-of-school laboratory work with regard to gene technology (Scharfenberg et al., in press) - many studies often lack the Solomon four-group design and indeed fail to include any special retest control (e.g., Killermann, 1998;Yager, Engen & Snider, 1969). Wilson and Putnam (1982) claimed that 'nonrandomized studies with pretests must be viewed with additional suspicion' (p. 256). They assume a potential bias due to pretesting likely be caused by the quasi-experimental selection of subjects per se. In order to counter this potential pretest effects we decided to incorporate in our quasi-experimental design two external control groups with no intervention: a three-test and a one-test group (Table 1). The specific objective of this present study thus is to investigate the suitability of such external control groups in science education research.

Methods

Design of the study

Our main study followed a quasi-experimental design (Cook & Campell, 1979) providing a modified Solomon four-group plan (Solomon, 1949). We combined three treatment groups within a comprehensive study and two external control groups without intervention (Table 1).

Table 1:
Design providing a modified Solomon four-group plan

		Test schedule						
Group		T-1	One week	Treatment		T-2	Six weeks	T-3
Treatment 1	Hands-on lab	O_1	→	X_1	→	O_2	→	O_3
Treatment 2	Nonexperimental lab	O_1	→	X_2	→	O_2	→	O_3
Treatment 3	Nonexperimental school	O_1	→	X_3	→	O_2	→	O_3
External control 1	Three-test	O_1	→			O_2	→	O_3
External control 2	One-test					O_2		

Note. O_n Outcome measure at test schedule T-n: T-1 pre-, T-2 posttest, T-3 retention test.

The objective of our main study was a quasi-experimental comparison of three instructional approaches. Our main method of instruction was a hands-on approach with a sequence of minds-on and hands-on phases in a dedicated out-of-school laboratory offered by us at the university. Two parallel methods covered identical contents but without experimenting (either in the laboratory or at school); in both cases, the content of the experimental lesson was taught in a problem-oriented learning modus (Reigeluth & Moore, 1999), but theoretically. We monitored cognitive achievement with respect to the upgrade of existing prior knowledge and to the acquisition of new knowledge. This was done in order to focus on the learning location effect (school vs. out-of-school lab without experiments) and of the experimentation itself (with experiments in the lab vs. non-experimental instruction in the lab or at school; for more details see, Scharfenberg et al., in press).

Students' sample

In all, 34 biology courses with 12^{th} graders (N = 418; course size M = 12, SD = 3.7; age M = 18.0, SD = 0.68) participated in our main study. In order to establish similar courses in the different groups, we used only A-level ('Leistungskurs') students of the highest stratification level ('Gymnasium') in Bavaria (Germany). Additionally, all students have been enrolled in a regular half-year genetic course at school before participation in the study. This genetic education provided comparability of the courses: (i) The Bavarian Ministry of Education, Science, and Art (1991) obliges its content by the current syllabus; (ii) genetic education for all courses will be finished by a centralized formal exam at the end of high school. In general, the five groups did not differ in their prior achievement in biology (quantified as standard of written

school work), Kruskal-Wallis-test $\chi^2(4, N = 394) = 3.65$, $p = .454$, and experiences with experimentation at school, Kruskal-Wallis-test $\chi^2(4, N = 404) = 9.17$, $p = .057$.

The hands-on group ($n = 146$) attended our teaching unit at the out-of-school laboratory. The day-long module "marker genes in bacteria" integrated four experiments into a lab lesson conformant with the syllabus. In general, the students worked in groups, mainly 3- or 4-person groups dependent of the course size actually given ($M = 13$, $SD = 4.0$). They transformed bacteria with a recombinant plasmid (coding for the Green Fluorescent Protein, Tsien, 1998), they isolated the plasmid and analyzed it with common restriction enzymes. At least, they visualized their results by agarose gel electrophoresis. All experiments followed the criteria of authentic inquiry (Chinn & Malhotra, 2002). The nonexperimental lab group ($n = 72$) followed the same themes at the lab-site but without hands-on experiments. The school group ($n = 83$) was taught the identical content at school (again without experimental activities). A single teacher previously unknown to all students taught all lessons. A consistent problem of studies comparing experimental and nonexperimental instruction lies in the students' different time exposures. This problem has often been ignored (e.g., Killermann, 1998); others (e.g., Saunders and Dickinson 1972, p. 461) used actions like "discussion of material presented in lecture" in order to achieve identical time schedules. Following this rationale, we included a nonexperimental "lab+time" group in our pilot study one year ago and provided a typical lab working environment in combination with printed information which allowed repetition of the themes taught. However, cognitive learning outcomes were similar (Scharfenberg, 2005), and we omitted this kind of treatment in our main study. The results of the three treatment groups have been described elsewhere (Scharfenberg et al., in press).

With regard to the control groups, we modified the Solomon four-group plan by omitting the control group with treatment and one test (posttest after treatment) because of the impossibility of organising three such groups with regard to each treatment. Altogether, 11 courses ($N = 117$) were assigned to the two external control groups: a three-test ($n = 62$) and a one-test group ($n = 55$). They received no corresponding instruction to permit us to examine potential pretest effects or other external influences (Hofstein & Lunetta, 1982; Keeves, 1998), but proceeded with the regular lessons being taught by their teachers.

Assessment of the questionnaire used

Generally, questionnaires were applied three times, as pretest (T-1) one week before participating, as posttest (T-2) immediately after and as retention test (T-3) six weeks later. In contrast, the one-test group responded only once to the test. The questionnaire covered cognitive achievement items dealing with the lesson content, and consisted of 15 multiple-choice and one open item (see, e.g., Table 2). In our main study we applied two levels of analysis: (i) one dealing with a student's expected task performance such as reproduction (rendering of facts from memory; seven items), reorganization (self-acting rearrangement of facts to a new knowledge structure; four items) and transfer of knowledge (self-acting application of known facts to an unknown example; five items; all definitions by Deutscher Bildungsrat, 1970); (ii) the other with content relation referring to testing updated prior knowledge (seven items) and newly attained knowledge (nine items) validated by a latent class analysis on students' individual response pattern (for details, see, Scharfenberg et al., in press).

Table 2:
Listing of five item examples providing the achievement survey

	Item	Item difficulty[a]	Item characterization[b]	
			Expected task performance	Content relation
1	A plasmid for heterologous gene expression has to be constructed. Transformed bacteria will express the heterologous gene, if the following DNA segments are arranged in this way: a) antibiotic resistance gene → inserted gene → origin of replication. b) origin of replication → inserted gene → bacterial promotor sequence. c) bacterial promotor sequence → inserted gene → antibiotic resistance gene [correct]. d) inserted gene → origin of replication → bacterial promotor sequence.	30,1	Transfer	Updated prior knowledge
2	If an operon is positively controlled the 'switching molecule' starts a) translation of DNA. b) translation of m RNA. c) transcription of m RNA. d) transcription of DNA [correct].	49,7	Re-production	Updated prior knowledge
3	Green fluorescent protein (GFP) can be used in molecular biology in different ways because a) it is easy to detect its infrared fluorescent high emission. b) it is easy to generate fusion proteins with GFP and other proteins [correct]. c) it can diffuse in an organism form one cell to another. d) its luminescence component alone is useful.	52,1	Reorgani-sation	Newly attained knowledge
4	An electrophoresis apparatus consists of the following parts: a) one electrode, two buffer chambers, one gel carrier. b) two electrodes, two buffer chambers, two gel carriers. c) two electrodes, one buffer chamber, two gel carriers. d) two electrodes, two buffer chambers, one gel carrier [correct].	77,1	Re-production	Newly attained knowledge
5	A plasmid contains three recognition sites for the restriction enzyme *Bam* HI and one recognition site for the restriction enzyme *Eco* RI. How many fragments will result in a double digest with these two enzymes [four]?	46,9	Transfer	Newly attained knowledge

Note. [a]Item difficulty = % of correct answers (Bortz & Döring 2001).
[b] Two level of analysis: see text for details.

The consistency with the existing syllabus (Bavarian Ministry of Education, Science, and Art, 1991) provided content validity; in-service teachers (N = 12) provided an affirmative expert rating (similarity of the lesson to the syllabus as *good* or *excellent*). Positive correlations betweens students' test scores without any intervention with their prior achievement in school (quantified as their standard of written school work in biology) supported the convergent validity (as criterion-related validity type; see, e.g., Bortz & Döring, 2001) of the questionnaire for knowledge assessment at all (Spearman rank correlation coeffizient T-1 r_s = .320, p = .011, n = 62, T-2 r_s = .231, p = .013, N = 117, T-3 r_s = .344, p = .006, n = 62). Cronbach's alpha of the test scores was .68 (T-2, N = 418). The item difficulties were normally distributed (Kolmogorov-Smirnov test with Lilliefors modification p = .200, see Figure 1a) and the corrected item-total correlations were for 9 items > .3 (p < .001) and for 7 items > .2 (p < .001). We accepted the latter in spite of the low value (below .3) because of the complexity of the lesson content involved (Diehl & Kohr, 1999). Additionally, the corrected item-total correlations relate to item difficulties in a parabolic way (Lienert, 1969, see Figure 1b), a fact that item selection has to take into account (Bortz & Döring, 2001).

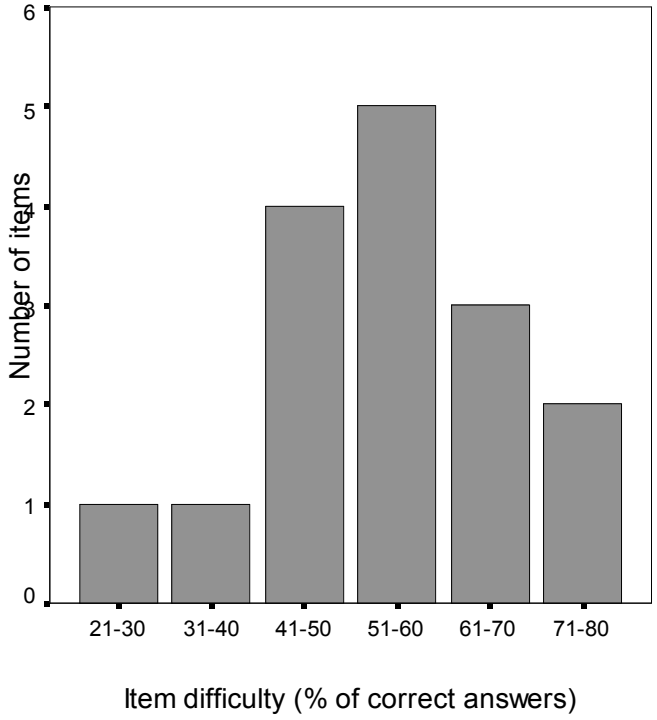

Figure 1a:
Distribution of questionnaire item difficulties with regard to patterns of ten units.

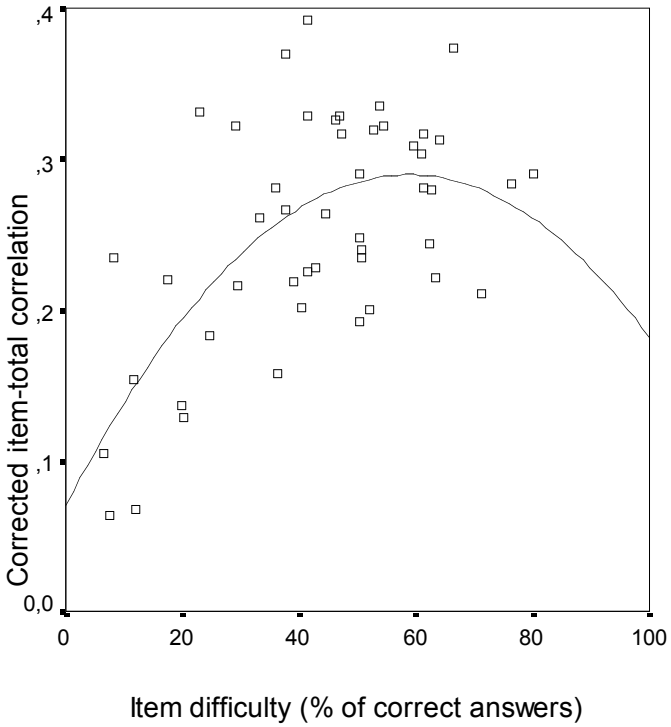

Figure1b:
Example of the parabolic relation between item difficulty and corrected item-total correlation (48 items of T-1, T-2, T-3, c.f. Lienert, 1969)

Statistical procedures

For each test and student, a total score was calculated as the number of correct answers. Due to the partial lack of normal distribution of our data, nonparametric methods were applied (Kolmogorov-Smirnov test with Lilliefors modification three-test group T-1, $p = .036$, T-2, $p = .195$, T-3, $p = .019$; one-test group T-2, $p = .001$). Consequently, we used boxplots as graphical charts. The statistical significance of changes of scores within all three test schedules was analysed using the Friedman-test, followed by pair-wise analyses from T-1 to T-2 and T-3 and from T-2 to T-3 using the Wilcoxon signed-rank test. The Mann-Whitney-U test was employed to test for pair-wise intergroup differences. An alpha level of .05 was used for all statistical tests.

Results

Our data analysis revealed an (unexpected) result with regard to the three-test control group, i.e. an increase of knowledge despite the lack of intervention, Friedman test $\chi^2(2, n = 62) = 8.673$, $p = .013$ (Figure 2).

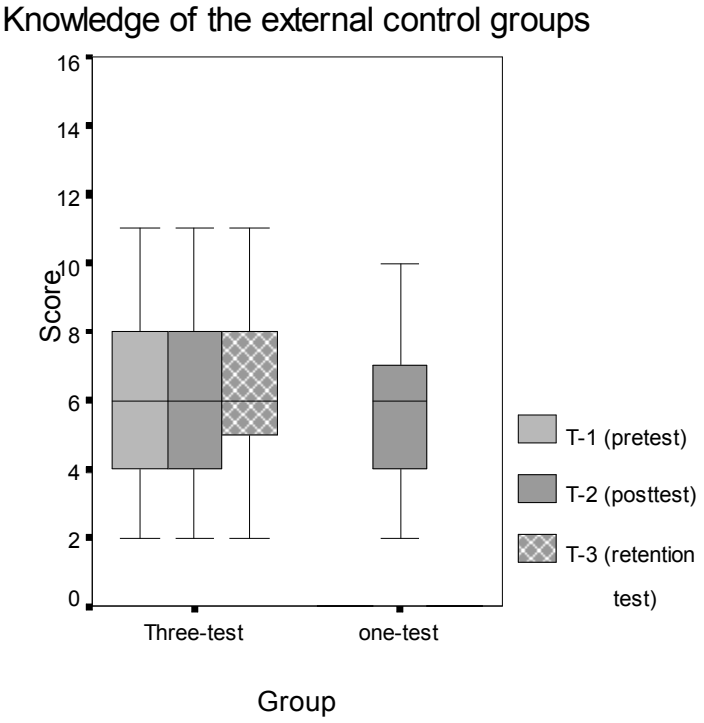

Figure 2:
Knowledge scores of the three-test and the one-test control groups without any intervention

Subsequent pair-wise analysis showed a statistically significant change from T-1 to T-3 (Wilcoxon signed-rank test Table 3).

Table 3:
Comparison of the one-test and the three-test group with regard to knowledge

	Control groups[a]			
Statistics	One-test	Three-test	Three-test	Three-test
Test dates	T-2	T-1 to T-2	T-2 to T-3	T-1 to T- 3
Mdn (grouped)	5.4	5.6 to 6.2	6.2 to 6.4	5.6 to 6.4
Z^a	-	1.857	1.204	2.137
p	-	.063	.228	.033

Note. [a]*In total N = 117, one-test group n = 55, three-test group n = 62.*
[b]*Wilcoxon signed-rank test (based on negative ranks).*

Comparison of the one-test group scores with the scores of the three-test group revealed a statistically significant difference only at the testing schedule T-3 (Mann-Whitney-U-test T-3, $p = 0.029$, Table 4).

Table 4:
Comparison of the two external control groups with regard to knowledge

	Comparison of the one-test group with the three-test group				
Statistics	as a whole			subsample-1	subsample-2
Test schedule	T-1	T-2	T-3	T-3	T-3
Mann-Whitney-*U*	1630.000	1442.500	1308.500	944.000	272.500
Z	0.414	1.448	2.188	0.378	4.521
p	.679	.148	.029	.705	<.001

Note: *Mann-Whitney-U-test shows statistically significant differences at T-3 between the one-test group and the three-test group as a whole and its subsample-2, in contrast to subsample-1 (see Figure 3).*

Subsequent analysis of the three-test group on the level of individual courses indicated a distinction of two separate subsamples (Figure 3), one with no significant change over the three survey schedules, the other with substantial change, Friedman-test subsample-1 χ^2 (2, $n = 36$) = 1.350, $p = .509$; subsample-2 χ^2 (2, $n = 26$) = 27.482, $p < .001$.

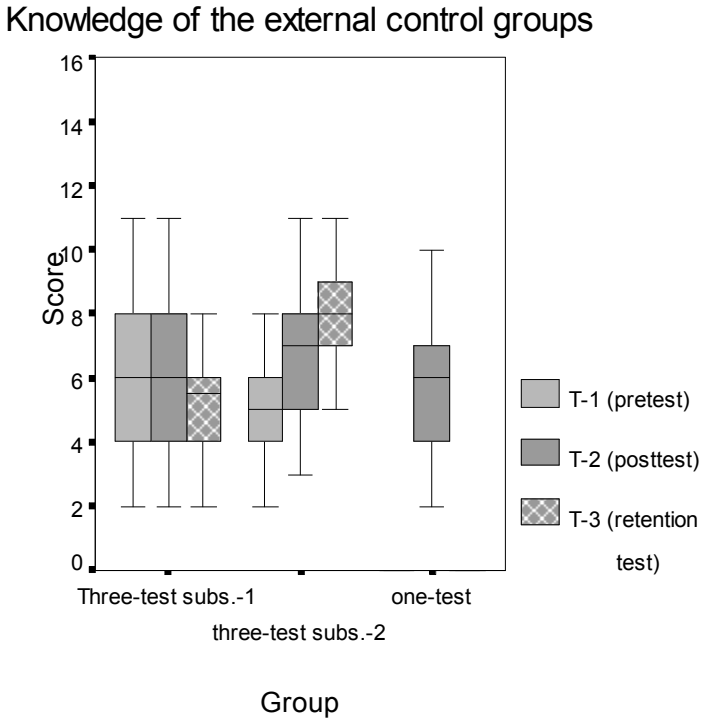

Figure 3:
Test scores of two subsamples (= subs.) extracted from the three-test group and compared to the score of the one-test group

Subsample-2 differed statistically significantly from the one-test group at T-3 (Mann-Whitney-U-test $p < .001$, Table 4). A comparison of both subsamples' scores on the three different tests also showed a significant difference at the testing schedule T-3 (Mann-Whitney-U-test $p < .001$, Table 5).

Table 5:
Comparison of both subsamples of the three-test group with regard to knowledge at the different test schedules

	Test schedule		
Statistics	T-1	T-2	T-3
Mann-Whitney-*U*	433.000	405.000	166.000
Z	0.504	0.907	4.353
p	.615	.365	<.001

Note. *Subsample-1 with no significant change over the three survey schedules (n = 36), subsample-2 with substantial change (n = 26).*

A subsequent pair-wise analysis of subsample-2 (Wilcoxon signed-rank tests, Table 6) showed significant changes across the three distinct survey dates. Despite the lack of intervention pupils of this specific subsample achieved a step by step increase in their level of knowledge, especially in the final T-3 (the retention test six weeks later). We conclude that some unknown factor must have affected the knowledge level of this subsample of the three-test control-group.

Table 6:
Cognitive achievement in subsample-2 of the three-test group

Statistics	Three-test group subsample-2[a]		
Test dates	T 1 to T 2	T 2 to T 3	T 1 to T 3
Mdn (grouped)	5.5 to 6.5	6.5 to 7.9	5.5 to 7.9
Z^a	3.063	3.689	3.811
p	.002	<.001	<.001

Note. [a] *n = 26*.
[b] *Wilcoxon signed-rank test (based on negative ranks).*

Discussion

The result of the three-test group in subsample-2 was quite unexpected, and we have no explanation for the surprising increase of achievement scores. However, we agree with Keeves (1998) that 'there is more information available in most well-designed evaluation studies', especially, when instruments used have the potential to be more generally introduced besides the original intent. Consequently, an external group scoring may enable us to gain additional insights into potential effects with regard to suitability of control-group.

Although many studies of the efficacies of laboratory activities rarely employ the Solomon four-group design or indeed any special retest control at all (e.g., Killermann 1998; Yager et al., 1969), we incorporated such a modified design to take account of potential pretest effects. Neither pupils nor teachers had any contact either with each other or with any course used as treatment group in our main study. Such contact was excluded due to the distances between the individual testing sites and the general survey schedule. Furthermore, the control courses did not take part nor planned to do so in any educational laboratory elsewhere.

The significant difference at T-3 between the one-test group and the three-test group as a whole firstly might hint at an effect of repeated testing. There may be a learning effect of participation in a pretest (T-1) that carries over to a following posttest (T-2). However, we found no difference between the one-test group and subsample-1 of the three-test group, suggesting that there is no bias due to repeated testing.

The unexpected gain in achievement level in subsample-2 of the three-test group may be attributed either to the students alone or to the teachers as well. With regard to the students, Cook and Campbell (1979) suggest that maturational factors may cause the increase in scores. This argument seems improbable since maturation may be assumed common to the sample as a whole and not specifically affect just this subgroup. Although students were unaware of the repeated testing schedules, the scores increased over time despite a gap of about six weeks between the posttest and the retention test. The second major explanation is the introduction of possible bias by the teacher: We have no evidence for this, but the hypothesis cannot be excluded as source of the disruptive influence.

Neither can we exclude the possibility of external influences via regional or supra-regional media. Schweiger and Brosius (1999), for instance, described the potential influence in their external control-group of news concerning the possible cloning of humans on pupils' specific attitudes towards gene technology. This too we find unlikely because all groups would be subject to such influences. Consequently, an intervention by the teacher seems more likely whether it occurred unconsciously or deliberately. Thus, the students may have been prepared for a routine assessment test perhaps by repetition of the selected knowledge necessary for the survey. A step-wise achievement effect is also feasible, resulting from a teacher either specifically preparing his/her pupils for the surveys or simply announcing that a second survey was to be conducted, thus motivating his/her pupils to recall previous test details in order to achieve better results.

A further explanation may be social desirability. Some indication of this is the fact that the investigator and the test administrator differed in the control-group surveys (in contrast to the treatment-groups). In all external control-groups the teachers acted as mediators. Some teachers in the three-test group may not have intervened with their samples. Mediators might have a specific interest in conflict with the investigators' interests in adhering to the standards of an empirical study. A teacher as mediator may intercede in order to get better results on his/her pupils as a desirable social objective, particularly if there is repeated testing. Two potential reasons may explain this. Firstly, a mediator may fear that investigators may get a poor impression of his/her capability. Secondly, he/she may have doubts about the anonymity of a survey, and hence fear bringing shame on his/her school.

Conclusions

With regard to our specific results of this study, we encourage the use of one-test groups in quasi-experimental research designs following a modified Solomon four-group plan. This might facilitate the identification of the above mentioned pretest effects in this design (Wilson & Putnam, 1982), especially such a form of retest effect as the one we observed in our subsample-2 of the three-test group. Furthermore, the selection of external control-groups needs careful action to exclude or at least reduce the influence of mediators' own interests. For instance, investigators could explicitly refer to the survey's anonymity and/or point out that the analysis of particular courses would be irrelevant to the study. We therefore emphasize to the complex issue of control-groups' selection in general, and to the necessity of ensuring that those control-groups function as intended. Taken this in account, a quasi-experimental design might provide more convincing empirical results with regard to the treatment groups as it is the case in our main study: Compared to the conventional learning location at school, hands-on activities with authentic experiments in out-of-school laboratories supported a substantial increase in knowledge (Scharfenberg et al., in press). As a consequence for science teaching, we suggest to offer teachers such out-of-school laboratories, especially, when authenticity is available (which is impossible to achieve at school). Nevertheless, any out-of-school experiment should be integrated within a teaching framework in a laboratory situation, thus, enabling students to actualize existing prior knowledge (as a precondition for the attainment of new knowledge). In our case, a specific teaching and learning unit assisted our students to develop individual hypotheses before participating in any hands-on activity and to verify/falsify his/her hypotheses. We appreciate, of course, the same frame in the context of demonstration experiments, either in a lab or in a school situation.

Acknowledgement

The study was funded by the Bavarian State Ministry of Regional Development and Environmental Affairs (Bayerisches Staatsministerium für Landesentwicklung und Umweltfragen), the Bavarian State Ministry of Education (Bayerisches Staatsministerium für Unterricht und Kultus) and the German Science Foundation (Deutsche Forschungsgemeinschaft; grant KL 664/5-1).

We welcome the cooperation of teachers and pupils involved in this study. Additionally, we appreciate the helpful and valuable discussion of the manuscript with C. Lehner, M. Wiseman and S. Tomkins.

References

Bavarian Ministry of Education, Science, and Art (Bayerisches Staatsministerium für Unterricht, Kultus, Wissenschaft und Kunst, Eds., 1991). Lehrplan für das bayerische Gymnasium, Fachlehrplan Biologie [Syllabus of the Bavarian secondary school at highest stratification level, Specific Syllabus of Biology]. *Amtsblatt des Bayerischen Staatsministeriums für Unterricht, Kultus, Wissenschaft und Kunst*, *So.-Nr. 7*, 1125-1172.

Blanchard, C., & Spence, J. C. (1999). The effect of pretesting on felling states and self-efficacy in acute exercise. *Research Update Alberta Centre for Active Living*, *7* (1), 1-2.

Bortz, J., & Döring, N. (2001). *Forschungsmethoden und Evaluation* (3rd ed.) [Methods of research and evaluation]. Berlin, Germany: Springer.

Campbell, D. T. (1963). From description to experimentation: interpreting trends as quasi-experiments. In: C. W. Harris (Ed.), *Problems in measuring change* (pp. 212-244). Madison, WI: The University of Wisconsin Press.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Dallas, TX: Houghton Mifflin Company Boston.

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

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation. Design & analysis issues for field settings*. Chicago: Rand McNally College Publishing Company.

Davies, N., & Gould, D. (2000). Updating cardiopulmonary resuscitation skills: a study to examine the efficacy of self-instruction on nurses' competence. *Journal of Clinical Nursing*, *9*, 400-410.

Deutscher Bildungsrat (Eds.) (1970). *Empfehlungen der Bildungskommission: Strukturplan für das Bildungswesen* [Recommendations of the Commision for Education: Curriculum for he educational system]. Stuttgart, Germany: Klett.

Diehl, J. M., & Kohr, H. U. (1999). *Deskriptive Statistik* (12th ed.) [Descriptive statistics]. Eschborn, Germany: Verlag Dietmar Klotz.

Hofstein, A., & Lunetta, V. N. (1982). The role of the laboratory in science teaching: neglected aspects of research. *Review of educational research*, *52,* 201-217.

Huck, S. W., & Chuang, I. C. (1977). A quasi-experimental design for the assessment of posttest sensitization. *Educational and Psychological Measurement*, *37*, 409-416.

Keeves, J. P. (1998). Methods and processes in research in science. In: B. J. Fraser & K. G. Tobin (Eds.), *International Handbook of Science Education, Part Two* (pp. 1127-1153). Dordrecht, Netherlands: Kluwer Academish Publishers.

Killermann, W. (1998). Research into biology teaching methods. *Journal of Biological Education*, *33* (1), 4-9.

Lienert, G. A. (1969). *Testaufbau und Testanalyse* (3rd ed.) [Development and analysis of tests]. Weinheim, Germany: Verlag Julius Beltz.

Marlatt, A. G., Demming, B., & Reid, J. B. (1973). Loss of control drinking in alcoholoics: a experimental analogue. *Journal of Abnormal Psychology*, *81* (3), 233-241.

McNemar, Q. (1963). *Psychological statistics* (3rd ed.). New York: John Wiley and Sons.

Michel, Y., & Haight, B. K. (1996). Using the Solomon four-group design. *Nursing research*, *45* (6), 367-369.

Morgan, W. (1997). *Physical activity and mental health*. Washington, DC: Taylor & Francis.

Reigeluth, C.M., & Moore, J. (1999). Cognitive Education and the Cognitive Domain. In: Reigeluth, C. M. (Ed.), *Instructional-Design Theories and Models Volume II* (pp. 51-68). Mahwah, NJ: Lawrence Erlbaum Associates.

Rosenthal, R., & Rosnow, R. L. (1997). *Essentials of behavioural research. Methods and data analysis* (2nd ed.). New York: McGraw-Hill.

Saunders, W. L., & Dickinson, D. H. (1976). A comparison of community college students' achievement and attitude. Changes in a lecture-only and lecture-laboratory approach to general education (biological science courses). *Journal of Research in Science Teaching 16* (5), 459-464.

Scharfenberg, F.-J., 2005 *Experimenteller Biologieunterricht zu Aspekten der Gentechnik im Lernort Labor: empirische Untersuchung zu Akzeptanz, Wissenserwerb und Interesse* [Hands-on teaching in biology in an out-of-school laboratory with regard to gene technology: empirical study with respect to acceptance, knowledge achievement and

interest]. Doctoral dissertation, University of Bayreuth, Germany. Retrieved December 4 from http://opus.ub.uni-bayreuth.de/volltexte/2005/176/ .

Scharfenberg, F.-J., Bogner, F., & Klautke, S. (in press). Learning in a gene technology lab with educational focus: Results of a teaching unit with authentic experiments. *Biochemistry and Molecular Biology Education*.

Schweiger, W. & Brosius, H.-B. (1999). *Von der 'Gentomate' zur Gentechnikakzeptanz. Eine Panelstudie zu Einstellungseffekten eines rollenden Genlabors an Gymnasien* [From the 'gene tomato' to acceptance of gene technology. A panel study to effects of a mobile gene technology lab at secondary schools with regard to attitudes]. Neuherberg, Germany: Gesellschaft für Strahlenschutzforschung.

Solomon, R. L. (1949). An extension of control group design. *Psychological Bulletin*, *46*, 137-150.

R. Y. Tsien (1998) The green fluorescent protein, *Annual Review of. Biochemistry*. *67*, 509-544.

Walton Braver, M. C., & Braver, S. (1988). Statistical treatment of the Solomon four-group design: a meta-analytic approach. *Psychological Bulletin*, *104* (1), 150-154.

Willson, V. L., & Putnam, R. R. (1982). A meta-analysis of pretest sensitization effects in experimental design. *American Educational Research Journal*, *19* (2), 249-258.

Yager, R. E., Engen, H. B., & Snider, B. C. (1969). Effects of the laboratory and demonstration methods upon the outcomes of instruction in secondary biology. *Journal of Research in Science Teaching*, *6* (5), 76-86.