What Works Clearinghouse™

ies INSTITUTE OF EDUCATION SCIENCES

October 2012

WWC Review of the Report "Learning the Control of Variables Strategy in Higher and Lower Achieving Classrooms: Contributions of Explicit Instruction and Experimentation"[1]

The findings from this review do not reflect the full body of research evidence on teaching the *control of variables strategy (CVS)*.

What is this study about?

The study examined three separate methods for teaching the *control of variables strategy (CVS)*, a procedure for conducting a science experiment so that only one variable is tested and all others are held constant, or "controlled."

The study analyzed data from a randomized controlled trial of 848 fourth-grade students in 39 classrooms in 12 schools in Fayette County, Kentucky. Half of the classrooms were from five schools that, in the previous year, scored the highest in the district on the science portion of the Kentucky Core Content Test (KCCT); the rest of the classrooms were from seven of the eight lowest achieving schools in the district.

Classrooms with similar achievement levels (within schools, when possible) were formed into triplets. Within each triplet, classrooms were randomly assigned to one of three conditions:

- *Instruct:* Teachers taught *CVS* in an interactive lecture format;
- *Manipulate:* Teachers taught *CVS* by providing time for students to design and run experiments in groups; or
- *Both:* Teachers taught *CVS* through both interactive lectures and by providing time for experimentation in groups.

The study assessed the impact of each of the three strategies by testing student understanding of the concepts at three points: the day before instruction, the day after instruction, and five months after instruction.

WWC Rating

The research described in this report meets WWC evidence standards without reservations

Strengths: The study is a well-implemented randomized controlled trial.

Features of Teaching the *Control of Variables Strategy (CVS)*

This study assessed the impact of three methods used to teach CVS. In this study, CVS is a means to teach students how to design and conduct controlled experiments that will give them valid and interpretable results.

In this study, students learned components of *CVS* through experiments that tested the effect of four control variables (ramp steepness, ramp smoothness, ball starting point, age of ball) on the distance a ball would roll according to the ramp specification. As a result of conducting a series of experimental trials with different combinations of the control variables, students learn how to test the effect of one of the control variables while holding all other variables constant, which is the basis for valid scientific experimentation.

What did the study find?

The study found, and the WWC confirmed, statistically significant differences in student performance on the *CVS comparison assessment* at posttest among the three conditions. Students in the *Both* condition outperformed students in the *Manipulate* condition and

the *Instruct* condition, and students in the *Instruct* condition outperformed students in the *Manipulate* condition. In addition, students in the *Both* condition outperformed students in the *Manipulate* condition on the *ramps test*. These findings indicate that using a combination of interactive lectures and manipulative experiments was the most effective method of teaching *CVS*.

Appendix A: Study details

Lorch, Jr., R. F., Lorch, E. P., Calderhead, W. J., Dunlap, E. E., Hodell, E. C., & Freer, B. D. (2010). Learning the control of variables strategy in higher and lower achieving classrooms: Contributions of explicit instruction and experimentation. *Journal of Educational Psychology, 102*(1), 90–101.

Setting

The study was conducted in 39 fourth-grade classrooms in 12 schools in Fayette County, Kentucky.

Study sample

Half of the classrooms in this study were from five schools that, in the previous year, scored the highest in the district on the science portion of the Kentucky Core Content Test (KCCT); the rest of the classrooms were from seven of the eight lowest achieving schools in the district. The study stratified schools by these high/low achievement levels, and classrooms were formed into triplets. If a school had three classrooms, the classrooms were treated as a triplet; otherwise, three classrooms from two or more schools with similar characteristics were treated as a triplet. Within each triplet, classrooms were randomly assigned to one of three conditions. On average, the student population of the sample schools included 43.4% minority students and 51.3% of students eligible for free or reduced-price lunch. The original sample included a total of 848 students in 39 classrooms (288 students in 13 classrooms in the *Both* condition, 294 students in 13 classrooms in the *Instruct* condition, and 266 students in 13 classrooms in the *Manipulate* condition). The analytic sample for the immediate posttest included 673 students (233 students in the *Both* condition, 227 students in the *Instruct* condition, and 213 students in the *Manipulate* condition) in 36 classrooms (12 in each condition).[2] The analytic sample for the delayed posttest included 617 students (215 students in the *Both* condition, 208 students in the *Instruct* condition, and 194 students in the *Manipulate* condition) in 36 classrooms (12 in each condition).

Intervention group

On Day 1 of the study, students in the *Both* condition took the *comparison test* as a pretest and then received instructions and a demonstration using ramps and balls on how to conduct an experiment to test the effect of four control variables (ramp steepness, ramp smoothness, ball starting point, age of ball) on the distance a ball would roll according to the ramp specification. They then spent from 30–45 minutes in groups conducting four experiments with the ramps, and ended the class by completing a *ramps test* as a group. On Day 2, the instructor delivered a 15–20 minute lesson on *CVS* that included two examples of valid and invalid experiments, as well as explicit instruction, after which students repeated the experiments from Day 1 and completed another *ramps test*. On Day 3, students took the *comparison posttest*, which was identical in format to the corresponding pretest on Day 1 but with different content.

The same procedure was followed in the *Instruct* condition as in the *Both* condition, except that students never conducted group experiments using the ramps (on Day 1 or 2), and therefore, the intervention session was shorter for these students on these days.

The procedure for the *Manipulate* condition was identical to the *Both* condition, except that on Day 2, students in the *Manipulate* condition did not receive the *CVS* lesson. In its place, students were reassembled into their groups and were told to do whatever experiments they wished to try to learn how the four variables affected how far the balls rolled for the remainder of the intervention session, followed by a *ramps test.*

Comparison group

The study involved a head-to-head comparison of each of the three teaching methods described above.

Outcomes and measurement

On both Day 1 (pre) and Day 3 (post), students took the *comparison test*, a researcher-developed paper-and-pencil assessment of students' ability to evaluate the validity of experimental comparisons. Students in the *Manipulate* and *Both* conditions also took a *ramps test*, which involved the manipulation of ramps to determine how far a ball would roll. Five months after the completion of the intervention, students in all conditions took a delayed *comparison test.* For a more detailed description of these outcome measures, see Appendix B.

Support for implementation

Teachers were not involved in the delivery of the intervention. Two graduate research assistants served as instructors for the intervention, with the same instructor teaching each of the three conditions to classes assigned to the same triplet. (One instructor taught eight triplets for a total of 24 classes, while the other taught the remaining four triplets for a total of 12 classes.) For each classroom, the instructors visited participating classrooms on three consecutive days during the fall semester. In each classroom, the instructor was assisted by one or two helpers who distributed materials and answered students' procedural questions during experiments.

Reason for review

This study was identified for review by the WWC because it was supported by a grant to the University of Kentucky (Principal Investigator: Elizabeth Lorch) from the National Center for Education Research (NCER) at the Institute of Education Sciences (IES).

Appendix B: Outcome measures for each domain

Science achievement	
Comparison test	This researcher-developed instrument was a paper-and-pencil assessment of students' ability to distinguish valid and invalid experimental comparisons that was expected to require 20 minutes to complete. The test consisted of five items that assessed three particular example domains (for example, baking cookies, exercise, and growing plants). For each domain of interest, there were three variables that would affect the outcome of interest (e.g., when the domain was "baking cookies," the variables of interest were whether cookies were baked for 5 or 10 minutes, whether they were sweetened with sugar or honey, and whether they were baked with one or three eggs). To assess students' understanding of valid or invalid experimental designs, comparisons were depicted that manipulated the three variables of interest. For each comparison, students indicated whether the comparison was a good (i.e., valid) or bad (i.e., invalid) test of the effects of a specific variable. Within each domain, two of the comparisons were valid (40%), while the other three invalid comparisons (60%) were composed of one doubly confounded comparison (the two pictures had different values on all three variables), one singly confounded comparison (the two pictures had different values on two variables), and one noncontrastive comparison (the two pictures differed on only one variable, but it was not the variable being tested). For each test administration, the instructor read the first two pages of each item domain and then let students do the last four items in the domain on their own. Performance was scored for the total number of correctly answered questions (maximum = 15).
Ramps test	This researcher-developed test was administered to students in the *Manipulate* and *Both* conditions only. To complete this test, students rolled a ball down a "down" ramp onto an "up" ramp with numbered lines to indicate how far the ball rolled. The ramps could be manipulated in four ways: (1) the steepness of the ramps could be changed, (2) the surface of the ramps could be rough or smooth, (3) two starting points for the balls were presented, and (4) students could roll either a new yellow ball or an old white ball during the trial. Students worked in groups of three or four to run experiments using the ramps and recorded their work in booklets. Each of the three test booklets had the same format. At the top of the page in each booklet, students were presented the focal variable to test, and each page included a section to plan the experiment and a table to record results of the trials. Students were asked to predict how far each ball would roll, to draw a conclusion about whether the focal variable had an effect, and to record their confidence in their results. All four focal variables were included in the posttest. The outcome for the *ramps test* was a "group" score (based on the 3–4 students working together), rather than an individual score. The test results indicate the percentage of valid ramps experiments conducted.

Appendix C: Study findings for science achievement

Domain and outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Science achievement								
Comparison test (Both v. Instruct)	Grade 4	24 classrooms/ 460 students	10.62 (3.47)	9.57 (3.41)	1.05	0.30	+12	< 0.01
Comparison test (Both v. Manipulate	Grade 4	24 classrooms/ 446 students	10.72 (3.47)	8.37 (2.75)	2.35	0.75	+27	nr
Comparison test (Instruct v. Manipulate	Grade 4	24 classrooms/ 440 students	9.67 (3.47)	8.37 (2.75)	1.30	0.42	+16	< 0.01
Ramps test (Both v. Manipulate)	Grade 4	24 classrooms/ 148 students	61%	28%	33%	0.68	+25	< 0.05

Table Notes: For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on student outcomes, representing the change (measured in standard deviations) in an average student's outcome that can be expected if the student is given the intervention. The improvement index is an alternate presentation of the effect size, reflecting the change in an average student's percentile rank that can be expected if the student is given the intervention. The study is characterized as having a statistically significant positive effect because univariate statistical tests are reported for each outcome measure, the effect for at least one measure within the domain is positive and statistically significant, and no effects are negative and statistically significant. nr = not reported.

Study Notes: A correction for multiple comparisons was needed but did not affect significance levels. The *p*-values presented here were reported in the original study. The WWC calculated the intervention group mean for all comparison test outcomes by adding the hierarchical linear modeling (HLM) estimate of the impact of the program (i.e., difference in adjusted mean gains between the intervention and comparison groups) to the unadjusted comparison group posttest means (obtained via email from the authors). For the *ramps test*, the WWC calculated the intervention group mean by adding the difference-in-differences adjusted estimate of the average impact of the program (i.e., difference in mean gains between the intervention and comparison groups) to the unadjusted comparison group posttests means (obtained via email from the authors). Please see the *WWC Handbook* for more information. The impact of the program for the *Both v. Manipulate* contrast was derived by subtracting the HLM coefficients in Model 3, reported in Table 2 of the study. One school with three classrooms (one in each condition) was eliminated in the final analysis (obtained through email correspondence with the authors), which resulted in a total of 12 classrooms contributing to each condition in the table above.

Appendix D: Supplemental findings by achievement level

Domain and outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Science achievement: Delayed posttest								
Comparison test (Both v. Instruct)	Grade 4	24 classrooms/ 423 students	10.37 (3.53)	9.52 (3.56)	0.85	0.24	+9	0.01
Comparison test (Both v. Manipulate)	Grade 4	24 classrooms/ 409 students	10.54 (3.53)	8.72 (3.03)	1.81	0.55	+21	< 0.01
Comparison test (Instruct v. Manipulate)	Grade 4	24 classrooms/ 402 students	9.69 (3.56)	8.72 (3.03)	0.96	0.29	+11	< 0.01
Science achievement: Immediate posttest for students in low-achieving schools								
Comparison test (Both v. Instruct)	Grade 4: Low achieving schools	12 classrooms/ 247 students	9.46 (3.35)	8.83 (3.08)	0.63	0.19	+8	0.50
Comparison test (Both v. Manipulate)	Grade 4: Low achieving schools	12 classrooms/ 224 students	9.44 (3.35)	7.71 (2.06)	1.73	0.61	+23	0.04
Comparison test (Instruct v. Manipulate)	Grade 4: Low achieving schools	12 classrooms/ 233 students	8.81 (3.08)	7.71 (2.06)	1.10	0.41	+16	0.16
Ramps test (Both v. Manipulate)	Grade 4: Low achieving schools	12 classrooms/ 65 groups	52%	13%	39%	1.21	+39	< 0.01
Science achievement: Immediate posttest for students in high-achieving schools								
Comparison test (Both v. Instruct)	Grade 4: High achieving schools	12 classrooms/ 213 students	11.99 (2.98)	10.54 (3.58)	1.45	0.44	+17	0.13
Comparison test (Both v. Manipulate)	Grade 4: High achieving schools	12 classrooms/ 222 students	11.43 (2.98)	9.00 (3.17)	2.43	0.79	+28	0.01
Comparison test (Instruct v. Manipulate)	Grade 4: High achieving schools	12 classrooms/ 207 students	9.98 (3.58)	9.00 (3.17)	0.98	0.29	+11	0.32
Ramps test (Both v. Manipulate)	Grade 4: High achieving schools	12 classrooms/ 83 groups	68%	42%	26%	0.75	+27	0.03
Science achievement: Delayed posttest for students in low-achieving schools								
Comparison test (Both v. Instruct)	Grade 4: Low achieving schools	12 classrooms/ 221 students	9.42 (3.38)	8.80 (3.23)	0.62	0.19	+7	0.52
Comparison test (Both v. Manipulate)	Grade 4: Low achieving schools	12 classrooms/ 195 students	9.29 (3.38)	7.81 (2.25)	1.48	0.50	+19	0.09

Domain and outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Comparison test (Instruct v. Manipulate)	Grade 4: Low achieving schools	12 classrooms/ 204 students	8.67 (3.23)	7.81 (2.25)	0.86	0.30	+12	0.31
Science achievement: Delayed posttest for students in high-achieving schools								
Comparison test (Both v. Instruct)	Grade 4: High achieving schools	12 classrooms/ 202 students	11.39 (3.24)	10.41 (3.77)	0.98	0.28	+11	0.34
Comparison test (Both v. Manipulate)	Grade 4: High achieving schools	12 classrooms/ 214 students	10.92 (3.24)	9.50 (3.39)	1.43	0.43	+17	0.14

Table Notes: For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on student outcomes, representing the change (measured in standard deviations) in an average student's outcome that can be expected if the student is given the intervention. The improvement index is an alternate presentation of the effect size, reflecting the change in an average student's percentile rank that can be expected if the student is given the intervention.

Study Notes: The study did not report the *p*-values for any contrasts in this table; the *p*-values reported here were calculated by the WWC. A correction for multiple comparisons was needed, and the *p*-value for the *comparison test (Both v. Manipulate)* for the low-achieving schools subgroup was found to be nonsignificant. The WWC calculated the intervention group mean for the delayed posttest outcome by adding the HLM estimate of the impact of the program (i.e., difference in adjusted mean gains between the intervention and comparison groups) to the unadjusted comparison group posttest means (obtained via email from the authors). Please see the *WWC Handbook* for more information. The impact of the program for the *Both v. Manipulate* contrast was derived by subtracting the HLM coefficients in Model 3, reported in Table 2 of the study. For the *ramps test* and all subgroup estimates of the *comparison test*, the WWC calculated the intervention group mean by adding the difference-in-differences adjusted estimate of the average impact of the program (i.e., difference in mean gains between the intervention and comparison groups) to the unadjusted comparison group posttest means (obtained via email from the authors). Two subgroup contrasts for the delayed posttest outcome had high attrition: (1) Delayed posttest for students in low-achieving schools–*Manipulate v. Both*: this contrast demonstrated baseline equivalence and did not need statistical adjustment (and this contrast is therefore only eligible to meet WWC standards with reservations); (2) Delayed posttest for students in high-achieving schools–*Manipulate v. Instruct* contrast: this contrast was not reported because there was high attrition, and the authors' results did not adequately control for baseline differences (and therefore, this contrast does not meet WWC standards). One school with three classrooms (one in each condition) was eliminated in the final analysis (obtained through email correspondence with the authors), which resulted in a total of 12 classrooms contributing to each condition in the table above. For each of the low and high achieving subgroup contrasts described above, six classrooms contributed to each condition.

Endnotes

[1] Single study reviews examine evidence published in a study (supplemented, if necessary, by information obtained directly from the author[s]) to assess whether the study design meets WWC evidence standards. The review reports the WWC's assessment of whether the study meets WWC evidence standards and summarizes the study findings following WWC conventions for reporting evidence on effectiveness. This study was reviewed using the Science review protocol, version 2.0. The WWC rating applies only to the results that were eligible under this topic area and met WWC standards without reservations or met WWC standards with reservations, and not necessarily to all results presented in the study.

[2] One school with three classrooms (one in each condition) was eliminated in the final analysis. This information was obtained through email correspondence with the authors.

Recommended Citation

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2012, October). *WWC review of the report: Learning the control of variables strategy in higher and lower achieving classrooms: Contributions of explicit instruction and experimentation.* Retrieved from http://whatworks.ed.gov.

Glossary of Terms

Attrition Attrition occurs when an outcome variable is not available for all participants initially assigned to the intervention and comparison groups. The WWC considers the total attrition rate and the difference in attrition rates across groups within a study.

Clustering adjustment If intervention assignment is made at a cluster level and the analysis is conducted at the student level, the WWC will adjust the statistical significance to account for this mismatch, if necessary.

Confounding factor A confounding factor is a component of a study that is completely aligned with one of the study conditions, making it impossible to separate how much of the observed effect was due to the intervention and how much was due to the factor.

Design The design of a study is the method by which intervention and comparison groups were assigned.

Domain A domain is a group of closely related outcomes.

Effect size The effect size is a measure of the magnitude of an effect. The WWC uses a standardized measure to facilitate comparisons across studies and outcomes.

Eligibility A study is eligible for review if it falls within the scope of the review protocol and uses either an experimental or matched comparison group design.

Equivalence A demonstration that the analysis sample groups are similar on observed characteristics defined in the review area protocol.

Improvement index Along a percentile distribution of students, the improvement index represents the gain or loss of the average student due to the intervention. As the average student starts at the 50th percentile, the measure ranges from –50 to +50.

Multiple comparison adjustment When a study includes multiple outcomes or comparison groups, the WWC will adjust the statistical significance to account for the multiple comparisons, if necessary.

Quasi-experimental design (QED) A quasi-experimental design (QED) is a research design in which subjects are assigned to intervention and comparison groups through a process that is not random.

Randomized controlled trial (RCT) A randomized controlled trial (RCT) is an experiment in which investigators randomly assign eligible participants into intervention and comparison groups.

Single-case design (SCD) A research approach in which an outcome variable is measured repeatedly within and across different conditions that are defined by the presence or absence of an intervention.

Standard deviation The standard deviation of a measure shows how much variation exists across observations in the sample. A low standard deviation indicates that the observations in the sample tend to be very close to the mean; a high standard deviation indicates that the observations in the sample tend to be spread out over a large range of values.

Statistical significance Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups. The WWC labels a finding statistically significant if the likelihood that the difference is due to chance is less than 5% ($p < 0.05$).

Substantively important A substantively important finding is one that has an effect size of 0.25 or greater, regardless of statistical significance.

Please see the WWC Procedures and Standards Handbook (version 2.1) for additional details.