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* ≤ 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.

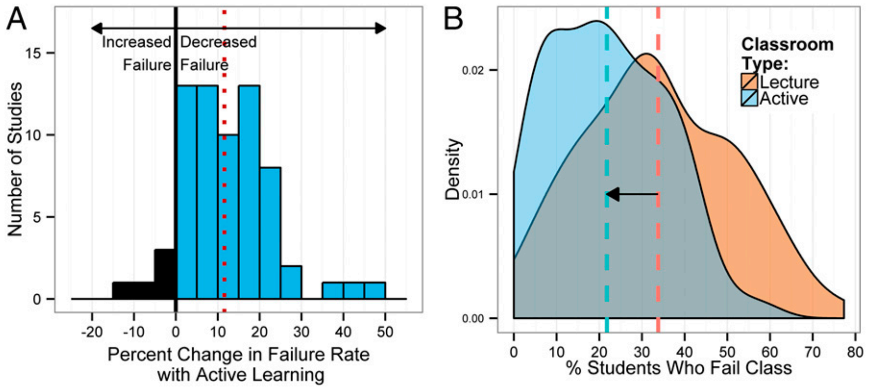


Fig. 1. Changes in failure rate. (*A*) Data plotted as percent change in failure rate in the same course, under active learning versus lecturing. The mean change (12%) is indicated by the dashed vertical line. (*B*) Kernel density plots of failure rates under active learning and under lecturing. The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

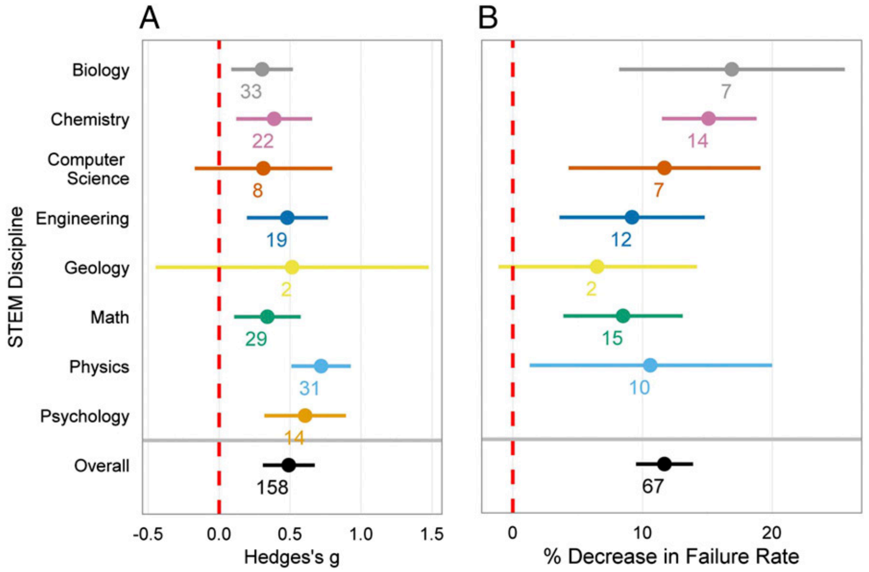


Fig. 2. Effect sizes by discipline. (*A*) Data on examination scores, concept inventories, or other assessments. (*B*) Data on failure rates. Numbers below data points indicate the number of independent studies; horizontal lines are 95% confidence intervals.

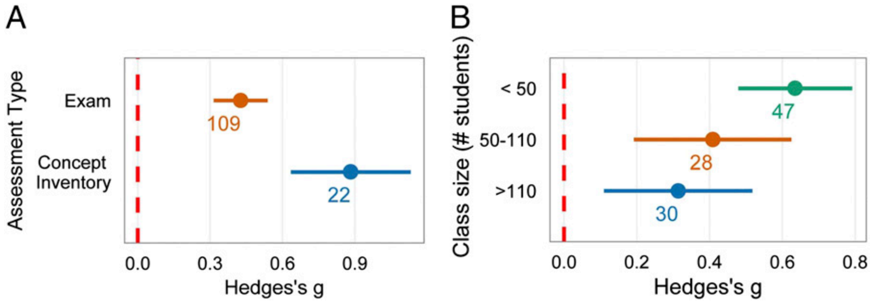


Fig. 3. Heterogeneity analyses for data on examination scores, concept inventories, or other assessments. (*A*) By assessment type—concept inventories versus examinations. (*B*) By class size. Numbers below data points indicate the number of independent studies; horizontal lines are 95% confidence intervals.

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 metaanalysis 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.

Supporting Information

Freeman et al. 10.1073/pnas.1319030111

The question we sought to answer was, "Does active learning work better, in terms of improving student performance in undergraduate science, technology, engineering, and mathematics (STEM) courses, than traditional lecturing?" We evaluated performance using two metrics: (*i*) scores on identical or formally equivalent examinations, concept inventories, or other assessments, and (*ii*) failure rates—in most cases measured as the percentage of Ds, Fs, and/or withdrawals. These were relevant criteria for failure because students with a D, F, or W in a STEM course are usually barred from receiving credit in the major.

SI Materials and Methods

Literature Search. Searching the unpublished or gray literature in addition to peer-reviewed papers is recommended to mitigate the file-drawer effect, i.e., a systematic bias against studies that may have been conducted carefully but were not published because the results failed to show statistical significance and/or had low effect sizes (1, 2).

We had no starting time limit for admission to the study; the ending cutoff for consideration was completion or publication before January 1, 2010. We used four approaches (3) to find papers for consideration:

i) Hand searching: We read the titles of every research paper published between June 1, 1998 and January 1, 2010 in the 55 journals listed in Table S4. If the titles indicated that one of the five criteria for admission might be met, we also read the abstract or summary. Papers that appeared to fulfill most or all of the criteria were downloaded for later coding.

ii) Database searches: We searched Web of Science, Expanded Academic Index, PubMed, Education Resources Information Center, Compendex, ProQuest Science, and ProQuest Dissertations and Theses. Each database was queried using the following terms: active learning, audience response system, case-based learning, clickers, collaborative learning, constructivism or constructivist learning, cooperative learning, peer instruction, peer teaching, personal response device, problem-based learning, reform calculus or calculus reform, studio physics, traditional lecturing, workshop calculus, and workshop physics. To reduce the number of irrelevant hits, these terms were modified by "and student" and/or "and lecture" in certain databases, and/or limited by selecting the modifiers "education," "math," or "college students."

iii) Mining reviews and bibliographies: We collected 33 bibliographies, qualitative reviews, and metaanalyses on undergraduate STEM education and searched their citation lists for papers relevant to this study (4–36).

iv) Snowballing: We searched the citation lists of all of the publications that we coded and checked for papers relevant to this study.

We did not contact individual researchers for unpublished work, and we did not include conference proceedings that included only abstracts.

Criteria for Admission. Five criteria for admission to the study—for coding by at least one researcher—are listed in *Materials and Methods, Criteria for Admission*. These criteria were designed to exclude:

- Experiments with volunteers or paid participants that occurred outside of a normal course.
- Studies on the impact of laboratory exercises, homework, and other assignments; take-home quizzes, supplementary instruction sessions; or other activities that were meant to be completed outside of normal class periods.
- Studies of student affect.

Coding. There were 244 studies classified as easy rejects and 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. Of these, 23 were assigned to one of the second coders. In 22 of 23 cases (96%), the second coder independently evaluated them as easy rejects; in the one case of conflict, a discussion between the two coders resulted in rejection from the study. Thus, it does not appear that a significant number of the 244 easy rejects would actually be admitted if they were subjected to a second, independent coding.

The two coders met to discuss each of the remaining 398 papers (37, 38). The goal was to review and reach consensus on the five issues itemized in *Materials and Methods*, *Criteria for Admission*. On the issue of examination equivalence, it is important to note that we rejected studies where authors either provided no data on examination equivalence or stated that examinations were similar or comparable without providing data to back the claim.

Coding: Student equivalence. To evaluate student equivalence, coders checked whether the experimental design was based on randomization or quasirandomization among treatments. In quasirandomized designs, students register for lecturing or active learning sections blind to the treatment type. However, students may differ among sections or between years in terms of their ability and/or preparation (39, 40). If the experiment was quasirandom, coders noted whether the authors included statistical tests on data concerning prior academic performance—usually college grade point average (GPA) at the time of entering the course, Scholastic Aptitude Test (SAT) or American College Testing (ACT) scores, or pretests directly relevant to the topic in question. Pretests were often a concept inventory; if prescores were statistically indistinguishable, authors frequently reported only postscores. In many cases, we used these postscores to compute an effect size. Note that if a statistical test indicated that students in the lecture section performed significantly worse in terms of the preparation/ability metric analyzed, the study was rejected from the metaanalysis.

This coding system allowed us to create four categories for the quality of studies, in terms of controlling for student equivalence in the active learning and lecture treatments. They are, in order from least well controlled to best controlled, as follows:

- Quasirandom studies where authors provided no data on student equivalence in terms of academic performance. Studies were assigned to this category even if authors claimed that students were indistinguishable but failed to provide a relevant statistical test, or if they provided data showing that students were equivalent in terms of sex, ethnicity, prior courses taken, major field of study, or similar non–performance-related criteria.
- Quasirandom studies where data indicated no statistical difference on a pretest that was directly related to the topic covered in the course and/or to the assessment used to calculate the effect size.
- Quasirandom studies where data indicated no statistical difference on a reliable metric of overall academic preparedness/ability, or where the metric was used as a covariate in the data analysis. These data were usually SAT scores, ACT scores, or college GPA. We did not accept high school GPA to qualify studies for this category.

- Randomized studies, where students were assigned at random (e.g., by the registrar's office) to the two classes being compared, and cross-over designs where the same students experienced both active learning and traditional lecturing in the same course.

Note that in the results on student equivalence given in Table 1*A*, the total sample size is 157 instead of 158. This is because one study had combined outcomes that we used to calculate the effect size (see *Data analysis: Multiple controls/treatments and multiple outcomes*), and because the student controls on the two outcomes used were different. As a result, this study could not be assigned to one of the four categories and had to be dropped from this heterogeneity analysis.

Coding: Identifying studies that were independent, as the unit of analysis. During the coding phase, we took several steps to eliminate pseudoreplication, which can artificially inflate sample sizes in metaanalyses. The coding protocol called for computing a single effect size from each published study, except in instances where a single paper reported the results of studies that were conducted in independent courses and populations (41). The exceptions occurred when experiments reported in the same paper were on the same campus but involved different courses with different student populations (e.g., business calculus versus mathematics for preservice teachers), or courses on different campuses (i.e., that the instructors and students involved did not overlap). In cases where a single paper reported results from multiple courses that may have been taken in sequence by at least some of the same students, we recorded effect size data from only the initial course in the sequence. Note that both coders had to agree that the study populations were completely independent for multiple effect sizes to be reported from the same paper.

Coding: Multiple controls/treatments and multiple outcomes. In cases where papers included data from multiple treatments or multiple controls—from the same course on the same campus, with or without the same instructor(s)—we averaged the replicates to compute a single effect size, using the approach described in *Data analysis: Combining multiple controls and/or treatments* and *Data analysis: Combining multiple outcomes*.

In cases where papers reported multiple outcome variables, the protocol called for coders to choose the assessment that was (*i*) most summative—for example, comprehensive final examinations over midterm examinations or the percentage of students receiving a D or F grade or withdrawing from the course in question (DFW rate) instead of percentage of withdrawals, and/or (*ii*) most comparable to other studies—for example, a widely used concept inventory versus an instructor-written examination. In cases where the assessments were determined to be equivalent—e.g., four hour-long examinations, over the course of a semester—the coders recorded all of the relevant data. We also combined outcomes in cases where summative examination scores (e.g., from a comprehensive final) as well as a concept inventory scores were reported. When coding was complete, we combined data from multiple outcomes to produce a single effect size, using the approaches described in *Data analysis: Combining multiple outcomes*.

Coding: Missing data. A total of 91 papers were judged to be admissible by both coders but lacked some or all of the data required to compute an effect size. In most cases, the missing data were SDs and/or sample sizes. We wrote to every author on all 91 papers for whom we could find an e-mail address, requesting the missing data. If we did not receive a response within 4 wk, we dropped the paper from the study. Via follow-up correspondence, we were able to obtain missing data for 19 of the 91 studies.

Coding: Categorizing class sizes. We coded class size as a continuous variable in all papers where it was mentioned. In cases where the authors reported a range of class sizes, we used the midpoint value in the range. To create class size categories, we made a histogram of the class size distribution, after coding was complete. Our interpretation of this distribution, shown in Fig. S4, was that class sizes fell into three natural groupings: classes with 50 or fewer students, classes of more than 50 up to 110, and classes with more than 110. We designated these as small, medium, and large classes, respectively.

Data Analysis. Data on examination or concept inventory performance were continuous and were reported either as means, SDs, and sample sizes for each treatment group, or as the value and df of Student's *t*. We used these data to compute an effect size from each study as the standardized mean difference weighted by the inverse of the pooled variance, with the correction for clustering in quasirandom studies explained in the *Materials and Methods, Data Analysis*, and with Hedges' correction for small sample sizes (42). Thus, effect sizes for examination-points data are in units of SDs. Effect size calculations, heterogeneity analyses, and publication bias analyses were done in the Comprehensive Meta-Analysis software package; the cluster correction was done in a Microsoft Excel program based on methods developed by Hedges (see *Materials and Methods, Data Analysis*).

Data on failure rates are dichotomous and were reported as the percentage of withdrawals/drops, Ds, Fs, Ds and Fs, DFWs, or students progressing to subsequent courses. We used the odds ratio to compute an effect size for failure rates (43).

It is important to note that papers reporting anything other than DFW—for example, only Ds and Fs or only withdrawals—may underreport the actual overall "raw" failure rate, as it is typical for D and F grades as well as withdrawals to prevent students from getting course credit toward a STEM major and progressing to subsequent courses. However, because experiments admitted to the study had to have the same metric for failing a traditional lecture section and an active learning section, the contrast in failure rate should be consistent across metrics. As a result, the model-based estimate of the change in percentage of students failing should accurately capture the differences observed across different metrics of failure.

The independent studies analyzed in the experiment and the categories used in heterogeneity analyses are identified in Table S4, along with the effect size and associated data from each study analyzed here. Forest plots showing the effect sizes from all of the experiments in the study, organized by discipline, are provided in Fig. S2 for data on examination scores, concept inventories, and other assessments, and in Fig. S3 for failure rate data. Note that due to small sample sizes or in keeping with common departmental organizations, we combined astronomy with physics, statistics with mathematics, and natural resources/nutrition with biology.

Data analysis: Combining multiple controls and/or treatments. Some studies reported multiple control and/or experimental treatments for the same course, meaning that data were reported from more than one section within or across terms/years. In these instances we combined groups (e.g., all of the control terms) to create a single pair-wise comparison (38). For achievement data this involved creating a pooled mean, sample size, and SD. For the dichotomous data on failure rates, we simply summed the sample size and the number of students in each category to create a single group.

Data analysis: Combining multiple outcomes. When studies had multiple outcomes from the same set of students, and when the outcomes were equivalent—in terms of how summative they were or how widespread the use of the assessment was—we computed a summary effect size that combined all of the equivalent outcomes (44). In doing so, we assumed that the correlation between different outcomes within a comparison was 1, because the same students were sampled for each outcome. This is a conservative measure, leading to a less precise estimate of the summary effect, as the actual correlation between outcomes is likely lower than 1.

Because we combined multiple outcomes when they were judged to be equivalent in terms of measuring student-learning gains, each independent study was represented by one effect size in the metaanalysis; 28% of the studies analyzed here had multiple outcomes that we combined to estimate a single effect size.

Discussion/Data Interpretation. To estimate how much average course grades would change if examination scores increased by 0.47 SDs in a STEM course, we obtained data on SDs for examinations in the courses in the introductory biology, chemistry, and physics sequences for majors at the University of Washington (UW). We then calculated how many more total examination points students would receive, on average, with an increase of 0.47 SDs, and compared this increase to the number of points needed to raise a final grade by 0.1 on the 4-point, decimal-based system used at this institution. Average or typical SDs on 100-point examinations were 12.5 in biology, 17.5 in chemistry, and 18 in physics. An increase of 0.47 SDs would raise total examination points by about 23 in biology, 28.7 in chemistry, and 51 in physics, and raise final grades an average of 0.30 in biology, 0.28 in chemistry, and 0.34 in physics. In a letter-based grading system, these increases would move course medians or averages from a B− to a B in biology and chemistry, and from a B to a B+ in physics.

Our estimate of US$3,500,000 in lost tuition and course fees was based on an analysis of income from introductory biology courses at UW. Tuition and fees at this institution are about average for its peer comparison group of 10 other large US public universities. At UW, each course in the introductory biology sequence represents one-third of a normal course load in a term. Based on tuition rates per credit hour for the 2012–2013 academic year, published percentages of students receiving full financial aid and students paying in-state versus out-of-state tuition, we estimated that an average student pays $1,000 in tuition and fees per five-credit STEM course. The resulting estimated total of US $5,000,000 in lost tuition and fees due to failure would be much higher if all of the studies analyzed here included DFWs, or if the estimate were based on tuition at private institutions. It would be lower, however, if it were based on courses that lack laboratory sessions and/or are awarded fewer credit hours.

Additional Results. The data on failure rate reported in Fig. 4 can also be visualized, on a study-by-study basis, as shown in Fig. S1.

Assessing publication bias. Publication bias may be the most serious threat to the validity of a metaanalysis. Specifically, it is possible for a metaanalysis to report an inflated value for the overall effect size because the literature search failed to discover studies with low effect sizes (41). Several analyses support the hypothesis that the so-called file-drawer effect is real: Studies with low effect sizes or that failed to show statistical significance are less likely to be published, and thus substantially harder for the metaanalyst to find (2).

We performed two analyses to evaluate the impact of publication bias on this study: (*i*) assessing funnel plots and (*ii*) calculating a fail-safe *N*.

Assessing funnel plots. Plots of effect sizes versus SE (or sample size) are called funnel plots because in the absence of publication bias, they should be symmetrical about the overall mean and flare out at the bottom. The flare is due to increased variation in effect size estimates in studies with a large SE (or small sample size). Publication bias produces a marked asymmetry in a funnel plot—a "bite" out of the lower left of the plot (45). Visual inspection of the funnel plots for assessment and failure rate data indicates some asymmetry in the data for scores, due to five data points with unusually large effect sizes and/or SEs (Fig. S5*A*), but little-to-no asymmetry in the data for failure rate (Fig. S5*B*).

Statistical analyses (46) are consistent with this conclusion—indicating asymmetry in the funnel plot for assessment data but no asymmetry in the funnel plot for failure rate data. Specifically, both nonparametric and parametric tests indicate a statistically significant association between effect size and SE for the examination score data (Kendall's Tau with continuity correction 0.11, one-tailed $P = 0.02$; Egger's regression intercept 0.55, one-tailed $P = 0.00032$) but no association for the failure rate data (Kendall's Tau with continuity correction 0.14, one-tailed $P = 0.10$; Egger's regression intercept –0.43, one-tailed $P = 0.37$).

However, Duval and Tweedie's trim and fill method (46) for adjusting for publication bias indicates that the degree of asymmetry observed in this study has virtually no impact on the estimates of mean overall effect size: Under trim and fill, the random effects estimate for the standardized mean difference is 0.47 (95% confidence interval 0.37–0.56) for the assessment data and an odds ratio of 1.94 (95% confidence interval 1.71–2.20) for the failure rate data. A sensitivity analysis focusing on extreme values for examination scores also failed to discern an effect (see *Other sensitivity analyses*). Thus, there is no indication that publication bias has affected the effect size estimates reported here.

Calculating a fail-safe **N.** The fail-safe *N* is the number of studies with effect size 0 that would have to be added to the study to reduce the observed effect size to a predetermined value that experts would consider inconsequential. Conservatively, we set the inconsequential value to a standardized mean difference of 0.20—still considered a pedagogically significant effect size in the K–12 literature (47)—for the examination score data and an odds ratio of 1.1 for the failure rate data. Orwin's fail-safe *N* is 114 for the assessment data and 438 for the failure rate data.

Like the trim and fill analyses, the large fail-safe *N*s suggest that the effect sizes reported here are not inflated by publication bias. To bring the effect sizes reported here down to values that might be insignificant to students and instructors, there would have to be an unreasonably large number of undetected studies with 0 effect.

Other sensitivity analyses: Assessing the influence of extreme values. The funnel plots of examination and concept inventory data suggest that there are several studies with extreme effect size values. To quantify this impression, we calculated the lower fence and upper fence for effect size values in the study, based on the interquartile range. The interquartile range was 0.59, producing a lower fence of −0.76 and an upper fence of 1.60. The two studies below the lower fence and the 10 studies above the upper fence can be considered outlying values. To evaluate the impact of these outliers on the metaanalysis, we recalculated the overall effect size with the 12 studies removed and found that it was 0.45, with a 95% confidence interval bounded by 0.38 and 0.51, meaning that it was statistically indistinguishable from the value computed with the outliers.

To investigate the outlying values further, we analyzed the characteristics of each study that generated an extreme value. Both of the outliers with low effect sizes were from studies that tested problem-based learning (PBL) as an active learning technique but involved extremely small sample sizes (totals of 16 and 25 students). The outliers with large effect sizes tested PBL, flipped classrooms with in-class activities, studio models, peer instruction, or vaguely defined interactive engagement with large numbers of students participating. One unifying feature of these high-effect-size studies appeared to be the high intensity of the active learning component. The percentage of class time devoted to active learning versus lecturing was reported as 25 (one study), 33 (one study), and 100 (seven studies), with one study missing data. This observation suggests that varying the intensity of active learning may be a productive experimental design for second-generation, discipline-based education research.

To evaluate the failure-rate data's sensitivity to values from individual studies, we ran a one-study-removed analysis—meaning that we recalculated the overall effect size, as the log-odds, with each of the 67 studies removed. In the 67 recalculations, the overall log-odds ranged from 1.49 to 1.60 with a mean of 1.55, suggesting that no one study had a disproportionate impact on the overall estimate.

Heterogeneity analyses: Assessment data. Table S1 *A*–*C* provides details on the heterogeneity analyses summarized in the *Results*–specifically, metaanalyzing the assessment data by discipline, assessment type (concept inventories versus examinations), and class size.

In addition, we evaluated how coders characterized the active learning interventions in the metaanalysis, and found that there was no heterogeneity in effect sizes for examination data based on the type of active learning intervention used ($Q = 11.173$, df = 7, $P = 0.131$).

For several reasons, however, we urge further work on the question of heterogeneity in effect sizes based on the type of active learning intervention. The sample sizes in our analysis are highly unbalanced, with some types of interventions (e.g., interactive demonstrations and case histories) poorly represented. In addition, a large suite of experiments was coded simply as "worksheets" because authors referred to "cooperative in-class exercises," "group problem solving," or "in-class tutorials." Coding was imprecise in these and many other cases because at present there is no consensus about what various active learning types actually entail in practice. For example, papers routinely use the term "problem-based learning" for what one author considers eight distinct course designs (48). Similar issues can arise in implementation of general strategies such as "cooperative group learning." The problem was exacerbated because authors rarely provided details on the intensity, duration, questioning level, group composition, and tasks involved in active learning interventions. Further, few studies noted whether assigned exercises had direct consequences for students—specifically, whether they were ungraded or graded, graded for participation or correctness, and for what percentage of total course grade.

We include the data on heterogeneity by intervention type here only to urge further work on the issue, as discipline-based education research begins its second generation. Progress will depend on the research community creating an objective and repeatable classification of intervention type—preferably grounded in theory and empirical work from cognitive science and educational psychology. To understand heterogeneity in effect sizes, we need a reliable taxonomy of active learning types.

Heterogeneity analyses: Failure rate data. Table S2*A* provides details on the heterogeneity analysis for failure rate data summarized in the *Results*–specifically, metaanalyzing the failure rate data by discipline. In addition, we metaanalyzed the failure rate data by the type of failure metric used, class size, and course level.

There was evidence for heterogeneity based on the type of failure metric used ($Q = 27.60$, df = 7, $P < 0.0001$; Table S2*B*), although the result should be interpreted cautiously because five of the categories are represented by tiny samples. Because it is the most comprehensive and interpretable metric of failure, we urge future investigators to use DFW as the sole metric for quantifying failure rates in STEM courses. Given these caveats, it is interesting to note that the analysis (Table S2*B*) indicates that the log-odds of students withdrawing from a course are much higher under lecturing. If these students were performing poorly before withdrawing but would be retained in active learning sections, it suggests that the overall effect size of 0.47 on assessment performance may be conservative, as noted in the *Discussion*.

The analysis based on class size just misses statistical significance ($Q = 5.91$, df = 2, $P = 0.052$). The data in Table S2*B* are consistent with the result for the examination scores (assessment) data. Active learning appears to work slightly better in classes with 50 or fewer students, but is effective in lowering failure rates in any class size. As with the assessment data, there is no impact on failure rate based on using active learning in introductory versus upper division classes ($Q = 0.71$, df = 1, $P = 0.40$).

The effect sizes and moderator variables for each independent study analyzed here are provided in Table S4 (49–234).

1. 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.
2. 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.
3. 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.
4. Alfieri L, Brooks PJ, Aldrich NJ, Tenenbaum HR (2011) Does discovery-based instruction enhance learning? *J Educ Psychol* 103(1):1–18.
5. Bailey JM, Slater TF (2004) A review of astronomy education research. *Astron Educ Rev* 2:20–45.
6. Bailey JM, Slater TF (2005) Resource letter AER-1: Astronomy education research. *Am J Phys* 73(8):677–685.
7. 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.
8. Cain J, Robinson E (2008) A primer on audience response systems: Current applications and future considerations. *Am J Pharm Educ* 72(4):77.
9. Caldwell JE (2007) Clickers in the large classroom: Current research and best-practice tips. *CBE Life Sci Educ* 6(1):9–20.
10. 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 Educ* 98(4):335–348.
11. Daempfle PA (2006) The effects of instructional approaches on the improvement of reasoning in introductory college biology: A quantitative review of research. *BioScene* 32:22–31.
12. Darken B, Wynegar R, Kuhn S (2000) Evaluating calculus reform: A review and a longitudinal study. *CBMS Issues Math Ed* 8:16–41.
13. Davidson N (1985) Small-group learning and teaching in mathematics. *Learning to Cooperate, Cooperating to Learn: Second Conference of the IASCE*, ed Slavin R (Plenum, New York), pp 211–230.
14. Evans KM (1965) An annotated bibliography of British research on teaching and teaching ability. *Educ Res* 4:67–80.
15. Fies C, Marshall J (2006) Classroom response systems: A review of the literature. *J Sci Educ Technol* 15:101–109.
16. Froyd JE (2008) *Evidence for the Efficacy of Student-Active Learning Pedagogies* (Project Kaleidoscope, Washington, DC).
17. Geer UC, Rudge DW (2002) A review of research on constructivist-based strategies for large lecture science classes. *Electron J Sci Educ* 7(2):1–22.
18. Glick JG (1994) *Effective Methods for Teaching Nonmajors Introductory College Biology: A Critical Literature Review* (Education Resources Information Center, Washington).
19. Hsu L, Brewe E, Foster TM, Harper KA (2004) Resource letter RPS-1: Research in problem-solving. *Am J Phys* 72:1147–1156.
20. Johnson DW, Maruyama G, Johnson R, Nelson D, Skon L (1981) Effects of cooperative, competitive, and individualistic goal structures on achievement: A meta-analysis. *Psychol Bull* 89:47–62.
21. Johnson DW, Johnson RT, Smith KA (1998) Cooperative learning returns to college. *Change* 30:26–35.
22. Judson E, Sawada D (2002) Learning from the past and present: Electronic response systems in college lecture halls. *J Comput Math Sci Teach* 21:167–181.
23. Kay RH, LeSage A (2009) Examining the benefits and challenges of using audience response systems: A review of the literature. *Comput Educ* 53:819–827.
24. Major CH, Palmer B (2001) Assessing the effectiveness of problem-based learning in higher education: Lessons from the literature. *Acad Exch Q* 5:4–9.
25. McDermott LC, Redish EF (1999) Resource letter: PER1: Physics education research. *Am J Phys* 67:755–767.
26. Michael J (2006) Where's the evidence that active learning works? *Adv Physiol Educ* 30(4):159–167.
27. Olds BM, Moskal BM, Miller RL (2005) Assessment in engineering education: Evolution, approaches and future collaborations. *J Eng Educ* 94:13–25.
28. Prince M (2005) Does active learning work? A review of the research. *J Eng Educ* 93(3):223–231.
29. Prince M, Felder R (2006) Inductive teaching and learning methods: Definitions, comparisons, and research bases. *J Eng Educ* 95:123–138.
30. Roschelle J, Penuel WR, Abrahamson L (2004) Classroom response and communication systems: Research review and theory. *American Educational Research Association, Annual Meeting* (AERA, San Diego).
31. Schoenfeld AH (1994) Some notes on the enterprise (research on collegiate mathematics education, that is). *CBMS Issues Math Ed* 4:1–19.
32. 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.
33. Tomcho TT, Foels R (2008) Assessing effective teaching of psychology: A meta-analytic integration of learning outcomes. *Teach Psychol* 35:286–296.
34. Tomcho TT, et al. (2008) Review of ToP teaching strategies: Links to students' scientific inquiry skills development. *Teach Psychol* 35:147–159.

35. Weller K, et al. (2003) Student performance and attitudes in courses based on APOS theory and the ACE teaching cycle. *CBMS Issues Math Ed* 12:97–131.
36. Zieffler A, et al. (2008) What does research suggest about the teaching and learning of introductory statistics at the college level? A review of the literature. *J Stat Educ* 16:1–25.
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, Baltimore). Available at www.cochrane-handbook.org. Accessed May 2, 2013.
39. Freeman S, et al. (2007) Prescribed active learning increases performance in introductory biology. *CBE Life Sci Educ* 6(2):132–139.
40. Freeman S, Haak D, Wenderoth MP (2011) Increased course structure improves performance in introductory biology. *CBE Life Sci Educ* 10(2):175–186.
41. Lipsey MW, Wilson DB (2001) *Practical Meta-Analysis* (Sage Publications, Thousand Oaks, CA).
42. Gurevitch J, Hedges LV (1999) Statistical issues in ecological meta-analyses. *Ecology* 80(4):1142–1149.
43. 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.
44. 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.
45. 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.
46. 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.
47. Slavin RE (1996) *Education for All* (Swets & Zeitlinger, Lisse, The Netherlands).
48. Barrows HS (1986) A taxonomy of problem-based learning methods. *Med Educ* 20(6): 481–486.
49. Beichner R, et al. (1999) Case study of the physics component of an integrated curriculum. *Am J Phys* 67(7):S16–S24.
50. Crider A (2005) Hot seat questioning: A technique to promote and evaluate student dialogue. *Astron Educ Rev* 3(2):137–147.
51. Beichner RJ, et al. (2007) The student-centered activities for large enrollment undergraduate programs (SCALE-UP) project. *Phys Educ Res* 1(1):42.
52. Cummings K, Marx J, Thornton R, Kuhl D (1999) Evaluating innovation in studio physics. *Am J Phys* 67(1):S38.
53. Anderson WL, Mitchell SM, Osgood MP (2005) Comparison of student performance in cooperative learning and traditional lecture-based biochemistry classes. *Biochem Mol Biol Educ* 33(6):387–393.
54. Armbruster P, Patel M, Johnson E, Weiss M (2009) Active learning and student-centered pedagogy improve student attitudes and performance in introductory biology. *CBE Life Sci Educ* 8(3):203–213.
55. Armstrong GM, Hendrix LJ (1999) Does traditional or reformed caculus prepare students better for subsequent courses? A preliminary study. *J Comput Math Sci Teach* 18(2):95–103.
56. Armstrong N, Chang S-M, Brickman M (2007) Cooperative learning in industrial-sized biology classes. *CBE Life Sci Educ* 6(2):163–171.
57. Asiala M, Cottrill J, Dubinsky E, Schwingendorf KE (1997) The development of students' graphic understanding of derivatives. *J Math Behav* 16(4):399.
58. Asiala M, Dubinsky E, Mathews DM, Morics S, Oktaç A (1997) Development of students' understanding of cosets, normality, and quotient groups. *J Math Behav* 16 (3):241–309.
59. Austin DA (1995) Effect of cooperative learning in finite mathematics on student achievement and attitude. PhD thesis (Illinois State Univ, Normal, IL).
60. Banerjee AC, Vidyapati TJ (1997) Effect of lecture and cooperative learning strategies on achievement in chemistry in undergraduate classes. *Int J Sci Educ* 19(8): 903–910.
61. Bardar EM, Brecher K (2008) Project LITE educational materials and their effectiveness as measured by the light and spectroscopy concept inventory. *Astron Educ Rev* 6(2):85–98.
62. Barg M, et al. (2000) Problem-based learning for foundation computer science courses. *Comput Sci Educ* 10(2):109–128.
63. Barnard JD (1942) The lecture-demonstration versus the problem-solving method of teaching a college science course. *Sci Educ* 26(3-4):121–132.
64. Basili PA, Sanford JP (1991) Conceptual change strategies and cooperative group work in chemistry. *J Res Sci Teach* 28:293–304.
65. Beck LL, Chizhik AW (2008) An experimental study of cooperative learning in CS1. *SIGCSE Bull* 40(1):205–209.
66. Bilgin I (2006) Promoting pre-service elementary students' understanding of chemical equilibrium through discussions in small groups. *Int J Sci Math Educ* 4(3): 467–484.
67. Bilgin I, Senocak E, Sozbilir M (2009) The effects of problem-based learning instruction on university students' performance of conceptual and quantitative problems in gas concepts. *Eurasia J Math Sci & Tech* 5(2):153–164.
68. Blue JM (1997) Sex differences in physics learning and evaluations in an introductory course PhD thesis. Univ of Minnesota, Twin Cities, Minneapolis, MN.
69. Bookman J, Friedman CP (1994) A comparison of the problem solving performance of students in lab based and traditional calculus. *CBMS Issues Math Ed* 4:101–116.
70. Booth KM, James BW (2001) Interactive learning in a higher education Level 1 mechanics module. *Int J Sci Educ* 23(9):955–967.
71. Born DG, Gledhill SM, Davis ML (1972) Examination performance in lecture-discussion and personalized instruction courses. *J Appl Behav Anal* 5(1):33–43.
72. Boyles MP, Killian PW, Rileigh KK (1994) Learning by writing in introductory psychology. *Psychol Rep* 75(1):563–568.
73. Bradley AZ, Ulrich SM, Jones M, Jr., Jones SM (2002) Teaching the sophomore organic course without a lecture. Are you crazy? *J Chem Educ* 79(4):514–519.
74. Brown JD (1972) An evaluation of the Spitz student response system in teaching a course in logical and mathematical concepts. *J Exp Educ* 40(3):12–20.
75. Buck JR, Wage KE (2005) Active and cooperative learning in signal processing courses. *IEEE Signal Process Mag* 22(2):76–81.
76. Bullard L, Felder R, Raubenheimer D (2008) Effects of active learning on student performance and retention. *ASEE Annual Conference Proceedings*, (American Society for Engineering Education, Washington).
77. Bunting CF, Cheville RA (2009) (VECTOR: A hands-on approach that makes electromagnetics relevant to students. *IEEE Trans Educ* 52(3):350–359.
78. Burrowes PA (2003) A student-centered approach to teaching general biology that really works: Lord's constructivist model put to a test. *Am Biol Teach* 65(7):491–502.
79. Cahyadi V (2004) The effect of interactive engagement teaching on student understanding of introductory physics at the faculty of engineering, University of Surabaya, Indonesia. *High Educ Res Dev* 23(4):455–464.
80. Carmichael J (2009) Team-based learning enhances performance in introductory biology. *J Coll Sci Teach* 38(4):54–61.
81. Chaplin S (2009) Assessment of the impact of case studies on student learning gains in an introductory biology course. *J Coll Sci Teach* 39(1):72–79.
82. Cheng KK, Thacker BA, Cardenas RL, Crouch C (2004) Using an online homework system enhances students' learning of physics concepts in an introductory physics course. *Am J Phys* 72(11):1447–1453.
83. Christensen T (2005) Changing the learning environment in large general education astronomy classes. *J Coll Sci Teach* 35(3):34.
84. Crouch CH, Mazur E (2001) Peer instruction: Ten years of experience and results. *Am J Phys* 69(9):970–977.
85. Davis M, Hult RE (1997) Effects of writing summaries as a generative learning activity during note taking. *Teach Psychol* 24(1):47–50.
86. Day JA, Foley JD (2006) Evaluating a web lecture intervention in a human-computer interaction course. *IEEE Trans Educ* 49(4):420–431.
87. Dees RL (1991) The role of cooperative learning in increasing problem-solving ability in a college remedial course. *J Res Math Educ* 22(5):409–421.
88. Demetry C, Groccia JE (1997) A comparative assessment of students' experiences in two instructional formats of an introductory materials science course. *J Eng Educ* 86(3):203–210.
89. Dennis EC (2001) An investigation of the numerical experience associated with the global behavior of polynomial functions in the traditional lecture method and cooperative learning method classes PhD thesis. (Illinois State Univ, Normal, IL).
90. Dollman J (2005) A new peer instruction method for teaching practical skills in the health sciences: An evaluation of the 'Learning Trail.' *Adv Health Sci Educ Theory Pract* 10(2):125–132.
91. Dougherty RC, et al. (1995) Cooperative learning and enhanced communication: Effects on student performance, retention, and attitudes in general chemistry. *J Chem Educ* 72(9):722–726.
92. Doymus K (2007) Effects of a cooperative learning strategy on teaching and learning phases of matter and one-component phase diagrams. *J Chem Educ* 84(11): 1857–1860.
93. Ellington AJ (2005) A modeling-based college algebra course and its effect on student achievement. *PRIMUS (Terre Ht Ind)* 15(3):193–214.
94. Enriquez AG (2008) Impact of Tablet PC-enhanced interactivity on student performance in sophomore-level engineering dynamics course. *Comp Educ J* 18(3): 69–84.
95. Erwin TD, Rieppi R (1999) Comparing multimedia and traditional approaches in undergraduate psychology classes. *Teach Psychol* 26(1):58–61.
96. Fallahi CR (2008) Redesign of a life span development course using Fink's taxonomy. *Teach Psychol* 35(3):169–175.
97. Felder RM, Felder GN, Dietz JE (1998) A longitudinal study of engineering student performance and retention. V. Comparisons with traditionally-taught students. *J Eng Educ* 87(4):469–470.
98. Forbes CA (1997) Analyzing the growth of the critical thinking skills of college calculus students PhD thesis. (Iowa State Univ, Ames IA).
99. Froelich AG, Duckworth WM (2008) Assessment of materials for engaging students in statistical discovery. *J Stat Educ* 16(2):1–29.
100. Gifford VD, Vicks J (1982) A comparison of the personalized system of instruction and a conventional biology course on the achievement of junior college freshmen. *J Res Sci Teach* 19(8):659–664.
101. Giraud G (1997) Cooperative learning and statistics instruction. *J Stat Educ* 5(3):1–13.
102. Gonzalez G (2006) A systematic approach to active and cooperative learning in CS1 and its effects on CS2. *SIGCSE Bull* 38(1):133–137.
103. Graham RB (1999) Unannounced quizzes raise test scores selectively for mid-range students. *Teach Psychol* 26(4):271–273.
104. Gruner HM (1997) A study of the cognitive and affective impact of the cockpit physics curriculum on students at the United States Air Force Academy. PhD thesis (Kansas State Univ, Manhattan, KS).
105. Gutwill-Wise JP (2001) The impact of active and context-based learning in introductory chemistry courses: An early evaluation of the modular approach. *J Chem Educ* 78(5):684–690.

106. Haberyan KA (2003) Do weekly quizzes improve student performance on general biology exams? *Am Biol Teach* 65(2):110–114.
107. Hagen JP (2000) Cooperative learning in organic II. Increased retention on a commuter campus. *J Chem Educ* 77(1):1441–1444.
108. Hake RR (1998) Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *Am J Phys* 66(1): 64–74.
109. Heron PRL, Loverude ME, Shaffer PS, McDermott LC (2003) Helping students develop an understanding of Archimedes principle. II. Development of research-based instructional materials. *Am J Phys* 71(11):1188–1195.
110. Hersam MC, Luna M, Light G (2004) Implementation of interdisciplinary group learning and peer assessment in a nanotechnology engineering course. *J Eng Educ* 93(1):49–57.
111. Hoellwarth C, Moelter MJ, Knight RD (2005) A direct comparison of conceptual learning and problem solving ability in traditional and studio style classrooms. *Am J Phys* 73(5):459–462.
112. Hsieh C, Knight L (2008) Problem-based learning for engineering students: An Evidence-based comparative study. *J Acad Librariansh* 34(1):25–30.
113. Hurley JD (2002) Effect of extended use of systematic instruction model on student achievement and content coverage in a "C" programming class. PhD thesis (Texas A&M Univ, College Station, TX).
114. Jensen MS, Finley FN (1996) Changes in students' understanding of evolution resulting from different curricular and instructional strategies. *J Res Sci Teach* 33(8): 879–900.
115. Johnson SD, Fischbach RM (1992) *Teaching Problem Solving and Technical Mathematics Through Cognitive Apprenticeship at the Community College Level* (National Center for Research in Vocational Education, Univ of California, Berkeley, CA).
116. Keeler CM, Steinhorst RK (1994) Cooperative learning in statistics. *Teach Stat* 16(3): 81–84.
117. Kellum KK, Carr JE, Dozier CL (2001) Response-card instruction and student learning in a college classroom. *Teach Psychol* 28(2):101–104.
118. Kitchen E, Bell JD, Reeve S, Sudweeks RR, Bradshaw WS (2003) Teaching cell biology in the large-enrollment classroom: Methods to promote analytical thinking and assessment of their effectiveness. *Cell Biol Educ* 2(3):180–194.
119. Klionsky DJ (2002) Constructing knowledge in the lecture hall. *J Coll Sci Teach* 31(4): 246–251.
120. Knight JK, Wood WB (2005) Teaching more by lecturing less. *Cell Biol Educ* 4(4): 298–310.
121. Fortenberry NL, Sullivan JF, Jordan PN, Knight DW (2007) Retention. Engineering education research aids instruction. *Science* 317(5842):1175–1176.
122. Krause PA (1997) Promoting active learning in lecture-based courses: Demonstrations, tutorials, and interactive tutorial lectures. PhD thesis (Univ of Washington, Seattle, WA).
123. Krawiec S, Salter D, Kay E (2005) A "hybrid" bacteriology course: The professor's design and expectations; the students' performance and assessment. *Microbiol Educ* 6(1).
124. Kurbanoglu NI, Taskesenligil Y, Sozbilir M (2006) Programmed instruction revisited: A study on teaching stereochemistry. *Chem Educ Res Pract* 7(1):13.
125. Lasry N, Mazur E, Watkins J (2008) Peer instruction: From Harvard to the two-year college. *Am J Phys* 76(11):1066–1069.
126. Lawson TJ, Bodle JH, Houlette MA, Haubner RR (2006) Guiding questions enhance student learning from educational videos. *Teach Psychol* 33(1):31–33.
127. Lee AH (2009) Development and evaluation of clicker methodology for introductory physics courses. PhD thesis (Ohio State Univ, Columbus OH).
128. LeJeaune NF (2002) Problem-based learning instruction versus traditional instruction on self-directed learning, motivation, and grades of undergraduate computer science students. PhD thesis (Univ of Colorado at Denver, Denver, CO).
129. Lenaerts J, Wieme W, Zele EV (2003) Peer instruction: A case study for an introductory magnetism course. *Eur J Phys* 24(1):7.
130. Lewis SE, Lewis JE (2008) Seeking effectiveness and equity in a large college chemistry course: An HLM investigation of peer-led guided inquiry. *J Res Sci Teach* 45(7):794–811.
131. Lindaman BJ (2007) Making sense of the infinite: A study comparing the effectiveness of two approaches to teaching infinite series in calculus. PhD thesis (Univ of Kansas, Lawrence, KS).
132. Linsey J, Talley A, White C, Jensen D, Wood K (2009) From Tootsie Rolls to broken bones: An innovative approach for active learning in mechanics of materials. *Adv Eng Educ* 1(3):1–23.
133. Lohse B, Nitzke S, Ney DM (2003) Introducing a problem-based unit into a lifespan nutrition class using a randomized design produces equivocal outcomes. *J Am Diet Assoc* 103(8):1020–1025.
134. Lord TR (1997) A comparison between traditional and constructivist teaching in college biology. *Innovative High Educ* 21(3):197–216.
135. Lord TR (1999) A comparison between traditional and constructivist teaching in environmental science. *J Environ Educ* 30(3):22–27.
136. Lovelace TL, McKnight CK (1980) The effects of reading instruction on calculus students' problem solving. *J Read* 23(4):305–308.
137. Lucas CA (1999) A study of the effects of cooperative learning on the academic achievement and self-efficacy of college algebra students. PhD thesis. (Univ of Kansas, Lawrence, KS).
138. Marbach-Ad G, Sokolove PG (2000) Can undergraduate biology students learn to ask higher level questions? *J Res Sci Teach* 37(8):854–870.
139. Martin M (2009) The effect of active techniques combined with didactic lecture on student achievement. MS thesis (Arkansas State Univ, Jonesboro).
140. Mathew E (2008) Learning physics: A comparative analysis between instructional design methods. PhD thesis (Capella Univ, Minneapolis, MN).
141. Mauk HV, Hingley D (2005) Student understanding of induced current: Using tutorials in introductory physics to teach electricity and magnetism. *Am J Phys* 73(12): 1164–1171.
142. Mayer RE, et al. (2009) Clickers in college classrooms: Fostering learning with questioning methods in large lecture classes. *Contemp Educ Psychol* 34(1):51–57.
143. McConnell DA, Steer DN, Owens KD, Knight CC (2005) How students think: Implications for learning in introductory geoscience courses. *J Geosci Educ* 53(4):462.
144. McConnell DA, et al. (2006) Using concept tests to assess and improve student conceptual understanding in introductory geoscience courses. *J Geosci Educ* 54(1):61.
145. McCormick BD (2000) Attitude, achievement, and classroom environment in a learner-centered introductory biology course. PhD thesis (Univ of Texas at Austin, Austin, TX).
146. McDaniel CN, Lister BC, Hanna MH, Roy H (2007) Increased learning observed in redesigned introductory biology course that employed web-enhanced, interactive pedagogy. *CBE Life Sci Educ* 6(3):243–249.
147. McFarlin BK (2008) Hybrid lecture-online format increases student grades in an undergraduate exercise physiology course at a large urban university. *Adv Physiol Educ* 32(1):86–91.
148. Meel DE (1998) Honors students' calculus understandings: Comparing calculus & mathematica and traditional calculus students. *CBMS Issues Math Ed* 7:163–215.
149. Miller ML (1999) A study of the effects of reform teaching methodology and Ven Hiele level on conceptual learning in calculus. PhD thesis (Oklahoma State Univ–Tulsa, Tulsa, OK).
150. Mohamed A-R (2008) Effects of active learning variants on student performance and learning perceptions. *IJ-SOTL* 2(2):1–14.
151. Montelone BA, Rintoul DA, Williams LG (2008) Assessment of the effectiveness of the studio format in introductory undergraduate biology. *CBE Life Sci Educ* 7(2): 234–242.
152. Morling B, McAuliffe M, Cohen L, DiLorenzo TM (2008) Efficacy of personal response systems ("clickers") in large, introductory psychology classes. *Teach Psychol* 35(1): 45–50.
153. Nasr KJ, Ramadan BH (2008) Impact assessment of problem-based learning in an engineering science course. *J STEM Edu* 9(3/4):16–24.
154. Nehm RH, Reilly L (2007) Biology majors' knowledge and misconceptions of natural selection. *Bioscience* 57(3):263–272.
155. Nelson J, Robison DF, Bell JD, Bradshaw WS (2009) Cloning the professor, an alternative to ineffective teaching in a large course. *CBE Life Sci Educ* 8(3):252–263.
156. Norwood KS (1995) The effects of the use of problem solving and cooperative learning on the mathematics achievement of underprepared college freshmen. *PRIMUS (Terre Ht Ind)* 5(3):229–252.
157. O'Brien TH (1993) The effects of cooperative learning versus lecture on the attitudes, achievement, and attrition rates of college algebra students. PhD thesis (Univ of Arkansas at Little Rock, Little Rock, AR).
158. Oliver-Hoyo M, Allen D, Hunt WF, Hutson J, Pitts A (2004) Effects of an active learning environment: Teaching innovations at a research I institution. *J Chem Educ* 81(3):441–448.
159. O'Sullivan DW, Copper CL (2003) Evaluating active learning: A new initiative for a general chemistry curriculum. *J Coll Sci Teach* 32(7):448–452.
160. Overlock TH, Sr. (1994) Comparison of effectiveness of collaborative learning methods and traditional methods in physics classes at Northern Maine Technical College. PhD thesis (Nova Southeastern Univ, Fort Lauderdale, FL).
161. Pandy MG, Petrosino AJ, Austin BA, Barr RE (2004) Assessing adaptive expertise in undergraduate biomechanics. *J Eng Educ* 93(3):211–222.
162. Park K, Travers KJ (1996) A comparative study of a computer-based and a standard college first-year calculus course. *CBMS Issues Math Ed* 6:155–176.
163. Paschal CB (2002) Formative assessment in physiology teaching using a wireless classroom communication system. *Adv Physiol Educ* 26(1-4):299–308.
164. Poirier CR, Feldman RS (2007) Promoting active learning using individual response technology in large introductory psychology classes. *Teach Psychol* 34(3):194–196.
165. Pollock SJ, Finkelstein ND (2008) Sustaining educational reforms in introductory physics. *Phys Rev ST Phys Educ Res* 4(1):010110-010111–010110-010118.
166. Prather EE, Brissenden G, Schlingman WM, Rudolph AL (2009) A national study assessing the teaching and learning of introductory astronomy. Part I. The effect of interactive instruction. *Am J Phys* 77(4):320–330.
167. Priebe RL (1997) The effects of cooperative learning on content comprehension and logical reasoning in a second-semester university computer science course. PhD thesis (Univ of Texas at Austin, Austin, TX).
168. Randolph WM (1992) *The Effect of Cooperative Learning on Academic Achievement in Introductory College Biology.* PhD thesis (Washington State Univ, Pullman, WA).
169. Redish EF, Saul JM, Steinberg RN (1997) On the effectiveness of active-engagement microcomputer-based laboratories. *Am J Phys* 65(1):45.
170. Robinson WR, Niaz M (1991) Performance based on instruction by lecture or by interaction and its relationship to cognitive variables. *Int J Sci Educ* 13(2):203.
171. Roddick CD (2001) Differences in learning outcomes: Calculus and mathematica vs. traditional calculus. *PRIMUS (Terre Ht Ind)* 11(2):161–184.
172. Roselli RJ, Brophy SP (2006) Effectiveness of challenge-based instruction in biomechanics. *J Eng Educ* 95(4):311–324.
173. Ross MR, Fulton RB (1994) Active learning strategies in the analytical chemistry classroom. *J Chem Educ* 71(2):141–143.
174. Rutz E, et al. (2003) Student performance and acceptance of instructional technology: Comparing technology-enhanced and traditional instruction for a course in statics. *J Eng Educ* 92(2):133–140.

175. Rybarczyk BJ, Baines AT, McVey M, Thompson JT, Wilkins H (2007) A case-based approach increases student learning outcomes and comprehension of cellular respiration concepts. *Biochem Mol Biol Educ* 35(3):181–186.
176. Sadler KC (2002) The effectiveness of cooperative learning as an instructional strategy to increase biological literacy and academic achievement in a large, nonmajors college biology class. PhD thesis (Tennessee State Univ, Nashville, TN).
177. Schielak DJF (1988) A cooperative learning, laboratory approach in a mathematics course for prospective elementary teachers. PhD thesis (Texas A&M Univ, College Station TX).
178. Shatila A (2007) Assessing the impact of integrating POGIL in elementary organic chemistry. PhD thesis (Univ of Southern Mississippi, Hattiesburg, MS).
179. Sorensen CM, Churukian AD, Maleki S, Zollman DA (2006) The New Studio format for instruction of introductory physics. *Am J Phys* 74(12):1077–1082.
180. Steele JE (2003) Effect of essay-style lecture quizzes on student performance on anatomy and physiology exams. *Bioscene* 29(4):15–20.
181. Tanahoung C, Chitaree R, Soankwan C, Sharma MD, Johnston ID (2009) The effect of interactive lecture demonstrations on students' understanding of heat and temperature: A study from Thailand. *Res Sci Technol Educ* 27(1):61–74.
182. Tessier J (2007) Small-group peer teaching in an introductory biology Classroom. *J Coll Sci Teach* 36(4):64–69.
183. Tien LT, Roth V, Kampmeier JA (2002) Implementation of a peer-led team learning instructional approach in an undergraduate organic chemistry course. *J Res Sci Teach* 39(7):606–632.
184. Turns SR, Pauley LL, Zappe SE (2009) Active and collaborative learning in a first course in fluid mechanics: Implementation and transfer. *Int J Eng Educ* 25(5): 979–997.
185. Udovic D, Morris D, Dickman A, Postlethwait J, Wetherwax P (2002) Workshop biology: Demonstrating the effectiveness of active learning in an introductory biology course. *Bioscience* 52(3):272–281.
186. Usoh II (2003) An investigation into the effectiveness of problem-based learning in an engineering technology program at Nashville State Technical Community College. PhD thesis (Tennessee State Univ, Nashville, TN).
187. Valentino VR (1988) A study of achievement, anxiety, and attitude toward mathematics in college algebra students using small group interaction methods. PhD thesis (West Virginia Univ, Morgantown, WV).
188. Van Dijk LA, Van Der Berg GC, Van Keulen H (2001) Interactive lectures in engineering education. *Eur J Eng Educ* 26(1):15–28.
189. Van Gorp M, Grissom S (2001) An empirical evaluation of using constructive classroom activities to teach introductory programming. *Comput Sci Educ* 11(3): 247–260.
190. Van Heuvelen A (1991) Overview, case study physics. *Am J Phys* 59(10):898–907.
191. Walker JD, Cotner SH, Baepler PM, Decker MD (2008) A delicate balance: Integrating active learning into a large lecture course. *CBE Life Sci Educ* 7(4):361–367.
192. Weir JA (2004) Active learning in transportation engineering education. PhD thesis (Worcester Polytechnic Institute, Worcester, MA).
193. Wilke RR, Straits WJ (2001) The effects of discovery learning in a lower-division biology course. *Adv Physiol Educ* 25(1-4):134–141.
194. Williamson VM, Rowe MW (2002) Group problem-solving versus lecture in college-level quantitative analysis: The good, the bad, and the ugly. *J Chem Educ* 79(9): 1131–1134.
195. Yoder JD, Hochevar CM (2005) Encouraging active learning can improve students' performance on examinations. *Teach Psychol* 32(2):91–95.
196. Zimrot R, Ashkenazi G (2007) Interactive lecture demonstrations: A tool for exploring and enhancing conceptual change. *Chem Educ Res Pract* 8(2):197.
197. Al-Holou N, et al. (1999) First-year integrated curricula: Design alternatives and examples. *J Eng Educ* 88(4):435–448.
198. Alsardary S, Blumberg P (2009) Interactive, learner-centered methods of teaching mathematics. *PRIMUS (Terre Ht Ind)* 19(4):401–416.
199. Barak M, Harward J, Kocur G, Lerman S (2007) Transforming an introductory programming course: From lectures to active learning via wireless laptops. *J Sci Educ Technol* 16(4):325–336.
200. Barbarick KA (1998) Exam frequency comparison in introductory soil science. *J Nat Resour Life Sci Educ* 27:55–58.
201. Becvar JE, Dreyfuss AE, Flores BC, Dickson WE (2008) 'Plus two' peer-led team learning improves student success, retention, and timely graduation. *38th ASEE/IEEE Frontiers in Education Conference* (Institute of Electrical and Electronics Engineers, Saratoga Springs, NY).
202. Bernold LE, Bingham WL, McDonald PH, Attia TM (2000) Impact of holistic and learning-oriented teaching on academic success. *J Eng Educ* 89(2):191–199.
203. Biggers M, Yilmaz T, Sweat M (2009) Using collaborative, modified peer led team learning to improve student success and retention in intro cs. *SIGCSE Bull* 41(1):9–13.
204. Borglund D (2007) A case study of peer learning in higher aeronautical education. *Eur J Eng Educ* 32(1):35–42.
205. Broyles ML (1999) A comparison of the participation in cooperative learning on the success of physics, engineering and mathematics students. PhD thesis (Texas A&M Univ–Commerce, Commerce, TX).
206. Buxeda RJ, Moore DA (2000) Transforming a sequence of microbiology courses using student profile data. *Microbiol Educ* 1(1):1–6.
207. Casanova J (1971) An instructional experiment in organic chemistry. The use of a student response system. *J Chem Educ* 48(7):453–455.
208. Chambers SLK (2008) Improving student performance in an introductory biology majors course: A social action project in the scholarship of teaching. PhD thesis (Union Institute & Univ, Cincinnati, OH).
209. Chase JD, Okie EG (2000) Combining cooperative learning and peer instruction in introductory computer science. *SIGCSE Bull* 32(1):372.
210. DePree J (1998) Small-group instruction: Impact on basic algebra students. *J Dev Educ* 22(1):2–6.
211. Farrell JJ, Moog RS, Spencer JN (1999) A guided inquiry general chemistry course. *J Chem Educ* 76(4):570–574.
212. Gautreau R, Novemsky L (1997) Concepts first-a small group approach to physics learning. *Am J Phys* 65(5):418–428.
213. Gosser DK, Jr. (2011) The PLTL boost: A critical review of research. *J PLTL* 14(1):2–14.
214. Hoffman EA (2001) Successful application of active learning techniques to introductory microbiology. *Microbiol Educ* 2(1).
215. Howles T (2009) a study of attrition and the use of student learning communities in the computer science introductory programming sequence. *Comput Sci Educ* 19(1): 1–13.
216. Hoyt M, Ohland M (1998) The impact of a discipline-based introduction to engineering course on improving retention. *J Eng Educ* 87(1):79–85.
217. Kashy E, Thoennessen M, Tsai Y, Davis NE, Wolfe SL (1998) Using networked tools to promote student success in large classes. *Soc Educ* 62(7):385–390.
218. Keeler CM, Voxman M (1994) The effect of cooperative learning in remedial freshman level mathematics. *AMATYC Review* 16(1):37–44.
219. Koch LC (1992) Revisiting mathematics. *J Dev Educ* 16(1):12–14, 16, 18.
220. Kurki-Suonio T, Hakola A (2007) Coherent teaching and need-based learning in science: An approach to teach engineering students in basic physics courses. *Eur J Eng Educ* 32(4):367–374.
221. Lasserre P (2009) Adaptation of team-based learning on a first term programming class. *SIGCSE Bull* 41(3):186–190.
222. Maggelakis S, Lutzer C (2007) Optimizing student success in calculus. *PRIMUS (Terre Ht Ind)* 17(3):284–299.
223. Maher RJ (1998) Small groups for general student audiences-2. *PRIMUS (Terre Ht Ind)* 8(3):265–275.
224. Marrs KA, Novak G (2004) Just-in-Time Teaching in biology: Creating an active learner classroom using the Internet. *Cell Biol Educ* 3(1):49–61.
225. McConnell DA, Steer DN, Owens KD (2003) Assessment and active learning strategies for introductory geology courses. *J Geosci Educ* 51(2):205–216.
226. Mittag KC, Collins LB (2000) Relating calculus-I reform experience to performance in traditional calculus-II. *PRIMUS (Terre Ht Ind)* 10(1):82–94.
227. Nirmalakhandan N, Ricketts C, McShannon J, Barrett S (2007) Teaching tools to promote active learning: Case study. *J Prof Issues Eng Educ Pract* 133(1):31–37.
228. Nuutila E, Törmä S, Malmi L (2005) PBL and computer programming - The Seven Steps Method with adaptations. *Comput Sci Educ* 15(2):123–142.
229. Paulson DR (1999) Active learning and cooperative learning in the organic chemistry lecture class. *J Chem Educ* 76(8):1136–1140.
230. Poulis J, Massen C, Robens E, Gilbert M (1998) Physics lecturing with audience paced feedback. *Am J Phys* 66(5):439–441.
231. Preszler RW (2009) Replacing lecture with peer-led workshops improves student learning. *CBE Life Sci Educ* 8(3):182–192.
232. Quitadamo IJ, Brahler CJ, Crouch GJ (2009) Peer-led team learning: A prospective method for increasing critical thinking in undergraduate science courses. *Sci Educator* 18(1):29–39.
233. Rogerson BJ (2003) Effectiveness of a daily class progress assessment technique in introductory chemistry. *J Chem Educ* 80(2):160–164.
234. Steinberg RN, Donnelly K (2002) PER-based reform at a multicultural institution. *Phys Teach* 40(2):108–114.

Fig. S1. Changes in failure rate. Data points on failure rates under traditional lecturing and active learning from the same study—and thus from the same course and comparable student populations—are connected by a line. The overall means for each treatment are indicated by the green bars.

Fig. S1

Fig. S2. Forest plots for data on examinations, concept inventories, and other assessments, organized by discipline. Horizontal lines indicate 95% confidence intervals; detailed data are given in Table S4*A*.

Fig. S2

Fig. S3. Forest plots for data on failure rates, organized by discipline. The dashed, red, vertical line indicates the odds ratio indicating no effect. Horizontal lines indicate 95% confidence intervals; detailed data are given in Table S4*B*.

Fig. S3

Fig. S4. Distribution of class sizes in the study.

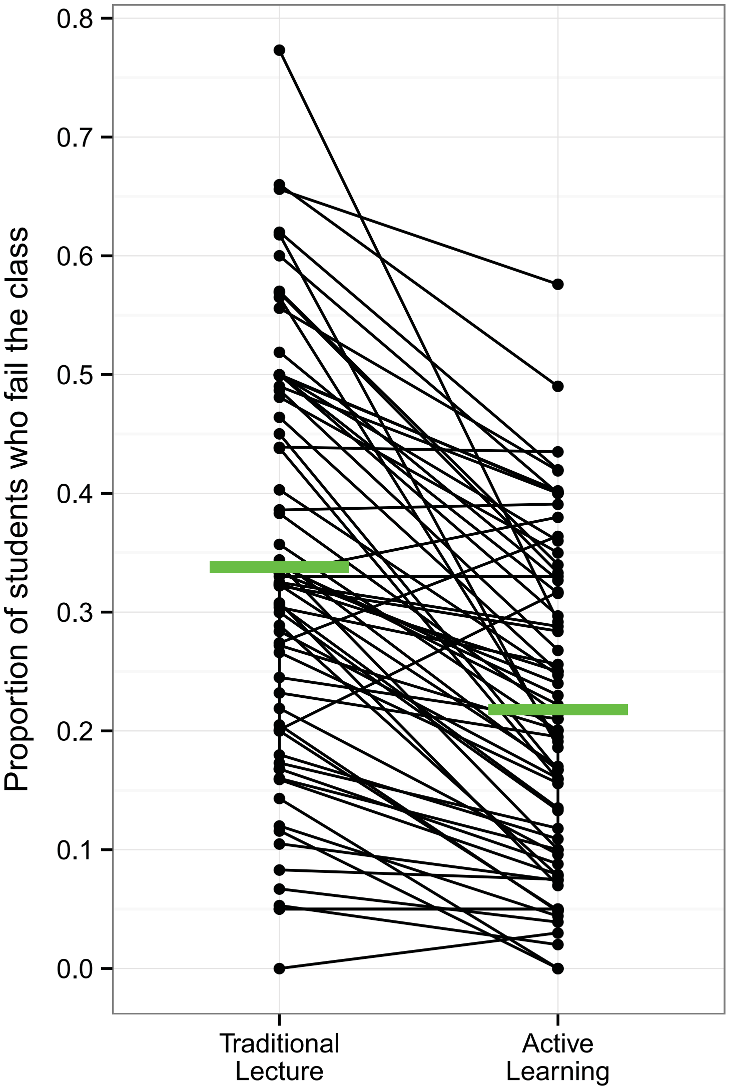
Fig. S4

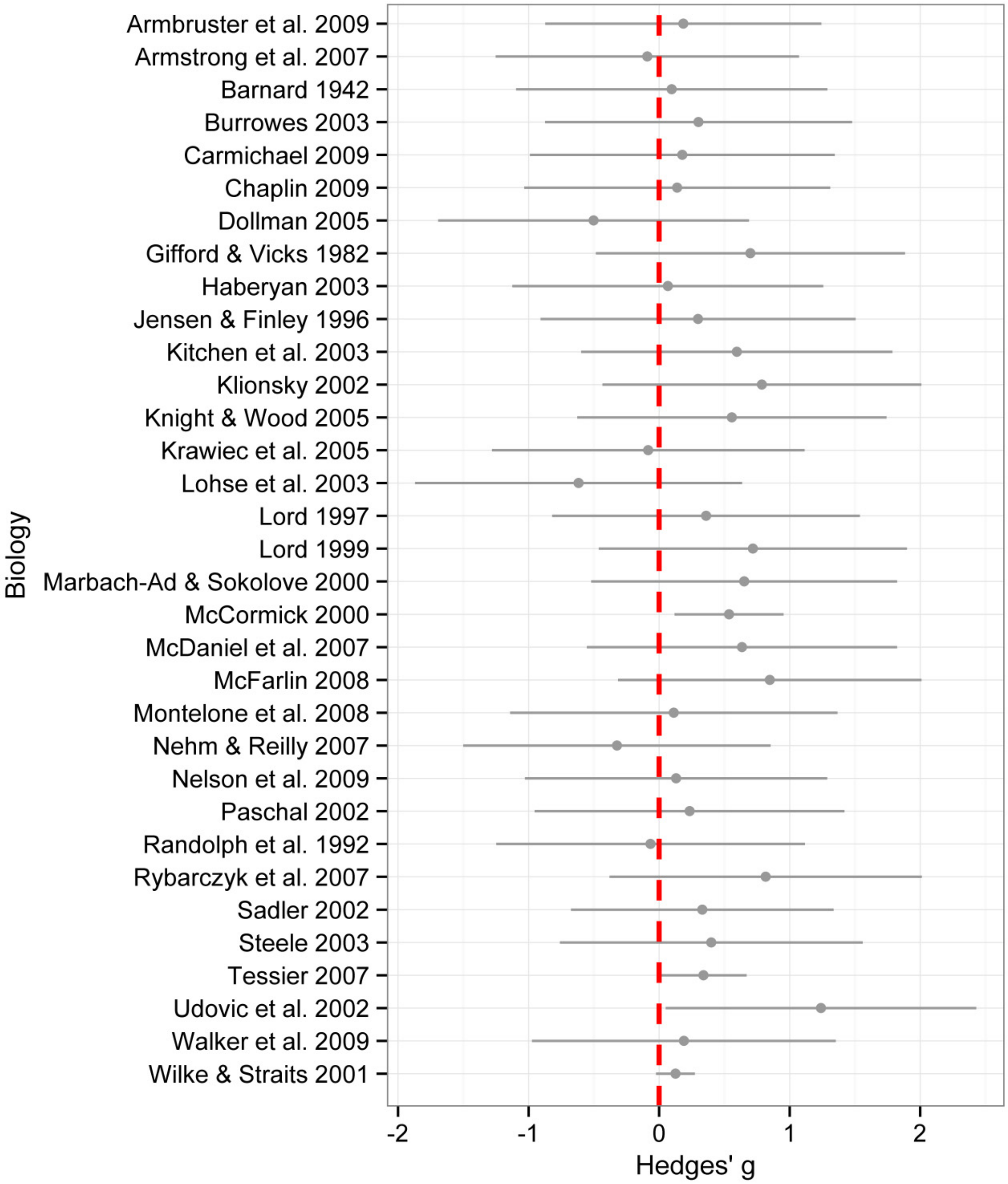
Fig. S5. Funnel plots for evaluating publication bias. (*A*) Funnel plot for examination score data (n = 160). (*B*) Funnel plot for failure rate data (n = 67). Note that the log odds ratio is plotted here.

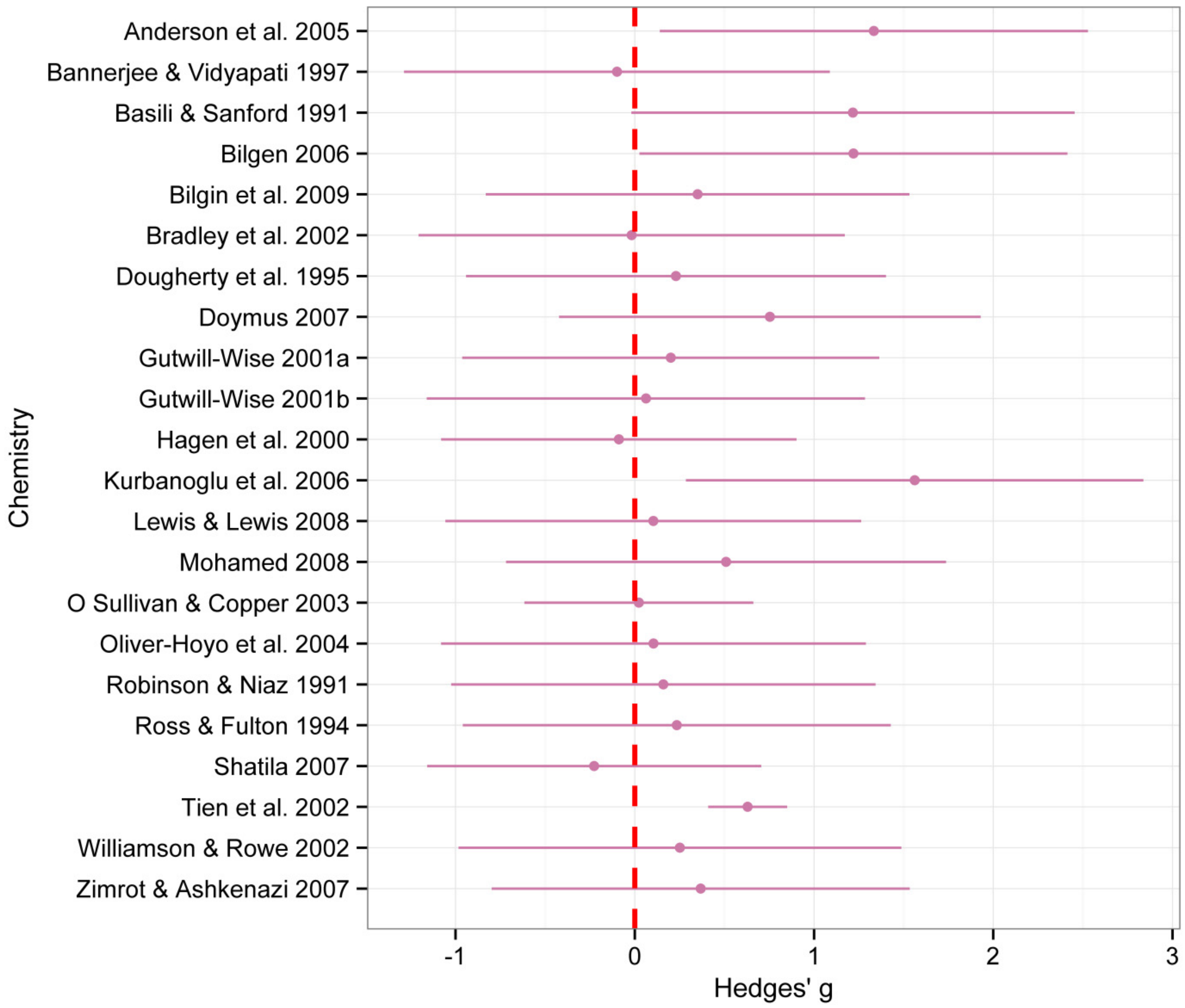
Fig. S5

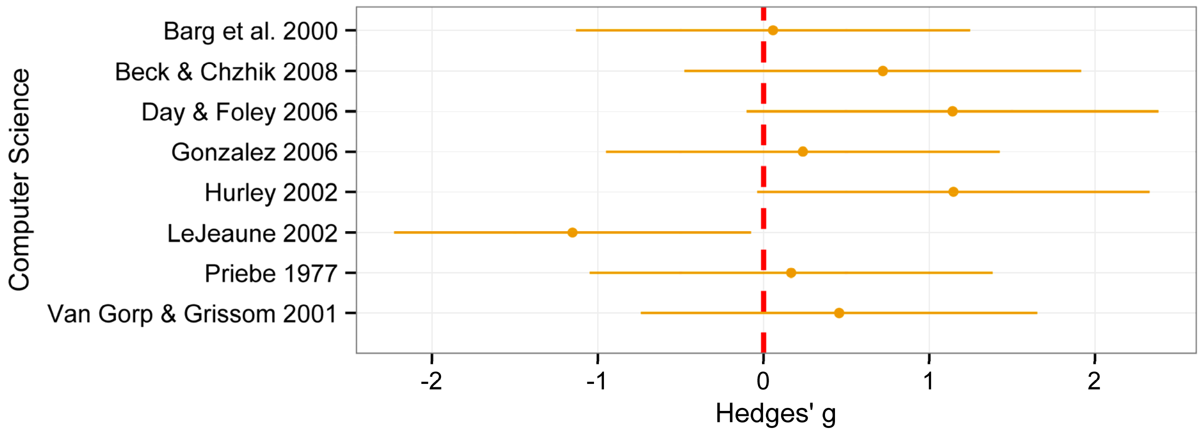
Other Supporting Information Files

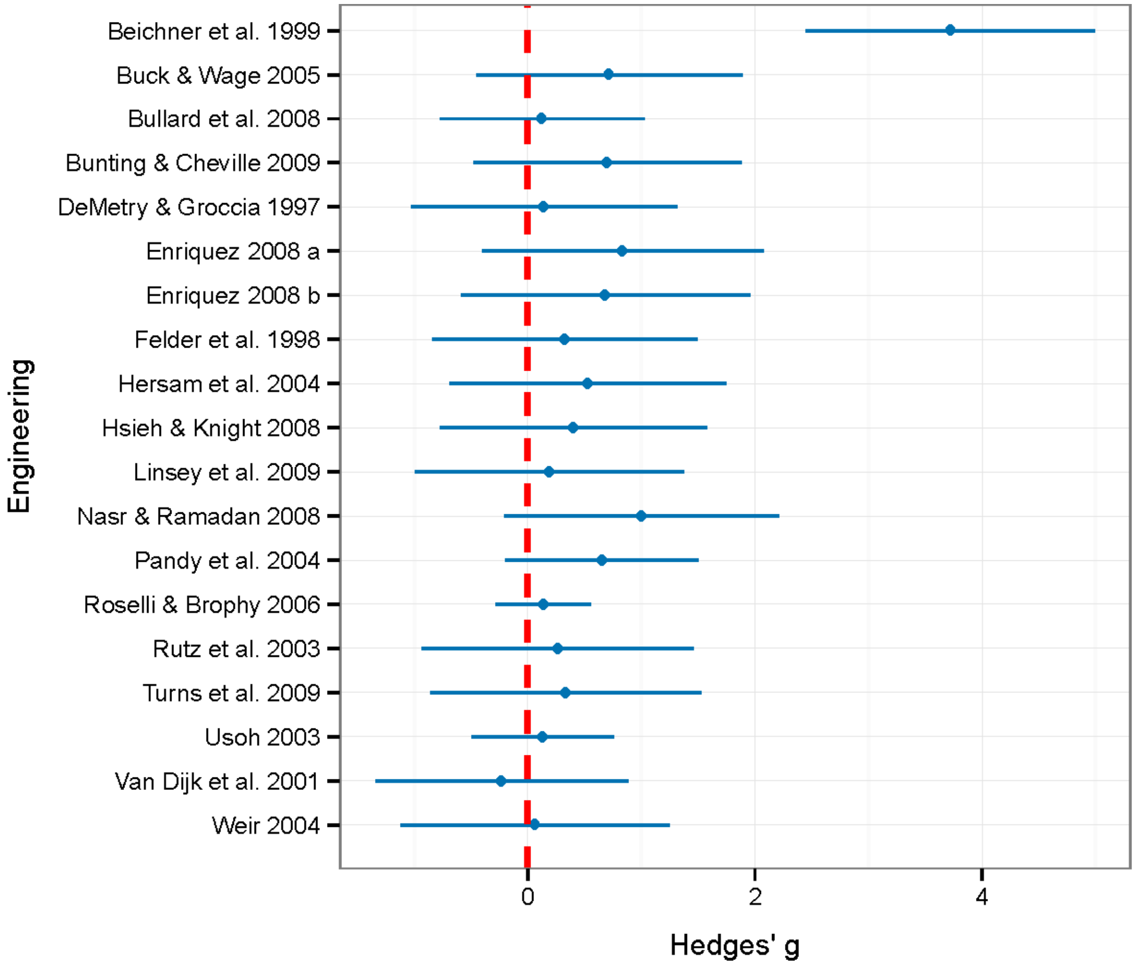
Table S1 (DOCX)
Table S2 (DOCX)
Table S3 (DOCX)
Table S4 (DOCX)

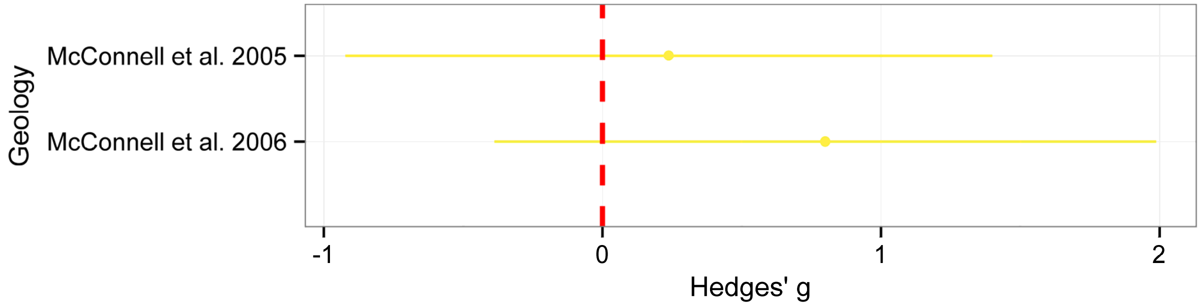

Proportion of students who fail the class
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Traditional Lecture
Active Learning

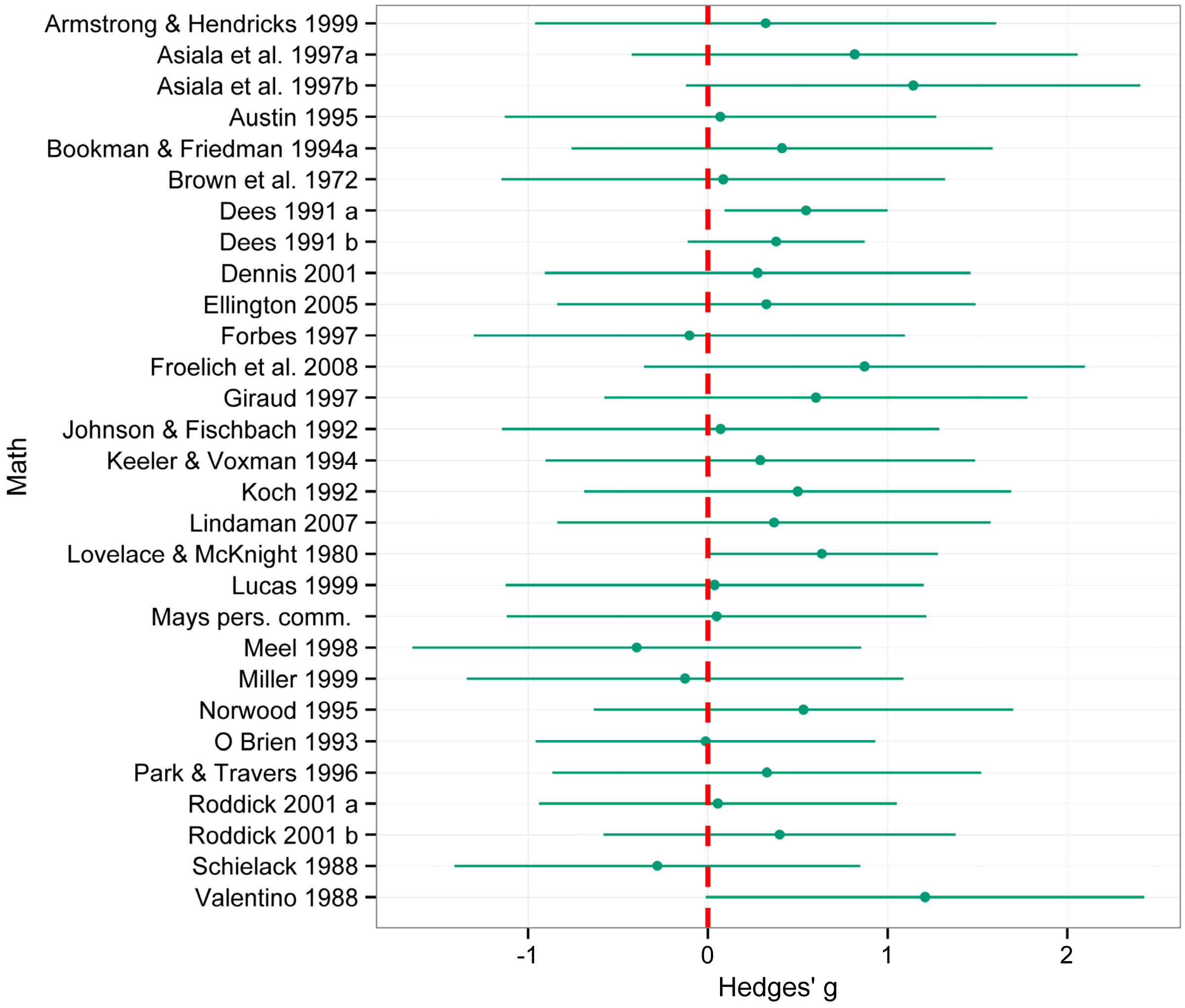

Armbruster et al. 2009
Armstrong et al. 2007
Barnard 1942
Burrowes 2003
Carmichael 2009
Chaplin 2009
Dollman 2005
Gifford & Vicks 1982
Haberyan 2003
Jensen & Finley 1996
Kitchen et al. 2003
Klionsky 2002
Knight & Wood 2005
Krawiec et al. 2005
Lohse et al. 2003
Lord 1997
Lord 1999
Marbach-Ad & Sokolove 2000
McCormick 2000
McDaniel et al. 2007
McFarlin 2008
Montelone et al. 2008
Nehm & Reilly 2007
Nelson et al. 2009
Paschal 2002
Randolph et al. 1992
Rybarczyk et al. 2007
Sadler 2002
Steele 2003
Tessier 2007
Udovic et al. 2002
Walker et al. 2009
Wilke & Straits 2001
Biology
-2
-1
0
1
2
Hedges' g

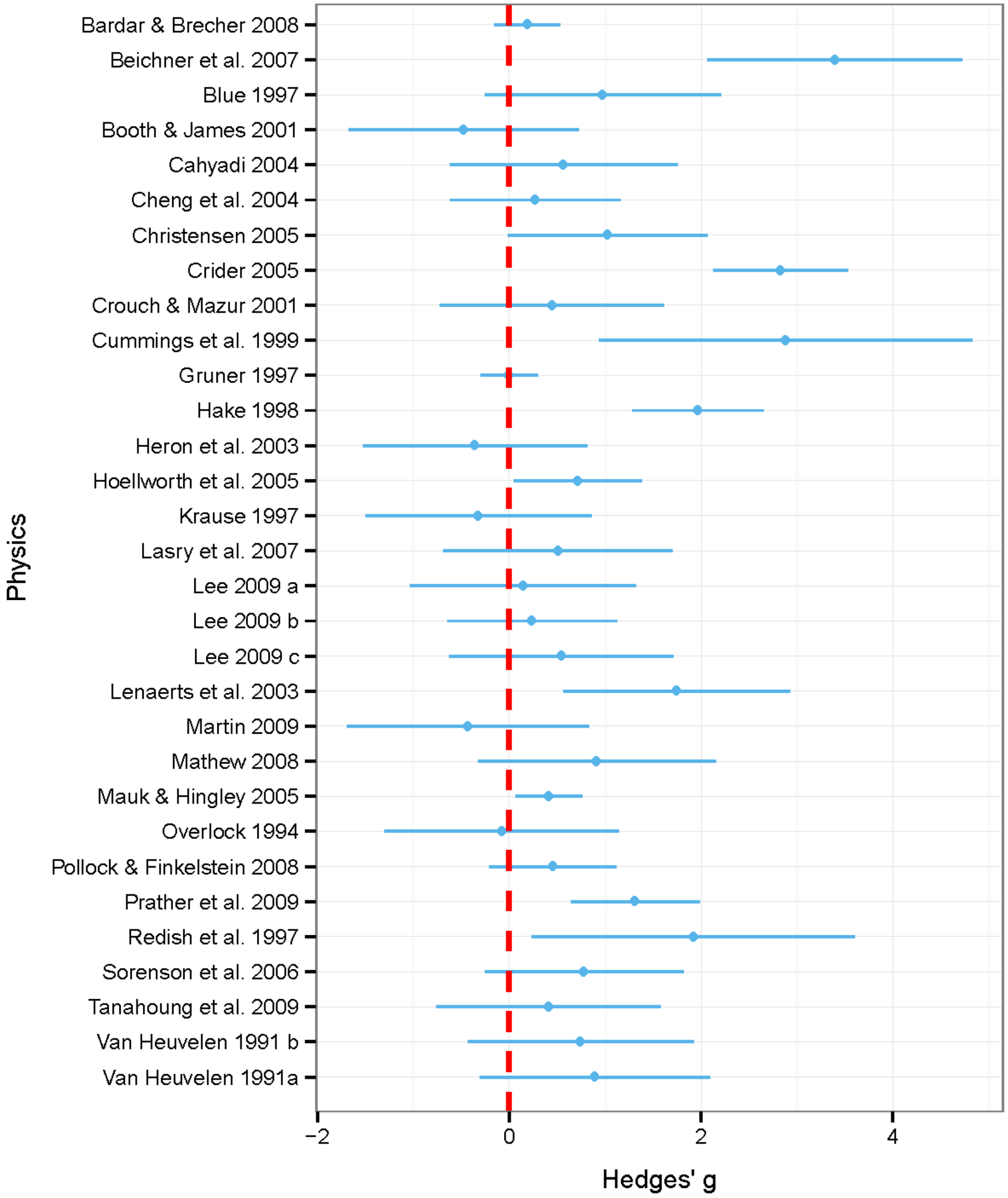

Chemistry
Anderson et al. 2005
Bannerjee & Vidyapati 1997
Basili & Sanford 1991
Bilgen 2006
Bilgin et al. 2009
Bradley et al. 2002
Dougherty et al. 1995
Doymus 2007
Gutwill-Wise 2001a
Gutwill-Wise 2001b
Hagen et al. 2000
Kurbanoglu et al. 2006
Lewis & Lewis 2008
Mohamed 2008
O Sullivan & Copper 2003
Oliver-Hoyo et al. 2004
Robinson & Niaz 1991
Ross & Fulton 1994
Shatila 2007
Tien et al. 2002
Williamson & Rowe 2002
Zimrot & Ashkenazi 2007
-1
0
1
2
3
Hedges' g

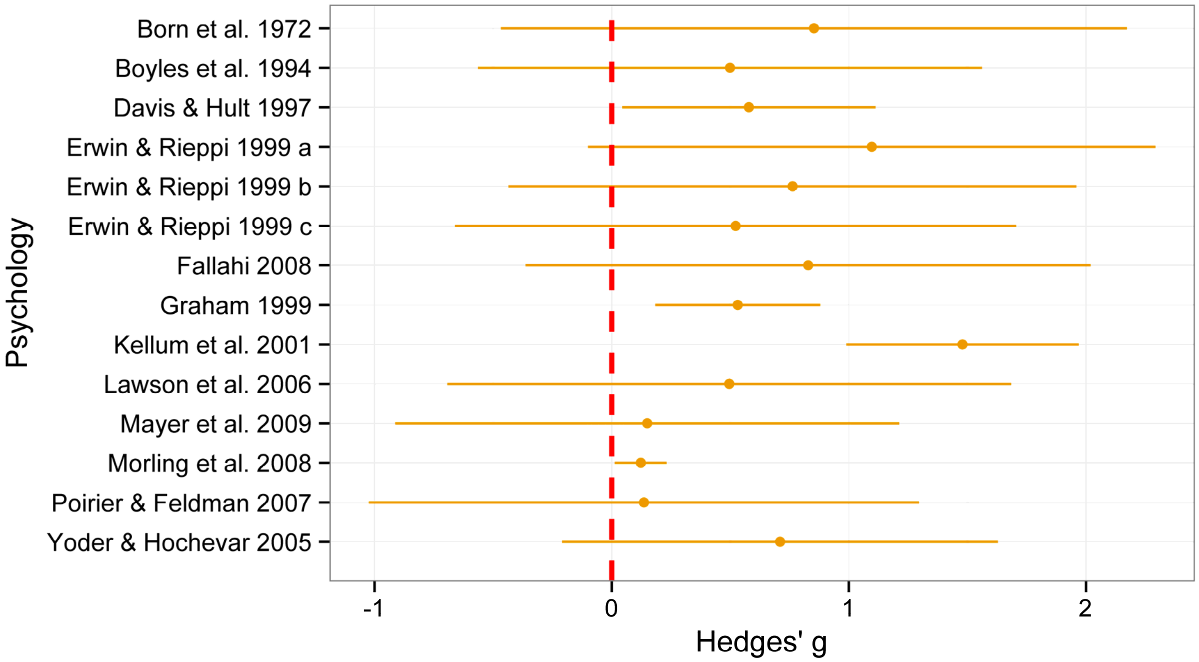

Computer Science
Barg et al. 2000
Beck & Chzhik 2008
Day & Foley 2006
Gonzalez 2006
Hurley 2002
LeJeaune 2002
Priebe 1977
Van Gorp & Grissom 2001
-2
-1
0
1
2
Hedges' g

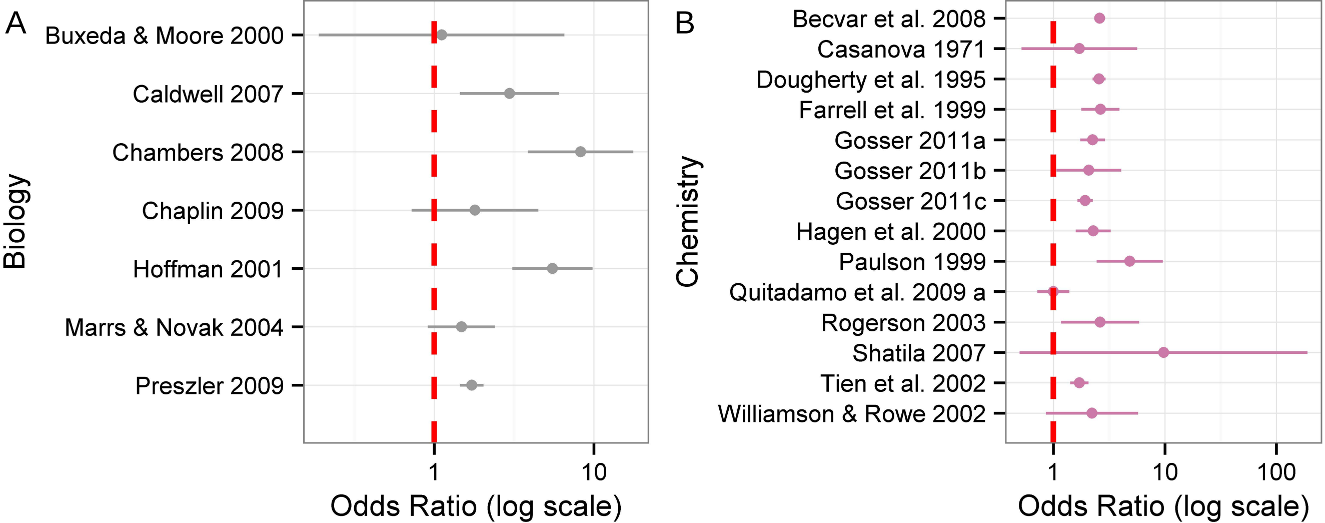

Beichner et al. 1999
Buck & Wage 2005
Bullard et al. 2008
Bunting & Cheville 2009
DeMetry & Groccia 1997
Enriquez 2008 a
Enriquez 2008 b
Felder et al. 1998
Hersam et al. 2004
Hsieh & Knight 2008
Linsey et al. 2009
Nasr & Ramadan 2008
Pandy et al. 2004
Roselli & Brophy 2006
Rutz et al. 2003
Turns et al. 2009
Usoh 2003
Van Dijk et al. 2001
Weir 2004
Engineering
0
2
4
Hedges' g

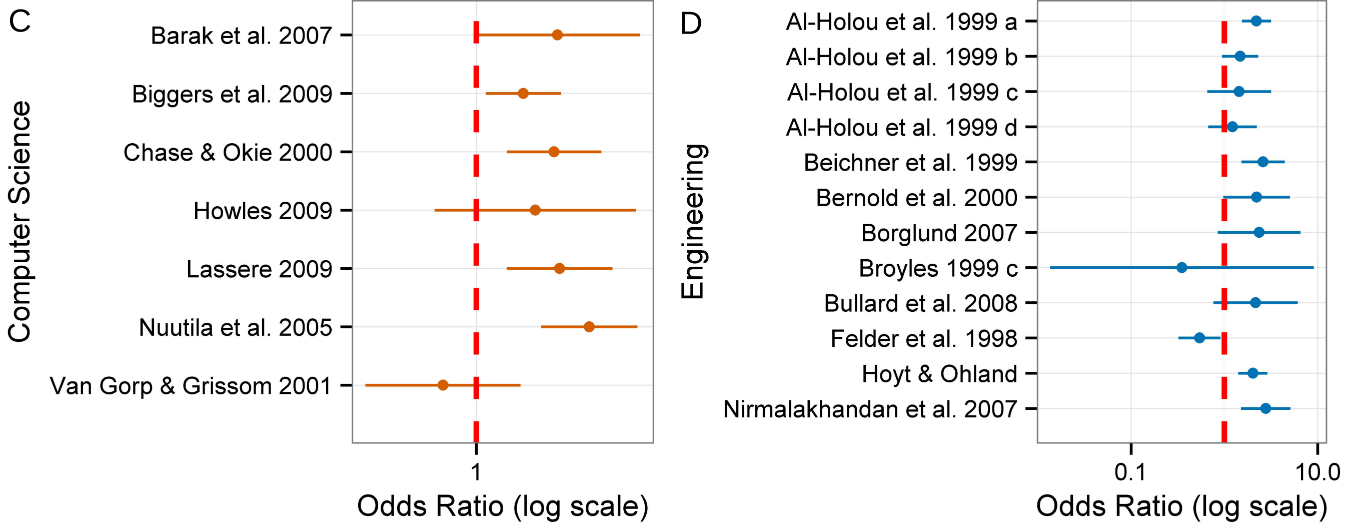

McConnell et al. 2005
McConnell et al. 2006
Geology
-1
0
1
2
Hedges' g

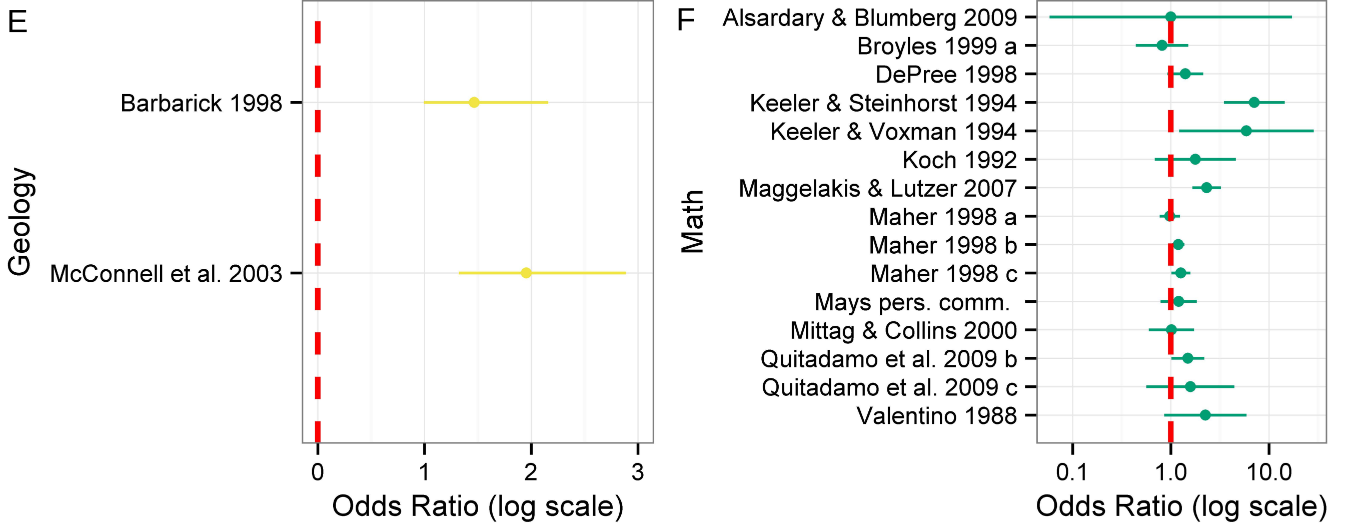

Armstrong & Hendricks 1999
Asiala et al. 1997a
Asiala et al. 1997b
Austin 1995
Bookman & Friedman 1994a
Brown et al. 1972
Dees 1991 a
Dees 1991 b
Dennis 2001
Ellington 2005
Forbes 1997
Froelich et al. 2008
Giraud 1997
Johnson & Fischbach 1992
Keeler & Voxman 1994
Koch 1992
Lindaman 2007
Lovelace & McKnight 1980
Lucas 1999
Mays pers. comm.
Meel 1998
Miller 1999
Norwood 1995
O Brien 1993
Park & Travers 1996
Roddick 2001 a
Roddick 2001 b
Schielack 1988
Valentino 1988
Math
-1
0
1
2
Hedges' g

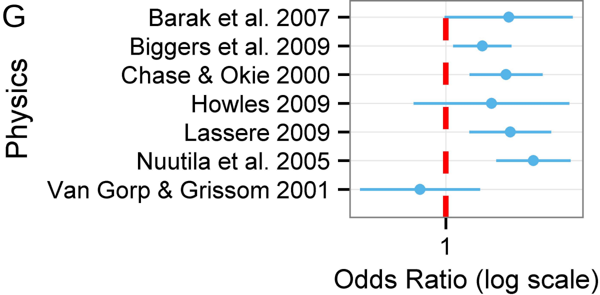

Physics
Bardar & Brecher 2008
Beichner et al. 2007
Blue 1997
Booth & James 2001
Cahyadi 2004
Cheng et al. 2004
Christensen 2005
Crider 2005
Crouch & Mazur 2001
Cummings et al. 1999
Gruner 1997
Hake 1998
Heron et al. 2003
Hoellworth et al. 2005
Krause 1997
Lasry et al. 2007
Lee 2009 a
Lee 2009 b
Lee 2009 c
Lenaerts et al. 2003
Martin 2009
Mathew 2008
Mauk & Hingley 2005
Overlock 1994
Pollock & Finkelstein 2008
Prather et al. 2009
Redish et al. 1997
Sorenson et al. 2006
Tanahoung et al. 2009
Van Heuvelen 1991 b
Van Heuvelen 1991a
−2
0
2
4
Hedges' g

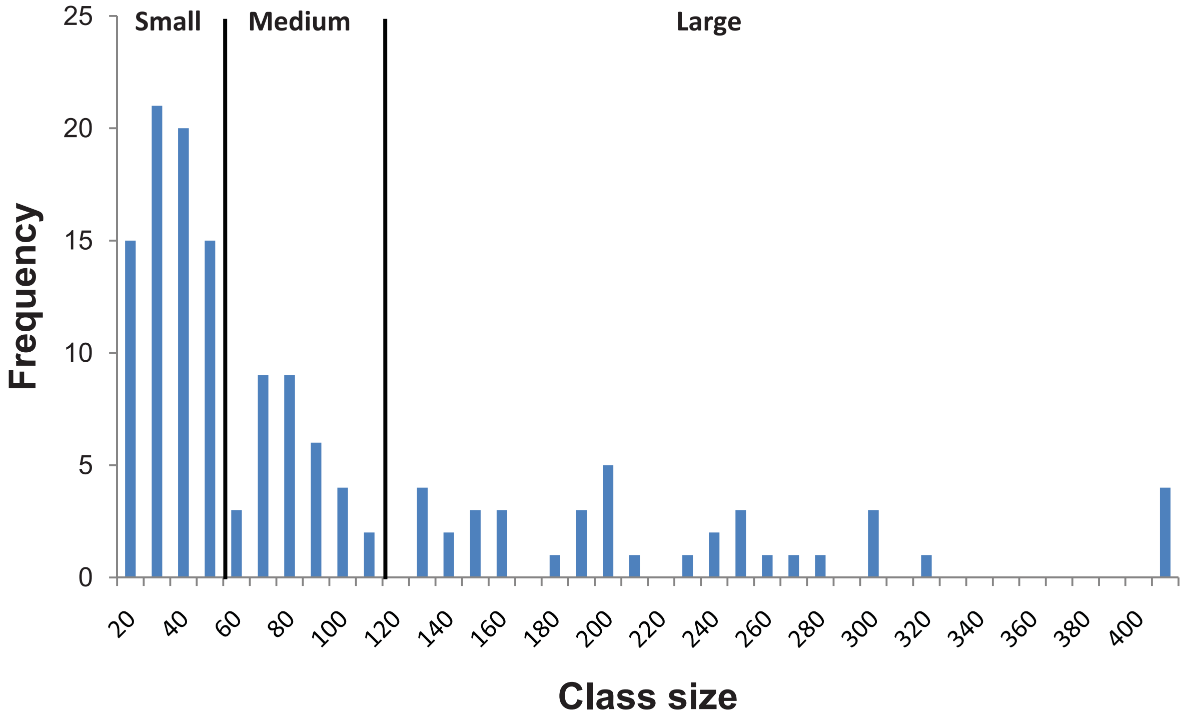

Psychology
Born et al. 1972
Boyles et al. 1994
Davis & Hult 1997
Erwin & Rieppi 1999 a
Erwin & Rieppi 1999 b
Erwin & Rieppi 1999 c
Fallahi 2008
Graham 1999
Kellum et al. 2001
Lawson et al. 2006
Mayer et al. 2009
Morling et al. 2008
Poirier & Feldman 2007
Yoder & Hochevar 2005
-1
0
1
2
Hedges' g

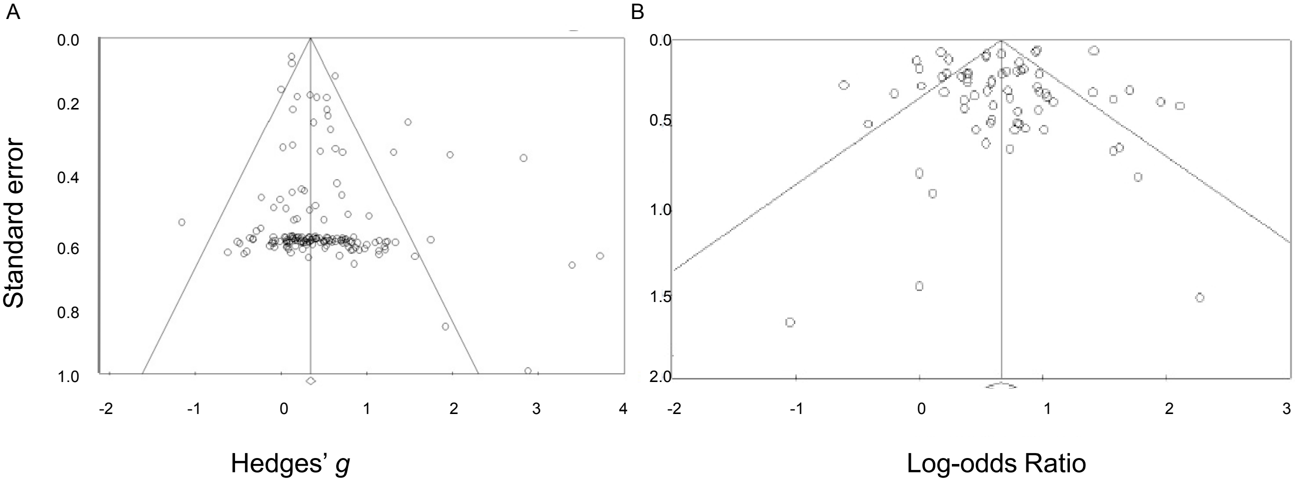

A
Buxeda & Moore 2000
Caldwell 2007
Chambers 2008
Chaplin 2009
Hoffman 2001
Marrs & Novak 2004
Preszler 2009
Biology
1
10
Odds Ratio (log scale)
B
Becvar et al. 2008
Casanova 1971
Dougherty et al. 1995
Farrell et al. 1999
Gosser 2011a
Gosser 2011b
Gosser 2011c
Hagen et al. 2000
Paulson 1999
Quitadamo et al. 2009 a
Rogerson 2003
Shatila 2007
Tien et al. 2002
Williamson & Rowe 2002
Chemistry
1
10
100
Odds Ratio (log scale)


C
Computer Science
Barak et al. 2007
Biggers et al. 2009
Chase & Okie 2000
Howles 2009
Lassere 2009
Nuutila et al. 2005
Van Gorp & Grissom 2001
1
Odds Ratio (log scale)
D
Engineering
Al-Holou et al. 1999 a
Al-Holou et al. 1999 b
Al-Holou et al. 1999 c
Al-Holou et al. 1999 d
Beichner et al. 1999
Bernold et al. 2000
Borglund 2007
Broyles 1999 c
Bullard et al. 2008
Felder et al. 1998
Hoyt & Ohland
Nirmalakhandan et al. 2007
0.1
10.0
Odds Ratio (log scale)


E
Geology
Barbarick 1998
McConnell et al. 2003
0
1
2
3
Odds Ratio (log scale)
F
Math
Alsardary & Blumberg 2009
Broyles 1999 a
DePree 1998
Keeler & Steinhorst 1994
Keeler & Voxman 1994
Koch 1992
Maggelakis & Lutzer 2007
Maher 1998 a
Maher 1998 b
Maher 1998 c
Mays pers. comm.
Mittag & Collins 2000
Quitadamo et al. 2009 b
Quitadamo et al. 2009 c
Valentino 1988
0.1
1.0
10.0
Odds Ratio (log scale)


G
Physics
Barak et al. 2007
Biggers et al. 2009
Chase & Okie 2000
Howles 2009
Lassere 2009
Nuutila et al. 2005
Van Gorp & Grissom 2001
1
Odds Ratio (log scale)


Small
Medium
Large
Frequency
25
20
15
10
5
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
320
340
360
380
400
Class size


A
B
Standard error
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
3
4
Hedges' *g*
0.0
0.5
1.0
1.5
2.0
-2
-1
0
1
2
3
Log-odds Ratio