Active learning increases student performance in science, engineering, and mathematics

Scott Freeman[a,1], Sarah L. Eddy[a], Miles McDonough[a], Michelle K. Smith[b], Nnadozie Okoroafor[a], Hannah Jordt[a], and Mary Pat Wenderoth[a]

[a]Department of Biology, University of Washington, Seattle, WA 98195; and [b]School of Biology and Ecology, University of Maine, Orono, ME 04469

Edited* by Bruce Alberts, University of California, San Francisco, CA, and approved April 15, 2014 (received for review October 8, 2013)

To test the hypothesis that lecturing maximizes learning and course performance, we metaanalyzed 225 studies that reported data on examination scores or failure rates when comparing student performance in undergraduate science, technology, engineering, and mathematics (STEM) courses under traditional lecturing versus active learning. The effect sizes indicate that on average, student performance on examinations and concept inventories increased by 0.47 SDs under active learning (n = 158 studies), and that the odds ratio for failing was 1.95 under traditional lecturing (n = 67 studies). These results indicate that average examination scores improved by about 6% in active learning sections, and that students in classes with traditional lecturing were 1.5 times more likely to fail than were students in classes with active learning. Heterogeneity analyses indicated that both results hold across the STEM disciplines, that active learning increases scores on concept inventories more than on course examinations, and that active learning appears effective across all class sizes—although the greatest effects are in small ($n \leq 50$) classes. Trim and fill analyses and fail-safe n calculations suggest that the results are not due to publication bias. The results also appear robust to variation in the methodological rigor of the included studies, based on the quality of controls over student quality and instructor identity. This is the largest and most comprehensive metaanalysis of undergraduate STEM education published to date. The results raise questions about the continued use of traditional lecturing as a control in research studies, and support active learning as the preferred, empirically validated teaching practice in regular classrooms.

constructivism | undergraduate education | evidence-based teaching | scientific teaching

Lecturing has been the predominant mode of instruction since universities were founded in Western Europe over 900 y ago (1). Although theories of learning that emphasize the need for students to construct their own understanding have challenged the theoretical underpinnings of the traditional, instructor-focused, "teaching by telling" approach (2, 3), to date there has been no quantitative analysis of how constructivist versus exposition-centered methods impact student performance in undergraduate courses across the science, technology, engineering, and mathematics (STEM) disciplines. In the STEM classroom, should we ask or should we tell?

Addressing this question is essential if scientists are committed to teaching based on evidence rather than tradition (4). The answer could also be part of a solution to the "pipeline problem" that some countries are experiencing in STEM education: For example, the observation that less than 40% of US students who enter university with an interest in STEM, and just 20% of STEM-interested underrepresented minority students, finish with a STEM degree (5).

To test the efficacy of constructivist versus exposition-centered course designs, we focused on the design of class sessions—as opposed to laboratories, homework assignments, or other exercises. More specifically, we compared the results of experiments that documented student performance in courses with at least some active learning versus traditional lecturing, by metaanalyzing 225 studies in the published and unpublished literature. The active learning interventions varied widely in intensity and implementation, and included approaches as diverse as occasional group problem-solving, worksheets or tutorials completed during class, use of personal response systems with or without peer instruction, and studio or workshop course designs. We followed guidelines for best practice in quantitative reviews (*SI Materials and Methods*), and evaluated student performance using two outcome variables: (*i*) scores on identical or formally equivalent examinations, concept inventories, or other assessments; or (*ii*) failure rates, usually measured as the percentage of students receiving a D or F grade or withdrawing from the course in question (DFW rate).

The analysis, then, focused on two related questions. Does active learning boost examination scores? Does it lower failure rates?

Results

The overall mean effect size for performance on identical or equivalent examinations, concept inventories, and other assessments was a weighted standardized mean difference of 0.47 (Z = 9.781, $P << 0.001$)—meaning that on average, student performance increased by just under half a SD with active learning compared with lecturing. The overall mean effect size for failure rate was an odds ratio of 1.95 (Z = 10.4, $P << 0.001$). This odds ratio is equivalent to a risk ratio of 1.5, meaning that on average, students in traditional lecture courses are 1.5 times more likely to fail than students in courses with active learning. Average failure rates were 21.8% under active learning but 33.8% under traditional lecturing—a difference that represents a 55% increase (Fig. 1 and Fig. S1).

Significance

The President's Council of Advisors on Science and Technology has called for a 33% increase in the number of science, technology, engineering, and mathematics (STEM) bachelor's degrees completed per year and recommended adoption of empirically validated teaching practices as critical to achieving that goal. The studies analyzed here document that active learning leads to increases in examination performance that would raise average grades by a half a letter, and that failure rates under traditional lecturing increase by 55% over the rates observed under active learning. The analysis supports theory claiming that calls to increase the number of students receiving STEM degrees could be answered, at least in part, by abandoning traditional lecturing in favor of active learning.

Author contributions: S.F. and M.P.W. designed research; S.F., M.M., M.K.S., N.O., H.J., and M.P.W. performed research; S.F. and S.L.E. analyzed data; and S.F., S.L.E., M.M., M.K.S., N.O., H.J., and M.P.W. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

See Commentary on page 8319.

[1]To whom correspondence should be addressed. E-mail: srf991@u.washington.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1319030111/-/DCSupplemental.

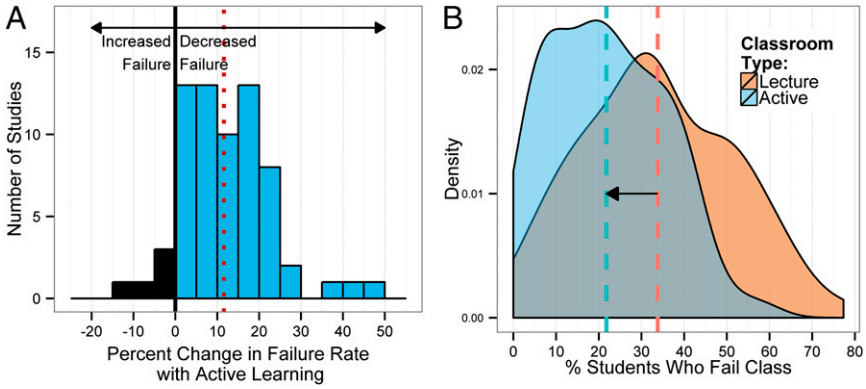


Fig. 1. Changes in failure rate. (*A*) Data plotted as percent change in failure rate in the same course, under active learning versus lecturing. The mean change (12%) is indicated by the dashed vertical line. (*B*) Kernel density plots of failure rates under active learning and under lecturing. The mean failure rates under each classroom type (21.8% and 33.8%) are shown by dashed vertical lines.

Heterogeneity analyses indicated no statistically significant variation among experiments based on the STEM discipline of the course in question, with respect to either examination scores (Fig. 2*A*; $Q = 910.537$, df = 7, $P = 0.160$) or failure rates (Fig. 2*B*; $Q = 11.73$, df = 6, $P = 0.068$). In every discipline with more than 10 experiments that met the admission criteria for the meta-analysis, average effect sizes were statistically significant for either examination scores or failure rates or both (Fig. 2, Figs. S2 and S3, and Tables S1*A* and S2*A*). Thus, the data indicate that active learning increases student performance across the STEM disciplines.

For the data on examinations and other assessments, a heterogeneity analysis indicated that average effect sizes were lower when the outcome variable was an instructor-written course examination as opposed to performance on a concept inventory (Fig. 3*A* and Table S1*B*; $Q = 10.731$, df = 1, $P << 0.001$). Although student achievement was higher under active learning for both types of assessments, we hypothesize that the difference in gains for examinations versus concept inventories may be due to the two types of assessments testing qualitatively different cognitive skills. This explanation is consistent with previous research indicating that active learning has a greater impact on student mastery of higher- versus lower-level cognitive skills (6–9), and the recognition that most concept inventories are designed to diagnose known misconceptions, in contrast to course examinations that emphasize content mastery or the ability to solve quantitative problems (10). Most concept inventories also undergo testing for validity, reliability, and readability.

Heterogeneity analyses indicated significant variation in terms of course size, with active learning having the highest impact on courses with 50 or fewer students (Fig. 3*B* and Table S1*C*; $Q = 6.726$, df = 2, $P = 0.035$; Fig. S4). Effect sizes were statistically significant for all three categories of class size, however, indicating that active learning benefitted students in medium (51–110 students) or large (>110 students) class sizes as well.

When we metaanalyzed the data by course type and course level, we found no statistically significant difference in active learning's effect size when comparing (*i*) courses for majors versus nonmajors ($Q = 0.045$, df = 1, $P = 0.883$; Table S1*D*), or (*ii*) introductory versus upper-division courses ($Q = 0.046$, df = 1, $P = 0.829$; Tables S1*E* and S2*D*).

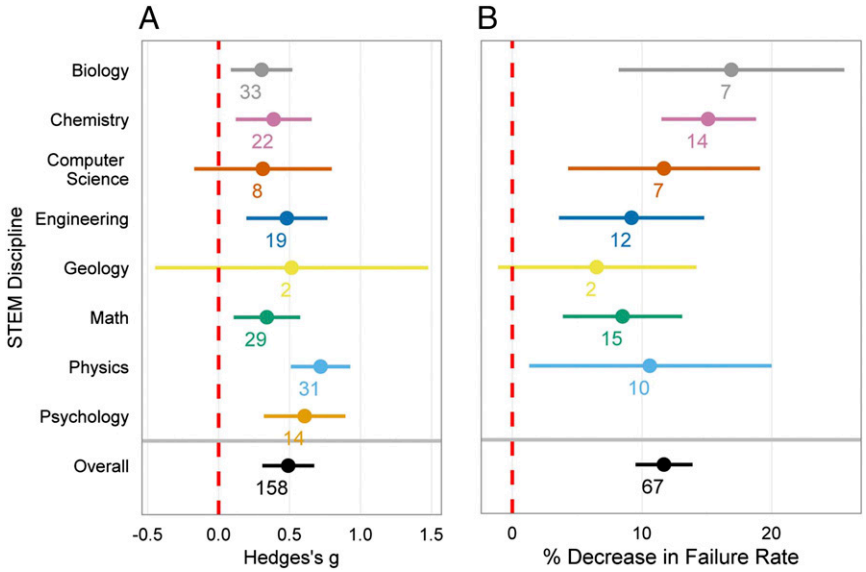


Fig. 2. Effect sizes by discipline. (*A*) Data on examination scores, concept inventories, or other assessments. (*B*) Data on failure rates. Numbers below data points indicate the number of independent studies; horizontal lines are 95% confidence intervals.

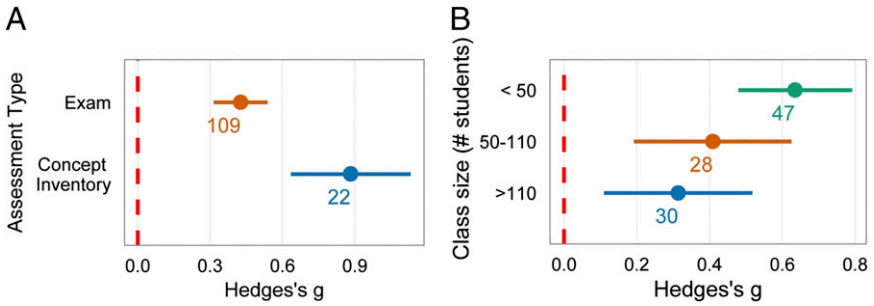


Fig. 3. Heterogeneity analyses for data on examination scores, concept inventories, or other assessments. (*A*) By assessment type—concept inventories versus examinations. (*B*) By class size. Numbers below data points indicate the number of independent studies; horizontal lines are 95% confidence intervals.

To evaluate how confident practitioners can be about these conclusions, we performed two types of analyses to assess whether the results were compromised by publication bias, i.e., the tendency for studies with low effect sizes to remain unpublished. We calculated fail-safe numbers indicating how many missing studies with an effect size of 0 would have to be published to reduce the overall effect sizes of 0.47 for examination performance and 1.95 for failure rate to preset levels that would be considered small or moderate—in this case, 0.20 and 1.1, respectively. The fail-safe numbers were high: 114 studies on examination performance and 438 studies on failure rate (*SI Materials and Methods*). Analyses of funnel plots (Fig. S5) also support a lack of publication bias (*SI Materials and Methods*).

To assess criticisms that the literature on undergraduate STEM education is difficult to interpret because of methodological shortcomings (e.g., ref. 11), we looked for heterogeneity in effect sizes for the examination score data, based on whether experiments did or did not meet our most stringent criteria for student and instructor equivalence. We created four categories to characterize the quality of the controls over student equivalence in the active learning versus lecture treatments (*SI Materials and Methods*), and found that there was no heterogeneity based on methodological quality ($Q = 2.097$, df = 3, $P = 0.553$): Experiments where students were assigned to treatments at random produced results that were indistinguishable from three types of quasirandomized designs (Table 1). Analyzing variation with respect to controls over instructor identity also produced no evidence of heterogeneity ($Q = 0.007$, df = 1, $P = 0.934$): More poorly controlled studies, with different instructors in the two treatment groups or with no data provided on instructor equivalence, gave equivalent results to studies with identical or randomized instructors in the two treatments (Table 1). Thus, the overall effect size for examination data appears robust to variation in the methodological rigor of published studies.

Discussion

The data reported here indicate that active learning increases examination performance by just under half a SD and that lecturing increases failure rates by 55%. The heterogeneity analyses indicate that (*i*) these increases in achievement hold across all of the STEM disciplines and occur in all class sizes, course types, and course levels; and (*ii*) active learning is particularly beneficial in small classes and at increasing performance on concept inventories.

Although this is the largest and most comprehensive meta-analysis of the undergraduate STEM education literature to date, the weighted, grand mean effect size of 0.47 reported here is almost identical to the weighted, grand-mean effect sizes of 0.50 and 0.51 published in earlier metaanalyses of how alternatives to traditional lecturing impact undergraduate course performance in subsets of STEM disciplines (11, 12). Thus, our results are consistent with previous work by other investigators.

The grand mean effect sizes reported here are subject to important qualifications, however. For example, because struggling students are more likely to drop courses than high-achieving students, the reductions in withdrawal rates under active learning that are documented here should depress average scores on assessments—meaning that the effect size of 0.47 for examination and concept inventory scores may underestimate active learning's actual impact in the studies performed to date (*SI Materials and Methods*). In contrast, it is not clear whether effect sizes of this magnitude would be observed if active learning approaches were to become universal. The instructors who implemented active learning in these studies did so as volunteers. It is an open question whether student performance would increase as much if all faculty were required to implement active learning approaches.

Assuming that other instructors implement active learning and achieve the average effect size documented here, what would

Table 1. Comparing effect sizes estimated from well-controlled versus less-well-controlled studies

				95% confidence interval	
Type of control	*n*	Hedges's *g*	SE	Lower limit	Upper limit
For student equivalence					
Quasirandom—no data on student equivalence	39	0.467	0.102	0.268	0.666
Quasirandom—no statistical difference in prescores on assessment used for effect size	51	0.534	0.089	0.359	0.709
Quasirandom—no statistical difference on metrics of academic ability/preparedness	51	0.362	0.092	0.181	0.542
Randomized assignment or crossover design	16	0.514	0.098	0.322	0.706
For instructor equivalence					
No data, or different instructors	59	0.472	0.081	0.313	0.631
Identical instructor, randomized assignment, or ≥3 instructors in each treatment	99	0.492	0.071	0.347	0.580

a shift of 0.47 SDs in examination and concept inventory scores mean to their students?

i) Students performing in the 50th percentile of a class based on traditional lecturing would, under active learning, move to the 68th percentile of that class (13)—meaning that instead of scoring better than 50% of the students in the class, the same individual taught with active learning would score better than 68% of the students being lectured to.

ii) According to an analysis of examination scores in three introductory STEM courses (*SI Materials and Methods*), a change of 0.47 SDs would produce an increase of about 6% in average examination scores and would translate to a 0.3 point increase in average final grade. On a letter-based system, medians in the courses analyzed would rise from a B− to a B or from a B to a B+.

The result for undergraduate STEM courses can also be compared with the impact of educational interventions at the precollege level. A recent review of educational interventions in the K–12 literature reports a mean effect size of 0.39 when impacts are measured with researcher-developed tests, analogous to the examination scores analyzed here, and a mean effect size of 0.24 for narrow-scope standardized tests, analogous to the concept inventories analyzed here (14). Thus, the effect size of active learning at the undergraduate level appears greater than the effect sizes of educational innovations in the K–12 setting, where effect sizes of 0.20 or even smaller may be considered of policy interest (14).

There are also at least two ways to view an odds ratio of 1.95 for the risk of failing a STEM course:

i) If the experiments analyzed here had been conducted as randomized controlled trials of medical interventions, they may have been stopped for benefit—meaning that enrolling patients in the control condition might be discontinued because the treatment being tested was clearly more beneficial. For example, a recent analysis of 143 randomized controlled medical trials that were stopped for benefit found that they had a median relative risk of 0.52, with a range of 0.22 to 0.66 (15). In addition, best-practice directives suggest that data management committees may allow such studies to stop for benefit if interim analyses have large sample sizes and P values under 0.001 (16). Both criteria were met for failure rates in the education studies we analyzed: The average relative risk was 0.64 and the P value on the overall odds ratio was $<< 0.001$. Any analogy with biomedical trials is qualified, however, by the lack of randomized designs in studies that included data on failure rates.

ii) There were 29,300 students in the 67 lecturing treatments with data on failure rates. Given that the raw failure rate in this sample averaged 33.8% under traditional lecturing and 21.8% under active learning, the data suggest that 3,516 fewer students would have failed these STEM courses under active learning. Based on conservative assumptions (*SI Materials and Methods*), this translates into over US$3,500,000 in saved tuition dollars for the study population, had all students been exposed to active learning. If active learning were implemented widely, the total tuition dollars saved would be orders of magnitude larger, given that there were 21 million students enrolled in US colleges and universities alone in 2010, and that about a third of these students intended to major in STEM fields as entering freshmen (17, 18).

Finally, increased grades and fewer failures should make a significant impact on the pipeline problem. For example, the 2012 President's Council of Advisors on Science and Technology report calls for an additional one million STEM majors in the United States in the next decade—requiring a 33% increase from the current annual total—and notes that simply increasing the current STEM retention rate of 40% to 50% would meet three-quarters of that goal (5). According to a recent cohort study from the National Center for Education Statistics (19), there are gaps of 0.5 and 0.4 in the STEM-course grade point averages (GPAs) of first-year bachelor's and associate's degree students, respectively, who end up leaving versus persisting in STEM programs. A 0.3 "bump" in average grades with active learning would get the "leavers" close to the current performance level of "persisters." Other analyses of students who leave STEM majors indicate that increased passing rates, higher grades, and increased engagement in courses all play a positive role in retention (20–22).

In addition to providing evidence that active learning can improve undergraduate STEM education, the results reported here have important implications for future research. The studies we metaanalyzed represent the first-generation of work on undergraduate STEM education, where researchers contrasted a diverse array of active learning approaches and intensities with traditional lecturing. Given our results, it is reasonable to raise concerns about the continued use of traditional lecturing as a control in future experiments. Instead, it may be more productive to focus on what we call "second-generation research": using advances in educational psychology and cognitive science to inspire changes in course design (23, 24), then testing hypotheses about which type of active learning is most appropriate and efficient for certain topics or student populations (25). Second-generation research could also explore which aspects of instructor behavior are most important for achieving the greatest gains with active learning, and elaborate on recent work indicating that underprepared and underrepresented students may benefit most from active methods. In addition, it will be important to address questions about the intensity of active learning: Is more always better? Although the time devoted to active learning was highly variable in the studies analyzed here, ranging from just 10–15% of class time being devoted to clicker questions to lecture-free "studio" environments, we were not able to evaluate the relationship between the intensity (or type) of active learning and student performance, due to lack of data (*SI Materials and Methods*).

As research continues, we predict that course designs inspired by second-generation studies will result in additional gains in student achievement, especially when the types of active learning interventions analyzed here—which focused solely on in-class innovations—are combined with required exercises that are completed outside of formal class sessions (26).

Finally, the data suggest that STEM instructors may begin to question the continued use of traditional lecturing in everyday practice, especially in light of recent work indicating that active learning confers disproportionate benefits for STEM students from disadvantaged backgrounds and for female students in male-dominated fields (27, 28). Although traditional lecturing has dominated undergraduate instruction for most of a millennium and continues to have strong advocates (29), current evidence suggests that a constructivist "ask, don't tell" approach may lead to strong increases in student performance—amplifying recent calls from policy makers and researchers to support faculty who are transforming their undergraduate STEM courses (5, 30).

Materials and Methods

To create a working definition of active learning, we collected written definitions from 338 audience members, before biology departmental seminars on active learning, at universities throughout the United States and Canada. We then coded elements in the responses to create the following consensus definition:

> Active learning engages students in the process of learning through activities and/or discussion in class, as opposed to passively listening

to an expert. It emphasizes higher-order thinking and often involves group work. (See also ref. 31, p. iii).

Following Bligh (32), we defined traditional lecturing as "...continuous exposition by the teacher." Under this definition, student activity was assumed to be limited to taking notes and/or asking occasional and unprompted questions of the instructor.

Literature Search. We searched the gray literature, primarily in the form of unpublished dissertations and conference proceedings, in addition to peer-reviewed sources (33, 34) for studies that compared student performance in undergraduate STEM courses under traditional lecturing versus active learning. We used four approaches (35) to find papers for consideration: hand-searching every issue in 55 STEM education journals from June 1, 1998 to January 1, 2010 (Table S3), searching seven online databases using an array of terms, mining reviews and bibliographies (*SI Materials and Methods*), and "snowballing" from references in papers admitted to the study (*SI Materials and Methods*). We had no starting time limit for admission to the study; the ending cutoff for consideration was completion or publication before January 1, 2010.

Criteria for Admission. As recommended (36), the criteria for admission to the coding and final data analysis phases of the study were established at the onset of the work and were not altered. We coded studies that (*i*) contrasted traditional lecturing with any active learning intervention, with total class time devoted to each approach not differing by more than 30 min/wk; (*ii*) occurred in the context of a regularly scheduled course for undergraduates; (*iii*) were largely or solely limited to changes in the conduct of the regularly scheduled class or recitation sessions; (*iv*) involved a course in astronomy, biology, chemistry, computer science, engineering, geology, mathematics, natural resources or environmental science, nutrition or food science, physics, psychology, or statistics; and (*v*) included data on some aspect of student academic performance.

Note that criterion *i* yielded papers representing a wide array of active learning activities, including vaguely defined "cooperative group activities in class," in-class worksheets, clickers, problem-based learning (PBL), and studio classrooms, with intensities ranging from 10% to 100% of class time (*SI Materials and Methods*). Thus, this study's intent was to evaluate the average effect of any active learning type and intensity contrasted with traditional lecturing.

The literature search yielded 642 papers that appeared to meet these five criteria and were subsequently coded by at least one of the authors.

Coding. All 642 papers were coded by one of the authors (S.F.) and 398 were coded independently by at least one other member of the author team (M.M., M.S., M.P.W., N.O., or H.J.). The 244 "easy rejects" were excluded from the study after the initial coder (S.F.) determined that they clearly did not meet one or more of the five criteria for admission; a post hoc analysis suggested that the easy rejects were justified (*SI Materials and Methods*).

The two coders met to review each of the remaining 398 papers and reach consensus (37, 38) on

i) The five criteria listed above for admission to the study;
ii) Examination equivalence—meaning that the assessment given to students in the lecturing and active learning treatment groups had to be identical, equivalent as judged by at least one third-party observer recruited by the authors of the study in question but blind to the hypothesis being tested, or comprising questions drawn at random from a common test bank;
iii) Student equivalence—specifically whether the experiment was based on randomization or quasirandomization among treatments and, if quasirandom, whether students in the lecture and active learning treatments were statistically indistinguishable in terms of (*a*) prior general academic performance (usually measured by college GPA at the time of entering the course, Scholastic Aptitude Test, or American College Testing scores), or (*b*) pretests directly relevant to the topic in question;
iv) Instructor equivalence—meaning whether the instructors in the lecture and active learning treatments were identical, randomly assigned, or consisted of a group of three or more in each treatment; and
v) Data that could be used for computing an effect size.

To reduce or eliminate pseudoreplication, the coders also annotated the effect size data using preestablished criteria to identify and report effect sizes only from studies that represented independent courses and populations reported. If the data reported were from iterations of the same course at the same institution, we combined data recorded for more than one control and/or treatment group from the same experiment. We also combined data from multiple outcomes from the same study (e.g., a series of equivalent midterm examinations) (*SI Materials and Methods*). Coders also extracted data on class size, course type, course level, and type of active learning, when available.

Criteria *iii* and *iv* were meant to assess methodological quality in the final datasets, which comprised 158 independent comparisons with data on student examination performance and 67 independent comparisons with data on failure rates. The data analyzed and references to the corresponding papers are archived in Table S4.

Data Analysis. Before analyzing the data, we inspected the distribution of class sizes in the study and binned this variable as small, medium, and large (*SI Materials and Methods*). We also used established protocols (38, 39) to combine data from multiple treatments/controls and/or data from multiple outcomes, and thus produce a single pairwise comparison from each independent course and student population in the study (*SI Materials and Methods*).

The data we analyzed came from two types of studies: (*i*) randomized trials, where each student was randomly placed in a treatment; and (*ii*) quasirandom designs where students self-sorted into classes, blind to the treatment at the time of registering for the class. It is important to note that in the quasirandom experiments, students were assigned to treatment as a group, meaning that they are not statistically independent samples. This leads to statistical problems: The number of independent data points in each treatment is not equal to the number of students (40). The element of nonindependence in quasirandom designs can cause variance calculations to underestimate the actual variance, leading to overestimates for significance levels and for the weight that each study is assigned (41). To correct for this element of nonindependence in quasirandom studies, we used a cluster adjustment calculator in Microsoft Excel based on methods developed by Hedges (40) and implemented in several recent metaanalyses (42, 43). Adjusting for clustering in our data required an estimate of the intraclass correlation coefficient (ICC). None of our studies reported ICCs, however, and to our knowledge, no studies have reported an ICC in college-level STEM courses. Thus, to obtain an estimate for the ICC, we turned to the K–12 literature. A recent paper reviewed ICCs for academic achievement in mathematics and reading for a national sample of K–12 students (44). We used the mean ICC reported for mathematics (0.22) as a conservative estimate of the ICC in college-level STEM classrooms. Note that although the cluster correction has a large influence on the variance for each study, it does not influence the effect size point estimate substantially.

We computed effect sizes and conducted the metaanalysis in the Comprehensive Meta-Analysis software package (45). All reported *P* values are two-tailed, unless noted.

We used a random effects model (46, 47) to compare effect sizes. The random effect size model was appropriate because conditions that could affect learning gains varied among studies in the analysis, including the (*i*) type (e.g., PBL versus clickers), intensity (percentage of class time devoted to constructivist activities), and implementation (e.g., graded or ungraded) of active learning; (*ii*) student population; (*iii*) course level and discipline; and (*iv*) type, cognitive level, and timing—relative to the active learning exercise—of examinations or other assessments.

We calculated effect sizes as (*i*) the weighted standardized mean difference as Hedges' *g* (48) for data on examination scores, and (*ii*) the log-odds for data on failure rates. For ease of interpretation, we then converted log-odds values to odds ratio, risk ratio, or relative risk (49).

To evaluate the influence of publication bias on the results, we assessed funnel plots visually (50) and statistically (51), applied Duval and Tweedie's trim and fill method (51), and calculated fail-safe *N*s (45).

Additional Results. We did not insist that assessments be identical or formally equivalent if studies reported only data on failure rates. To evaluate the hypothesis that differences in failure rates recorded under traditional lecturing and active learning were due to changes in the difficulty of examinations and other course assessments, we evaluated 11 studies where failure rate data were based on comparisons in which most or all examination questions were identical. The average odds ratio for these 11 studies was 1.97 ± 0.36 (SE)—almost exactly the effect size calculated from the entire dataset.

Although we did not metaanalyze the data using "vote-counting" approaches, it is informative to note that of the studies reporting statistical tests of examination score data, 94 reported significant gains under active learning whereas only 41 did not (Table S4*A*).

Additional results from the analyses on publication bias are reported in Supporting Information.

ACKNOWLEDGMENTS. We thank Roddy Theobald for advice on interpreting odds ratios; the many authors who provided missing data upon request (*SI Materials and Methods*); Colleen Craig, Daryl Pedigo, and Deborah Wiegand for supplying information on examination score standard deviations and grading thresholds; Kelly Puzio and an anonymous reviewer for advice on analyzing data from quasirandom studies; and Steven Kroiss, Carl Wieman, and William Wood for comments that improved the manuscript. M.S. was supported in part by National Science Foundation Grant 0962805.

1. Brockliss L (1996) Curricula. *A History of the University in Europe*, ed de Ridder-Symoens H (Cambridge Univ Press, Cambridge, UK), Vol II, pp 565–620.
2. Piaget J (1926) *The Language and Thought of the Child* (Harcourt Brace, New York).
3. Vygotsky LS (1978) *Mind in Society* (Harvard Univ Press, Cambridge, MA).
4. Handelsman J, et al. (2004) Education. Scientific teaching. *Science* 304(5670):521–522.
5. PCAST STEM Undergraduate Working Group (2012) *Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics*, eds Gates SJ, Jr, Handelsman J, Lepage GP, Mirkin C (Office of the President, Washington).
6. Haukoos GD, Penick JE (1983) The influence of classroom climate on science process and content achievement of community college students. *J Res Sci Teach* 20(7): 629–637.
7. Martin T, Rivale SD, Diller KR (2007) Comparison of student learning in challenge-based and traditional instruction in biomedical engineering. *Ann Biomed Eng* 35(8): 1312–1323.
8. Cordray DS, Harris TR, Klein S (2009) A research synthesis of the effectiveness, replicability, and generality of the VaNTH challenge-based instructional modules in bioengineering. *J. Eng Ed* 98(4).
9. Jensen JL, Lawson A (2011) Effects of collaborative group composition and inquiry instruction on reasoning gains and achievement in undergraduate biology. *CBE Life Sci Educ* 10(1):64–73.
10. Momsen JL, Long TM, Wyse SA, Ebert-May D (2010) Just the facts? Introductory undergraduate biology courses focus on low-level cognitive skills. *CBE Life Sci Educ* 9(4): 435–440.
11. Ruiz-Primo MA, Briggs D, Iverson H, Talbot R, Shepard LA (2011) Impact of undergraduate science course innovations on learning. *Science* 331(6022):1269–1270.
12. Springer L, Stanne ME, Donovan SS (1999) Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology. *Rev Educ Res* 69(1):21–51.
13. Bowen CW (2000) A quantitative literature review of cooperative learning effects on high school and college chemistry achievement. *J Chem Educ* 77(1):116–119.
14. Lipsey MW, et al. (2012) *Translating the Statistical Representation of the Effects of Educational Interventions into Readily Interpretable Forms* (US Department of Education, Washington).
15. Montori VM, et al. (2005) Randomized trials stopped early for benefit: A systematic review. *JAMA* 294(17):2203–2209.
16. Pocock SJ (2006) Current controversies in data monitoring for clinical trials. *Clin Trials* 3(6):513–521.
17. National Center for Education Statistics (2012) *Digest of Education Statistics* (US Department of Education, Washington).
18. National Science Board (2010) *Science and Engineering Indicators 2010* (National Science Foundation, Arlington, VA).
19. National Center for Education Statistics (2012) *STEM in Postsecondary Education* (US Department of Education, Washington).
20. Seymour E, Hewitt NM (1997) *Talking About Leaving: Why Undergraduates Leave the Sciences* (Westview Press, Boulder, CO).
21. Goodman IF, et al. (2002) *Final Report of the Women's Experiences in College Engineering (WECE) Project* (Goodman Research Group, Cambridge, MA).
22. Watkins J, Mazur E (2013) Retaining students in science, technology, engineering, and mathematics (STEM) majors. *J Coll Sci Teach* 42(5):36–41.
23. Slavich GM, Zimbardo PG (2012) Transformational teaching: Theoretical underpinnings, basic principles, and core methods. *Educ Psychol Rev* 24(4):569–608.
24. Dunlosky J, Rawson KA, Marsh EJ, Nathan MJ, Willingham DT (2013) Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psych Sci Publ Int* 14(1):4–58.
25. Eddy S, Crowe AJ, Wenderoth MP, Freeman S (2013) How should we teach tree-thinking? An experimental test of two hypotheses. *Evol Ed Outreach* 6:1–11.
26. Freeman S, Haak D, Wenderoth MP (2011) Increased course structure improves performance in introductory biology. *CBE Life Sci Educ* 10(2):175–186.
27. Lorenzo M, Crouch CH, Mazur E (2006) Reducing the gender gap in the physics classroom. *Am J Phys* 74(2):118–122.
28. Haak DC, HilleRisLambers J, Pitre E, Freeman S (2011) Increased structure and active learning reduce the achievement gap in introductory biology. *Science* 332(6034): 1213–1216.
29. Burgan M (2006) In defense of lecturing. *Change* 6:31–34.
30. Henderson C, Beach A, Finkelstein N (2011) Facilitating change in undergraduate STEM instructional practices: An analytic review of the literature. *J Res Sci Teach* 48(8): 952–984.
31. Bonwell CC, Eison JA (1991) *Active Learning: Creating Excitement in the Classroom* (George Washington Univ, Washington, DC).
32. Bligh DA (2000) *What's the Use of Lectures?* (Jossey-Bass, San Francisco).
33. Reed JG, Baxter PM (2009) Using reference databases. *The Handbook of Research Synthesis and Meta-Analysis*, eds Cooper H, Hedges LV, Valentine JC (Russell Sage Foundation, New York), pp 73–101.
34. Rothstein H, Hopewell S (2009) Grey literature. *The Handbook of Research Synthesis and Meta-Analysis*, eds Cooper H, Hedges LV, Valentine JC (Russell Sage Foundation, New York), pp 103–125.
35. White HD (2009) Scientific communication and literature retrieval. *The Handbook of Research Synthesis and Meta-Analysis*, eds Cooper H, Hedges LV, Valentine JC (Russell Sage Foundation, New York), pp 51–71.
36. Lipsey MW, Wilson DB (2001) *Practical Meta-Analysis* (Sage Publications, Thousand Oaks, CA).
37. Orwin RG, Vevea JL (2009) Evaluating coding decisions. *The Handbook of Research Synthesis and Meta-Analysis*, eds Cooper H, Hedges LV, Valentine JC (Russell Sage Foundation, New York), pp 177–203.
38. Higgins JPT, Green S, eds (2011) Cochrane Handbook for Systematic Reviews of Interventions, Version 5.1.0 (The Cochrane Collaboration, Oxford). Available at www.cochrane-handbook.org. Accessed December 14, 2012.
39. Borenstein M (2009) Effect sizes for continuous data. *The Handbook of Systematic Review and Meta-Analysis*, eds Cooper H, Hedges LV, Valentine JC (Russell Sage Foundation, New York), pp 221–235.
40. Hedges LV (2007) Correcting a significance test for clustering. *J Educ Behav Stat* 32(2): 151–179.
41. Donner A, Klar N (2002) Issues in the meta-analysis of cluster randomized trials. *Stat Med* 21(19):2971–2980.
42. Davis D (2012) Multiple Comprehension Strategies Instruction (MCSI) for Improving Reading Comprehension and Strategy Outcomes in the Middle Grades. (The Campbell Collaboration, Oxford). Available at http://campbellcollaboration.org/lib/project/167/. Accessed December 10, 2013.
43. Puzio K, Colby GT (2013) Cooperative learning and literacy: A meta-analytic review. *J Res Ed Effect* 6(4):339–360.
44. Hedges LV, Hedberg EC (2007) Intraclass correlation values for planning group-randomized trials in education. *Educ Eval Policy Anal* 29:60–87.
45. Borenstein M, et al. (2006) *Comprehensive Meta-Analysis* (Biostat, Inc., Englewood, NJ).
46. Hedges LV (2009) Statistical considerations. *The Handbook of Research Synthesis and Meta-Analysis*, eds Cooper H, Hedges LV, Valentine JC (Russell Sage Foundation, New York), pp 38–47.
47. Raudenbush SW (2009) Analyzing effect sizes: Random-effects models. *The Handbook of Research Synthesis and Meta-Analysis*, eds Cooper H, Hedges LV, Valentine JC (Russell Sage Foundation, New York), pp 295–315.
48. Gurevitch J, Hedges LV (1999) Statistical issues in ecological meta-analyses. *Ecology* 80(4):1142–1149.
49. Fleiss J, Berlin JA (2009) Effect sizes for dichotomous data. *The Handbook of Research Synthesis and Meta-Analysis*, eds Cooper H, Hedges LV, Valentine JC (Russell Sage Foundation, New York), pp 237–253.
50. Greenhouse JB, Iyengar S (2009) Sensitivity analysis and diagnostics. *The Handbook of Research Synthesis and Meta-Analysis*, eds Cooper H, Hedges LV, Valentine JC (Russell Sage Foundation, New York), pp 417–433.
51. Sutton AJ (2009) Publication bias. *The Handbook of Research Synthesis and Meta-Analysis*, eds Cooper H, Hedges LV, Valentine JC (Russell Sage Foundation, New York), pp 435–452.