What Works Clearinghouse

Science June 2012

Great Explorations in Math and Science® (GEMS®) Space Science

Program Description[1]

Great Explorations in Math and Science® (GEMS®) Space Science is an instructional sequence for grades 3–5 that covers fundamental concepts, including planetary sizes and distance, the Earth's shape and movement, gravity, and moon phases and eclipses. Part of the *GEMS®* core curriculum[2], *GEMS® Space Science* uses the solar system as the focal point for learning. The sequence utilizes models, hands-on investigations, peer-to-peer discussions, reflection, and informational student readings. Students complete four units, each lasting between four and nine sessions. Each unit builds upon knowledge from previous units and can be used independently or in conjunction with one another for an overall learning progression.

Research[3]

One study of *GEMS® Space Science* that falls within the scope of the Science review protocol meets What Works Clearinghouse (WWC) evidence standards without reservations. The study included 2,594 elementary school students from grades 4 and 5 in elementary schools in Florida. Based on this study, the WWC considers the extent of evidence for *GEMS® Space Science* on elementary school students to be small for the general science achievement domain, the only domain identified by the review protocol.

Effectiveness

GEMS® Space Science was found to have potentially positive effects on general science achievement for elementary school students.

Table 1. Summary of findings[4]

Outcome domain	Rating of effectiveness	Improvement index (percentile points)		Number of studies	Number of students	Extent of evidence
		Average	Range			
General science achievement	Potentially positive effects	+7	na	1	2,594	Small

na = not applicable

Report Contents

Program Information

Background

The *GEMS® Space Science* sequence was developed at the Lawrence Hall of Science in collaboration with the National Aeronautics and Space Administration (NASA) and astronomy educators, researchers, and assessment experts. The Lawrence Hall of Science is the public science and mathematics curriculum development and educational research center of the University of California, Berkeley. The *GEMS® Space Science* sequence is available from the *GEMS®* publisher, Carolina Curriculum. Address: Carolina Biological Supply Company, 2700 York Road, Burlington, NC 27215-3398. Email: curriculum@carolina.com. Web: http://www.carolinacurriculum.com/GEMS/. Telephone: (800) 334-5551.

Program details

The *GEMS® Space Science* sequence for grades 3–5 introduces students to fundamental concepts in space science using the solar system as the foundation. Students investigate size and scale relative to distance, the Earth's shape and gravity, how the Earth moves, and moon phases and eclipses. The sequence has 24 sixty-minute class sessions broken down into four units:

- Unit 1: How Big and How Far? (9 class sessions)
- Unit 2: Earth's Shape and Gravity (6 class sessions)
- Unit 3: How Does the Earth Move? (4 class sessions)
- Unit 4: Moon Phases and Eclipses (5 class sessions)

The activities in the curriculum target core space science concepts and common misconceptions that students might have about them. Students explore the role of models and evidence in science. Working in small groups, students are encouraged to evaluate alternative explanations, use evidence to support them, and critique the merits of an explanation.

Educators may implement all the units in a single grade during one school year or teach individual units in consecutive grades over two or three years. Not all of the units in a sequence must be taught—each can stand alone, if necessary. The *GEMS® Space Science* curriculum comes with a teacher's guide, a materials kit, and master copies for duplication or electronic presentation. The teacher's guide includes an assessment system and a CD-ROM, which offers a collection of resources, software programs, and web links.

More than 60 *GEMS®* network sites and centers provide ongoing training and support to teachers on how to use *GEMS® Space Science* within their larger curriculum.

Cost

The *GEMS® Space Science Curriculum Sequence for Grades 3–5* kit is available for $515. The teacher's guide costs $149.95 (rates effective December 2011). Additional information is available from the program publisher, Carolina Curriculum.

Research Summary

Two studies reviewed by the WWC investigated the effects of *GEMS® Space Science* on elementary school students. One study (Granger, Bevis, Saka, & Southerland, 2010) is a randomized controlled trial that meets WWC evidence standards without reservations. That study is summarized in this report. The other study does not meet WWC eligibility screens. (See references beginning on p. 5 for citations for both studies.)

Table 2. Scope of reviewed research

Grade	4, 5
Delivery method	Small group/Whole class
Program type	Curriculum
Studies reviewed	2
Meets WWC standards without reservations	1 study
Meets WWC standards with reservations	0 studies

Summary of study meeting WWC evidence standards without reservations

Granger et al. (2010) conducted a cluster randomized controlled trial that examined the effects of *GEMS® Space Science* on students in grades 4 and 5 attending elementary schools in central Florida. The study used a two-step assignment procedure. Volunteer teachers were first matched on demographic characteristics and grade levels taught. They were then randomly assigned to use either the *GEMS® Space Science* sequence or the regular space science sequence offered in the district.

The teacher analysis sample included 66 teachers in intervention classrooms and 59 teachers in comparison classrooms. The student analysis sample included 1,418 students who received the *GEMS® Space Science* sequence and 1,176 comparison group students who received the typical space science instruction available in the district.

The study reported students' outcomes immediately following completion of the space science unit, and then again five months later. The WWC rating of effectiveness is based on the immediate posttest findings.

Summary of studies meeting WWC evidence standards with reservations

No studies of *GEMS® Space Science* meet WWC evidence standards with reservations.

Effectiveness Summary

The WWC review of interventions for Science addresses student outcomes in one domain: general science achievement. The domain includes three outcome constructs: life science, earth/space science, and physical science. The study that contributes to the effectiveness rating in this report covers one construct: earth/space science. The findings below present the authors' estimates and WWC-calculated estimates of the size and the statistical significance of the effects of *GEMS® Space Science* on elementary school students. For a more detailed description of the rating of effectiveness and extent of evidence criteria, see the WWC Rating Criteria on p. 12.

Summary of effectiveness for the general science achievement domain

One study reported findings in the general science achievement domain.

Granger et al. (2010) reported, and the WWC confirmed, statistically significant positive effects of *GEMS® Space Science* on the Space Science Content test for students in grades 4 and 5.

Thus, for the general science achievement domain, one study showed statistically significant positive effects. This results in a rating of potentially positive effects, with a small extent of evidence.

Table 3. Rating of effectiveness and extent of evidence for the general science achievement domain

Rating of effectiveness	Criteria met
Potentially positive effects *Evidence of a positive effect with no overriding contrary evidence.*	The review of *GEMS® Space Science* in the general science achievement domain had one study showing a statistically significant positive effect and no studies showing a statistically significant or substantively important negative effect or indeterminate effects.
Extent of evidence	**Criteria met**
Small	The review of *GEMS® Space Science* in the general science achievement domain was based on one study that included 2,594 students.

References

Study that meets WWC evidence standards without reservations

Granger, E. M., Bevis, T. H., Saka, Y., & Southerland, S. A. (2010, March). *Large scale, randomized cluster design study of the relative effectiveness of reform-based and traditional/verification curricula in supporting student science learning.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, Philadelphia, PA.

Additional source:

Granger, E. M., Bevis, T. H., Saka, Y., & Southerland, S. A. (2009, April). *Comparing the efficacy of reform-based and traditional/verification curricula to support student learning about space science.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, Garden Grove, CA.

Studies that are ineligible for review using the Science Evidence Review Protocol

Granger, E. M., Bevis, T. H., Saka, Y., & Southerland, S. A. (2010, March). *Comparing reform-based and traditional curricula in a large-scale, randomized cluster design study: The interaction between curriculum and teachers' knowledge and beliefs.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, Philadelphia, PA. The study is ineligible for review because it does not include a student outcome.

Additional source:

Granger, E. M., Bevis, T. H., Saka, Y., & Southerland, S. (2009, April). *Learning about space science: Comparing the efficacy of reform based teaching with a traditional/verifications approach.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Appendix A: Research details for Granger et al., 2010

Granger, E. M., Bevis, T. H., Saka, Y., & Southerland, S. A. (2010, March). *Large scale, randomized cluster design study of the relative effectiveness of reform-based and traditional/verification curricula in supporting student science learning.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, Philadelphia, PA.

Table A. Summary of findings

Meets WWC evidence standards without reservations

		Study findings	
Outcome domain	Sample size	Average improvement index *(percentile points)*	Statistically significant
General science achievement	2,594 students	+7	Yes

Setting

The study was conducted in a county in central Florida during the 2007–08 and 2008–09 school years.

Study sample

The study used a randomized cluster experimental design. Volunteer teachers were matched on student demographics and grade level and then were randomly assigned to either the intervention group or the comparison group. Over a two-year period, 140 teachers were randomly assigned—70 to the intervention group and 70 to the comparison group.[5] The total analysis sample across both years included 66 teachers in intervention classrooms and 59 teachers in comparison classrooms. The overall and differential attrition rates of teachers (11% and 10%, respectively) met WWC standards for low attrition.[6]

The student analysis sample included 1,418 students who received the *GEMS® Space Science* sequence and 1,176 comparison group students who received the typical space science instruction available in the district. Attrition rates of students were unknown.[7] About 40% of these students were in the fourth grade; the rest were fifth graders. The students were evenly split between boys and girls (50% male, 50% female). Almost one-third (31%) were eligible for free and reduced-price lunch. About 62% of the students were White. Three percent of the sample were English language learners.

The study reported students' outcomes immediately following completion of the space science sequence. These findings can be found in Appendix C. Additional findings reflecting students' follow-up outcomes five months after the completion of the space science sequence can be found in Appendix D.

Intervention group

Students in the intervention group received *GEMS® Space Science* (Lawrence Hall of Science, 2007)[8] for grades 3–5, which was designed to address age-appropriate core concepts in space science and common misconceptions that students might have about them. Students investigated size and scale relative to distance, the Earth's shape and gravity, how the Earth moves, and moon phases and eclipses in four units, over 24 sixty-minute class sessions. The curriculum had an explicit focus on the role of models and evidence in science. Throughout the unit, students evaluated alternative explanations and used evidence to support explanations and to critique the merits of an explanation.

Comparison group

Comparison teachers used the standard district text for grades 4 and 5 to address the same space science content as the intervention group. The district curriculum was centered on a more didactic presentation of space science concepts, including direct instruction, reading of text, and students answering very focused questions.

Outcomes and measurement

Student outcomes were assessed with the Space Science Content test (Sadler, Coyle, Cook-Smith, & Miller, 2007).[9] The assessments were given to students prior to space science instruction, two weeks following completion of teaching the space science unit, and at the five-month follow-up. For a more detailed description of these outcome measures, see Appendix B. The study also used assessments that did not meet inclusion criteria as outcome measures for the Science topic area: the Homerton Science Attitudes survey, the Models and Evidence Questionnaire, and Views of Scientific Inquiry.

Support for implementation

Teachers in the intervention condition were given four days of preservice professional development to learn about the specific curriculum before the school year, a three-hour follow-up training before the curriculum was implemented, and access to a "science coach" midway through teaching the unit that was tested. In addition, all teachers were offered basic professional development related to the new textbook being used by the district in all space science classes. Teachers in both groups were instructed to refrain from adding any activities to those present in their assigned curriculum. The study does not provide information on the education or experience of teachers.

Appendix B: Outcome measures for each domain

General science achievement	
Earth/space science construct	
Space Science Content test	The Space Science Content test assesses students' understanding of key physical science ideas. The assessment was developed by researchers at MOSART: Misconceptions-Oriented Standards-Based Assessment Resources for Teachers. The test is aligned to national content standards and reviewed by science faculty to ensure validity (as cited in Granger et al., 2010).

Appendix C: Findings included in the rating for the general science achievement domain

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Granger et al., 2010[a]								
Space Science Content test	Grades 4–5	2,594 students	nr (2.83)	nr (2.83)	0.49	0.17	+7	0.00
Domain average for general science achievement (Granger et al., 2010)						**0.17**	**+7**	**Statistically significant**

Table Notes: Positive results for mean difference, effect size, and improvement index favor the intervention group; negative results favor 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 statistical significance of the study's domain average was determined by the WWC; a study is characterized as having a statistically significant positive effect when 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.

[a] For Granger et al. (2010), no corrections for clustering or multiple comparisons were needed. The *p*-values presented here were reported in the original study. For Granger et al. (2010), the mean difference is the program coefficient from the hierarchical linear modeling (HLM) analysis. The student characteristics controlled for in the HLM are ethnicity, free and reduced-price lunch, and the pretest measure. The effect size was obtained by dividing the raw score *GEMS®* coefficient by the outcome standard deviation.

Appendix D: Summary of follow-up findings for the general science achievement domain

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Granger et al., 2010[a]								
Space Science Content test	Grades 4–5	2,594 students	nr	nr	0.19	0.07	+3	0.19

Table Notes: The supplemental findings presented in this table are additional findings at the five-month follow-up from Granger et al. (2010) that do not factor into the determination of the intervention rating. Immediate posttest scores were used for rating purposes and are presented in Appendix C. Positive results for mean difference, effect size, and improvement index favor the intervention group; negative results favor 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. nr = not reported.

[a] For Granger et al. (2010), no corrections for clustering or multiple comparisons were needed. The *p*-values presented here were reported in the original study. The mean difference reported in the table is the program coefficient from the HLM analysis. The student characteristics controlled for in the HLM are ethnicity, free and reduced-price lunch, and the pretest measure. The effect size was obtained by dividing the raw score *GEMS®* coefficient by the outcome standard deviation.

Endnotes

[1] The descriptive information for this program was obtained from publicly available sources: the developer's website (http://www.lawrencehallofscience.org/gems, downloaded June 2011) and the program publisher's website (http://www.carolinacurriculum.com/GEMS/About+GEMS.asp, downloaded June 2011). The WWC requests developers review the program description sections for accuracy from their perspective. The program description was provided to the developer in August 2011, and we incorporated feedback from the developer. Further verification of the accuracy of the descriptive information for this program is beyond the scope of this review. The literature search reflects documents publicly available by May 2012.

[2] The *GEMS® Space Science* sequence offers two sequence levels, one for grades 3–5 and one for grades 6–8. This intervention report focuses on the *GEMS® Space Science* sequence for grades 3–5.

[3] The studies in this report were reviewed using WWC Evidence Standards, version 2.1, as described in the Science review protocol version 2.0. The evidence presented in this report is based on available research. Findings and conclusions may change as new research becomes available.

[4] For criteria used in the determination of the rating of effectiveness and extent of evidence, see the WWC Rating Criteria on p. 12.

[5] In the 2007–08 school year, half of 72 fourth- and fifth-grade teachers were randomized into intervention and comparison conditions. The analysis sample included 32 intervention and 29 comparison teachers. In the 2008–09 school year, half of the 68 fourth- and fifth-grade teachers were randomized into intervention and comparison conditions. This analysis sample included 34 intervention and 30 comparison teachers. The authors did not clarify if some of these teachers were the same as in the previous year, or if some of the fourth-grade students in the 2007–08 school year were included in either condition in the 2008–09 school year.

[6] Teachers left the study as a result of personal issues or inability to comply with instructional and/or data collection procedures.

[7] Student attrition is considered "low" under the assumption of an equal number of students assigned to each teacher at baseline. For example, if all teachers (n = 140 at baseline) on average taught 21.48 students (derived from the intervention group analysis sample: 1,418 students/66 teachers), then the resulting subcluster attrition is low.

[8] Lawrence Hall of Science. (2007). *Space Science Curriculum Sequence. Great Explorations in Math and Science.* Berkeley, CA: University of California Press.

[9] Sadler, P., Coyle, H., Cook-Smith, N., & Miller, J. (2007). *Misconceptions-Oriented Standards-based Assessment Resources for Teachers (MOSART)*. Cambridge, MA: Harvard College.

Recommended Citation

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2012, June). *Science intervention report: Great Explorations in Math and Science® (GEMS®) Space Science.* Retrieved from http://whatworks.ed.gov.

WWC Rating Criteria

Criteria used to determine the rating of a study

Study rating	Criteria
Meets WWC evidence standards without reservations	A study that provides strong evidence for an intervention's effectiveness, such as a well-implemented RCT.
Meets WWC evidence standards with reservations	A study that provides weaker evidence for an intervention's effectiveness, such as a QED or an RCT with high attrition that has established equivalence of the analytic samples.

Criteria used to determine the rating of effectiveness for an intervention

Rating of effectiveness	Criteria
Positive effects	Two or more studies show statistically significant positive effects, at least one of which met WWC evidence standards for a strong design, AND No studies show statistically significant or substantively important negative effects.
Potentially positive effects	At least one study shows a statistically significant or substantively important positive effect, AND No studies show a statistically significant or substantively important negative effect AND fewer or the same number of studies show indeterminate effects than show statistically significant or substantively important positive effects.
Mixed effects	At least one study shows a statistically significant or substantively important positive effect AND at least one study shows a statistically significant or substantively important negative effect, but no more such studies than the number showing a statistically significant or substantively important positive effect, OR At least one study shows a statistically significant or substantively important effect AND more studies show an indeterminate effect than show a statistically significant or substantively important effect.
Potentially negative effects	One study shows a statistically significant or substantively important negative effect and no studies show a statistically significant or substantively important positive effect, OR Two or more studies show statistically significant or substantively important negative effects, at least one study shows a statistically significant or substantively important positive effect, and more studies show statistically significant or substantively important negative effects than show statistically significant or substantively important positive effects.
Negative effects	Two or more studies show statistically significant negative effects, at least one of which met WWC evidence standards for a strong design, AND No studies show statistically significant or substantively important positive effects.
No discernible effects	None of the studies shows a statistically significant or substantively important effect, either positive or negative.

Criteria used to determine the extent of evidence for an intervention

Extent of evidence	Criteria
Medium to large	The domain includes more than one study, AND The domain includes more than one school, AND The domain findings are based on a total sample size of at least 350 students, OR, assuming 25 students in a class, a total of at least 14 classrooms across studies.
Small	The domain includes only one study, OR The domain includes only one school, OR The domain findings are based on a total sample size of fewer than 350 students, AND, assuming 25 students in a class, a total of fewer than 14 classrooms across studies.

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 and inclusion in this report 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.

Extent of evidence An indication of how much evidence supports the findings. The criteria for the extent of evidence levels are given in the WWC Rating Criteria on p. 12.

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.

Rating of effectiveness The WWC rates the effects of an intervention in each domain based on the quality of the research design and the magnitude, statistical significance, and consistency in findings. The criteria for the ratings of effectiveness are given in the WWC Rating Criteria on p. 12.

Single-case design 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.