The authors assessed the impact of three designs (randomized experiment, nonequivalent control group design, regression discontinuity design) on estimates of effect size of a university-level freshman remedial writing program. Designs were implemented within the same context, same time frame, and with the same population. The 375 freshman participants were either randomly assigned or self-selected into specific evaluation groups, according to design protocols. The three designs led to highly similar effect size estimates of the impact of a semester of remedial writing on writing outcomes following standard freshman composition. Specific design features contributed to the convergence of effect size estimates across designs.

COMPARISON OF A RANDOMIZED AND TWO QUASI-EXPERIMENTAL DESIGNS IN A SINGLE OUTCOME EVALUATION

Efficacy of a University-Level Remedial Writing Program

LEONA S. AIKEN
STEPHEN G. WEST
DAVID E. SCHWALM
Arizona State University

JAMES L. CARROLL
VIA, Incorporated

SHENGHWA HSIUNG
Arizona State University

The purposes of this research are both methodological and substantive. From a methodological perspective, we compare effect size estimates resulting from the implementation of a randomized experiment versus two different

AUTHORS' NOTE: *This research was supported by the College of Liberal Arts and Sciences and the Office of the Provost of Arizona State University. We acknowledge the cooperation of the Office of the Vice President for Student Affairs and the assistance of the Office of the Registrar (Louise Ann Denny), the Freshman Composition Program (Barbara Metcalf), the Undergraduate Admissions Office (Carla Cassity, Robert Francis, Jane Olsen, Timothy Desch, Robert Hancock,*
EVALUATION REVIEW, Vol. 22 No. 2, April 1998 207-244

quasi-experiments within a single evaluation context. In addition, we illustrate the use of design controls to render clear interpretations of an outcome evaluation in an institutional setting in the face of threats to both internal and external validity, as defined by Cook and Campbell (1979). From a substantive perspective, we evaluate the efficacy of a one-semester remedial English course of a large university. The remedial course was designed to ameliorate writing skill deficits of entering students. The remedial training provided by such courses is expected to prepare students for the writing demands of university-level courses.

There are no published accounts of the impact of a one-semester remedial writing course on college-level writing skills. Two sets of studies, both unpublished, examined the impact of placement testing and remediation on college retention. As summarized by White (1994, 1995), the California State University English Placement Study and the New Jersey Basic Skills Council Study both showed that a placement program followed by intensive basic skills training (for New Jersey, training in reading, writing, and mathematics) led to increased college retention. There are interpretational complexities associated with these studies in that students self-selected to participate in the programs. Other evaluation research on writing program impact has focused on standard, rather than remedial, college-level composition. The two historically largest projects are that of Faigley and Witte (Faigley et al. 1985; Witte and Faigley 1983) and that of White and Polin (1986).

Effect Sizes as a Function of Research Design

In terms of internal validity, the sine qua non of experimental design is the randomized experiment with random assignment of subjects to treatment conditions. Yet, applied settings often pose constraints that prevent the use of randomized experimental designs. Policy may dictate that individuals must decide whether they require special instruction (i.e., self-selection into treatment). Alternatively, policy may dictate that access to special instruction is determined by need. Ideally, need is determined by a formal assessment, with individuals on one side of a cut score having access to special instruction, whereas those on the other side are not eligible for such instruction.

---

and Nancy Hall), and the Office of Institutional Analysis (Nelle Moore). We especially acknowledge the efforts of Jane Hawthorne in field management and Virgil Sheets in data management. We acknowledge Professor Keith Miller for his role in evaluation of the writing samples, along with Jacqueline Wheeler and Allene Cooper. We thank the faculty of the many sections of both remedial and standard freshman composition who cooperated with our evaluation protocol. *Correspondence should be addressed to Leona S. Aiken, Department of Psychology, Arizona State University, Box 871104, Tempe, AZ 85287-1104; e-mail: iaclsa@asuvm.inre.asu.edu.*

An important question from the perspective of outcome evaluation is the impact of the investigator's choice of research design on estimates of treatment effect sizes. Dramatic illustrations of the differences between estimates of the effectiveness of treatments obtained from experimental and quasi-experimental designs do exist in the literature. For example, LaLonde (1986) assessed the increases in earnings that resulted from a training program that provided work experience and job counseling relative to a no-treatment control group using both experimental and econometric quasi-experimental methods. For female participants, the estimates from the quasi-experiment were substantially larger and positive, whereas for male participants, the estimates were smaller and negative relative to those from the randomized experiment. Similarly, Miao (1977) reported that a medical innovation, gastric freezing of ulcers, showed positive results in nonequivalent control group designs but failed to show similar positive benefits in carefully conducted randomized trials. These intriguing results illustrate the potential importance of assessments of the impact of design choice on estimates of treatment effects. We distinguish here between two classes of such assessments. The first and more common class compares effect sizes taken from different studies that examine the same general question but that use different designs (Between-study comparisons). The second and more rare class provides effect size estimates from different designs implemented within a single study in a single evaluation context (Within-study comparisons).

Between-study comparisons. Between-study comparisons have most often contrasted effect size estimates across two particular designs: (a) the true experiment with random assignment to treatment versus control conditions and (b) the nonequivalent control group design (Cook and Campbell 1979) in which assignment to treatment and comparison groups is on an unknown basis that is presumed to be nonrandom. Heinsman and Shadish (1996) and Shadish and Heinsman (1997), in their reviews of 98 studies in four substantive areas, reported that whether the randomized or the quasi-experimental design yields larger effect sizes appeared to depend on the substantive area, with an average advantage for randomized experiments. Lipsey and Wilson (1993) also reported variation across substantive areas in whether larger effect sizes obtained from true versus quasi-experiments; however, across 302 meta-analyses of educational and behavioral interventions, the average effect sizes of the two types of designs were essentially equal. Shadish and Heinsman (1997) summarized earlier contrasts between designs. In two earlier reviews of the literature evaluating medical innovations, randomized experiments were found to yield smaller effect size estimates than quasi-experiments (Colditz, Miller, and Mosteller 1988; Gilbert, McPeek, and Mosteller 1978). In contrast, randomized experiments evaluat-

ing the effect of coaching on SAT scores yielded larger effect sizes than quasi-experiments (Becker 1990). Smith, Glass, and Miller (1980) found no appreciable effect of design choice on effect size estimates in psychotherapy outcome evaluations, a null finding replicated by Shapiro and Shapiro (1982) and Hazelrigg, Cooper, and Borduin (1987). However, Shadish and Ragsdale (1996) reported an effect size advantage for randomized evaluations of marital psychotherapy and enrichment interventions.

Shadish and his coworkers (Heinsman and Shadish 1996; Shadish and Heinsman 1997; Shadish and Ragsdale 1996) argued that important methodological characteristics might covary with design type (true versus quasi-experiment). These characteristics, they argued, could account for discrepancies reported between effect size estimates from randomized versus quasi-experiments. These characteristics included the following: (a) the use of internal versus external comparison groups (i.e., from the same or a different population from the control group), (b) the use of blocking or matching in subject assignment, (c) self-selection versus other-selection of cases into conditions, (d) activity level in the comparison group (untreated versus some alternative treatment), (e) level of total and differential attrition, (f) control for pretest effect size (i.e., pretest differences between groups), and (g) the use of exact versus approximate effect size estimates (Ray and Shadish 1996). Heinsman and Shadish (1996) showed that much, but not all, of the observed difference in effect size estimates from true versus quasi-experiments could be accounted for by these design characteristics. They further showed that these design factors covaried with substantive area and could account for the variation in discrepancy between effect sizes from true versus quasi-experiments across these areas. In sum, they argued that previous literature may have attributed to design type (true versus quasi-experiment) differences in effect sizes more accurately attributable to other methodological characteristics of the investigations. Shadish and Ragsdale (1996) replicated the work of Heinsman and Shadish (1996) for marital therapy and enrichment programs, showing that approximately half the effect size advantage of randomized over quasi-experiments could be accounted for by such design factors.

Within-study comparisons. A variety of design characteristics in addition to those identified by Heinsman and Shadish (1996) may also influence effect sizes, among them the strength of particular interventions and the integrity of their implementation, the adequacy of coverage of constructs by measurement instruments, and the reliability of measurement (Cooper and Richardson 1986; Sechrest et al. 1979). Comparisons of design type (true versus quasi-experiments) within a single evaluation context should provide a more accurate picture of the impact of design type on effect size estimates, in that

other design and measurement characteristics would be held constant. In fact, Dennis and Boruch (1989) argued that adding quasi-experimental components to a fully randomized experiment would provide insight into the conditions under which effect size estimates from the two types of experiments would converge.

Within-study design comparisons are rare. Dennis and Boruch's (1989) review of several such experiments suggested that effect size estimates may vary widely over design types within a single setting. For example, substantially larger effect size estimates were gleaned from a pretest-posttest comparison in a treated group than from a between-group comparison of treated versus control subjects (Gomez 1985). Larger effect sizes were reported in the true experimental design over the regression discontinuity design in the evaluation of the Salk polio vaccine (Meier 1985).

Evaluation: Setting and Questions

Large-scale institutions, such as major universities, afford a unique opportunity for the implementation of multiple designs to evaluate program interventions. We illustrate the use of three evaluation designs in the evaluation of a remedial English program targeted toward students who matriculate into the university with inadequate writing skills. As at many universities, selection into the one-semester remedial writing class has been traditionally based on scores on standardized admissions tests. The remedial class was required for students with ACT English scores of 16 or below or SAT Verbal scores of 380 or below.[1] We examined two questions for those students who needed remediation:

1. Does the remedial course or the standard freshman composition course lead to better writing skills at the end of just one semester of composition training?
2. Does the sequence of remedial writing in the fall semester followed by standard freshman composition in the spring semester lead to improved writing relative to the standard freshman composition class alone?

Evaluation Designs

We implemented three classes of designs described in methodology texts and chapters (e.g., Boruch 1997; Cook and Campbell 1979; Judd and Kenny 1981; Reichardt and Mark 1997; Shadish, Cook, and Campbell 1997; West, Biesanz, and Pitts in press) that are relatively commonly used in the educational evaluation area. Design A, our basic design, was a randomized experi-

ment. Design B was a nonequivalent control group design. Design C was the regression discontinuity design (Trochim 1984) in which students are assigned to treatment on the basis of quantitative measure of need, here each student's SAT Verbal or ACT English test score. The three types of designs share overlapping groups of subjects so that our effect size estimates will not be independent. All design components are summarized in detail in Table 1; the complete schedule of testing and treatment for each group in the design is provided.

Design A: Randomized experiment. The base design for evaluation was a "tie-breaking" experiment (Boruch 1975; Campbell 1984), with random assignment to conditions within a fixed range of scores. Participants were students whose admissions test scores (ACT English, SAT Verbal) fell within a fixed range just below the normal cutoff scores for taking standard freshman composition. These students were offered participation in a lottery wherein they might be exempted from the remedial English requirement. According to the conditions of the lottery, those who were not exempted took remedial English in the fall semester followed by standard freshman composition in the immediately following spring semester (*remediation* group). Those who were exempted took standard freshman English composition in the fall semester (*nonremediation* group). Comparison of these two groups provided an estimate of treatment effect sizes from a randomized experiment.

Design B: Nonequivalent control group design. One additional matched comparison group of remedial English students naturally occurred during the fall semester. This group had admissions test scores in exactly the same range as participants in the tie-breaking experiment. We term this group the *accretion* group because they registered late or could not be contacted during the recruitment for the randomized experiment. The accretion group took remedial English in the fall semester and standard composition during the spring semester. In the nonequivalent control group design, the performance of the accretion group was compared to that of the nonremediation group from the randomized experiment to generate an additional estimate of effect size.

Design C: Regression discontinuity design. Because assignment to the remedial English course was normally based on standardized test scores, we were also able to implement a regression discontinuity design (Berk and Rauma 1983; Shadish, Cook, and Campbell 1997, chap. 5; Trochim 1984, 1990). The design involved two types of students: (a) those students whose placement test scores fell below the English Department's normal cut point and who, therefore, were required to take remedial writing followed by standard freshman composition (remediation plus accretion groups) and (b) those students whose placement test scores fell above the cut point and

TABLE 1: Experimental Groups, Course and Testing Schedule, and Available Cases for All Experimental Groups

Design/Groups	*Fall Pretest*	*Fall Course*	*Fall Posttest*	*Spring Course*	*Spring Posttest*	*Total Entering Evaluation*	*Total Passing Fall Course*	*Total Passing Spring Course*
A. Randomized experiment								
Remediation	Yes	Remedial	Yes	Standard	Yes	53	39[a]	34[c]
Nonremediation	Yes	Standard	Yes	—	—	76	68	—
B. Nonequivalent control group design								
Accretion (Nonremediation)	Yes	Remedial	Yes	Standard	Yes	75	60	38[c]
C. Regression discontinuity design								
Standard (Remediation, accretion)	Yes	Standard	Yes	—	—	141	119	—
Control for cyclic variation								
Spring (Standard)	No	—	—	Standard	Yes[b]	30	—	26

a. In all, only 48 of the 53 students originally recruited, who attended the university, actually took the remedial course, and 39 of these 48 passed the course.
b. Students in the spring group were posttested only at the end of spring semester.
c. All 39 remediation students entered the spring course, but only 44 of the 60 accretion students did so.

who, thus, took only standard freshman composition (*standard group*). The logic of this design is illustrated in Figures 1 and 2, in which hypothetical performance at the end of the standard composition course is plotted as a function of a placement test score, the ACT English score. Figure 1 illustrates the regression of writing outcomes on the ACT English score in the absence of any effect of the remedial English course. Figure 2 illustrates a hypothetical outcome in which all students below the cutoff score initially receive an effective remedial English course and then all students take standard freshman composition. In Figure 2, the vertical line represents the cutoff score. In the region below the cut score in which all students receive remediation, the regression line for the prediction of outcomes from admissions test scores is shifted upward by a constant amount. This shift represents the gain in performance attributable to the treatment, here the remedial English course.

Control for Cyclic Institutional Variation

Our evaluation involved comparison of writing outcomes across students who completed standard freshman composition during the fall (nonremediation, standard groups) versus spring semester (remediation, accretion groups). One plausible alternative explanation for any observed differences is that eligible students who selected standard freshman composition during their first (fall) semester at the university differed from those who deferred the course until the second (spring) semester of their freshman year. Differences in student capability in English across semesters might have led to differential instruction, grading, or both. A group of students who qualified for standard freshman composition through their standardized test scores who took the course during the fall semester (standard group) were compared with similar students who took the course during the spring semester (spring group). This comparison served as a test of institutional variation across semesters.

METHOD

Participants

Participants were drawn from entering freshman at a large university during 1988-1989.[2] They included students who met the requirement for standard freshman composition and selected to enroll in this course during the fall or spring semester. Participants also included students for whom

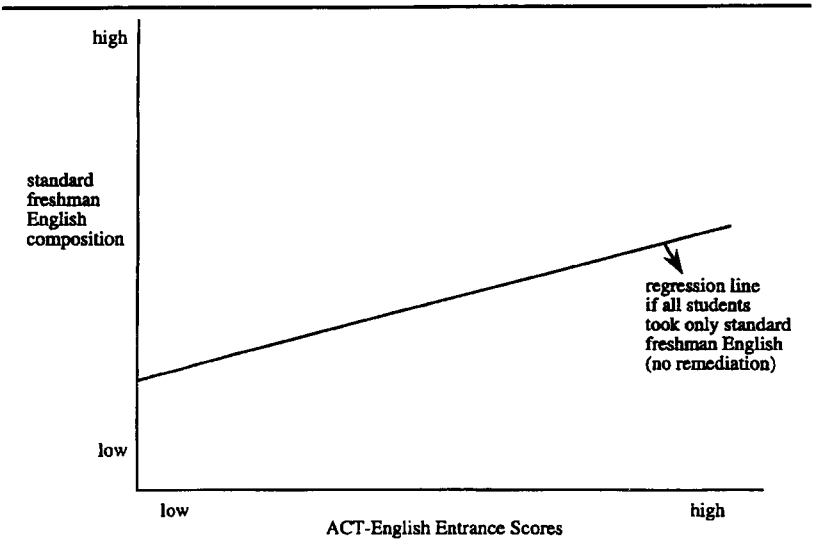

Figure 1: English Competency at the End of Standard Freshman English if There Were no Remedial English Course

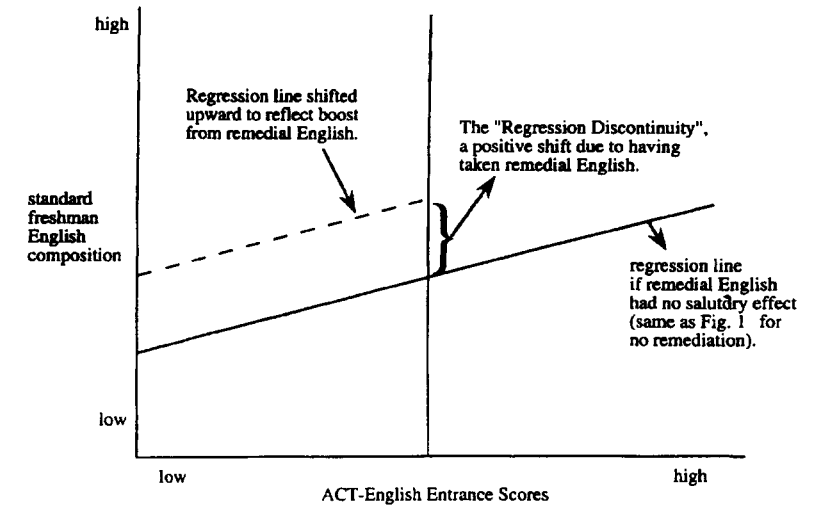

Figure 2: English Competency at the End of Standard Freshman English if Students With ACT-English Scores ≤ 16 Took Remedial English Before Standard Freshman English

TABLE 2: Demographic Characteristics and Placement Test Scores of Freshman Study Participants at Point of Recruitment

Characteristic	*Group* *Remediation*	*Nonremediation*	*Accretion*	*Standard*	*Spring*
Sample size	53	76	75	141	30
Percentage male	55%	42%	57%	57%	60%
Percentage Caucasian	76%	88%	60%	92%	87%
Age in years					
M	17.9	17.9	18.4	18.2	18.7
SD	.5	.5	1.7	2.2	1.9
SAT verbal					
M	357.6[a]	364.0	347.8[a]	462.0	480.0
SD	32.7	23.1	32.9	65.0	93.8
n	33	48	36	105	13
ACT English					
M	14.9	14.7	14.7	21.2	20.5
SD	1.2	1.8	1.6	2.6	2.8
n	34	42	61	78	21

a. Students who took both the ACT and SAT qualified for the study if either test fell within the criterion range. Some students who qualified based on the ACT had SAT scores below the SAT cutoff of 360, accounting for the SAT means below 360.

remedial writing was required who either were enrolled into the randomized experiment or who selected to register for remedial writing in the fall semester. The age, gender, and ethnic composition of all participants, stratified by experimental group, are provided in Table 2.

Placement Test Scores

Students whose ACT English scores were less than or equal to 16 or whose SAT Verbal scores were less than or equal to 380 have traditionally been required to take remedial English composition at this university. In that the randomized experiment involved exempting some students from this requirement, participation in the randomized experiment was limited to those whose ACT English scores fell between 12 and 16, or whose SAT verbal scores fell between 360 and 380. Students scoring below the lower cutoffs were considered in the absence of remediation to be at too great a risk for failure in standard freshman English composition to participate in the randomized experiment.

University policy permits students to take either the ACT or the SAT for admissions; some students reported scores on both tests. Scores on the SAT

Verbal and ACT English subtests are reported by the experimental group in Table 2. For those students who took both admissions tests, both their SAT and ACT scores are included in Table 2.

Recruitment Procedures

Subject recruitment procedures varied by experimental design. All students were assured complete confidentiality. Students were accurately informed that their composition instructors would be kept blind to their participation in the evaluation. Recruitment of each experimental group is described below.

Design A: Randomized experiment. Students who had been admitted into the university and whose placement test scores were in the remedial target range were recruited through nine different registration mechanisms (off-site, on-site, mail-in) during a 4-month period prior to the fall semester. Students were informed of the study and their option to participate in a lottery. By lottery, they might be permitted to take standard freshman English composition rather than the required remedial English composition during the fall semester. To participate in the lottery, they agreed *either* to take remedial composition followed by standard freshman composition during their freshman year *or* to take standard freshman composition in the fall semester of their freshman year, depending on the outcome of the lottery. All participants in the lottery were guaranteed admission to these courses as an inducement for participation and to ensure the integrity of the design.[3] Those who agreed to participate in the lottery were randomly assigned to the remediation (remedial English followed by standard freshman English composition) versus nonremediation (standard freshman English composition only) group; written parental permission was gathered from all students younger than 18 years of age. To ensure that no particular section of standard freshman composition contained an excess of remedial students who might slow class progress, no more than one nonremediation student was placed into any single section of standard freshman composition. Just before the outset of the fall semester, all participants in the randomized experiment were traced to their housing at the university. In all, 53 remediation and 76 nonremediation students were initially recruited, and 48 and 75, respectively, entered the appropriate composition course in the fall semester.

Attempts were made to contact a total of 363 newly admitted freshman students during the recruitment period. This group included all students eligible for participation in the randomized experiment who participated in the preregistration process for fall matriculation on campus or at any of the

remote sites in the United States. This process takes place during a 6-month period and closes at the end of July. The 363 students all had admissions test scores that fell within the target range of the study. Of these, 36% agreed to participate, 48% refused participation, and 16% could not be reached. Of those who refused, 35% had decided not to attend the university, 20% had exercised a university-provided option to retake the ACT English test and had improved their scores sufficiently to be no longer eligible for the study, 11% intended to retake the placement test, 14% had taken standard freshman composition at another college, and 16% wished to take remedial English composition.

Design B: Nonequivalent comparison group. The comparison group (accretion group) for the nonequivalent control group design was recruited from later registrants who had not been contacted for participation in the randomized experiment. At the end of the second week of the fall semester, students who had self-selected into the remedial English composition class and whose placement test scores were in the target range of the evaluation (i.e., ACT English of 12-16, or SAT Verbal of 360-380) were recruited. To participate, these students agreed to take standard freshman English composition during the spring semester and were guaranteed enrollment into this class upon passing the remedial English composition class. In all, 76 students agreed to participate, forming the accretion group. Though these students were enrolled into the study after the outset of the semester, pretest data were nonetheless available for them, because all fall remedial classes had received the pretest writing assessments as part of mass in-class testing at the outset of the semester. In the nonequivalent control group design, the accretion group was compared with the nonremediation group discussed above.

Design C: Regression discontinuity design. Students whose placement scores made them eligible for the standard freshman English composition class and who enrolled in this class during the fall semester were eligible for participation in the standard group of the regression discontinuity design. In lieu of recruiting individual students, a random sample of 12 of the 117 sections of standard freshman English composition in the fall semester, stratified across all times and days of instruction, was chosen for data collection. A one-in-three random sample ($n = 141$) of these freshmen English composition students was selected for the regression discontinuity design (standard group). The regression discontinuity design included the standard group and the remediation and accretion groups described above.

Control for cyclic institutional variation. A stratified random sample of five sections of standard freshman English composition was also identified at the end of the spring semester. We assessed a one-in-three sample ($n = 30$) of freshmen (spring group) in these sections. Comparison of the spring group

with the standard group provided a test for possible cyclic institutional variation in the classes from fall to spring.

Writing Courses

Remedial writing. The remedial writing course focused on preparing students for standard freshman English composition. It addressed three related areas: (a) paragraph and essay structure and development, (b) sentence structure, and (c) grammar usage and mechanics. Personal experience served as subject matter for writing assignments. Students were taught to write, edit, and revise coherent and unified paragraphs and essays. Prepared writing was stressed over extemporaneous writing, with five of the seven papers for the course written outside class. Persistent problems in sentence structure and grammar were addressed through revision of papers and in supplemental writing exercises.

Standard freshman composition. The freshman composition course focused on critical reading and writing, with emphasis on strategies of academic discourse. Information drawn from reading or observation served as the subject matter for writing, in contrast to writing from personal experience in the remedial course. Through a process of writing, revising, and editing nine multiparagraph writing assignments, students were taught to incorporate accurately new information drawn from reading or observation into focused and coherent essays. Because the assignments were reading-based or data-based, the course also stressed critical reading strategies. There was a shift from a stress on correctness in the remedial course to rhetorical effectiveness in standard freshman composition.

Measurement of Writing Outcomes

Two measures of writing competency were employed. Using the terminology from the writing competency area, the first was a *direct assessment* in the form of a written essay. The second was an *indirect assessment*, in the form of an objective test of knowledge of written expression. It has been argued that both these methods should be employed in the evaluations of writing competency (Breland et al. 1987; Breland and Gaynor 1979; White 1994).

Direct assessment. The direct assessment was an essay written in response to one of three prompts; each prompt was a table of data. For example, one table provided life expectancies at birth in the United States from 1900 to 1965 stratified by gender and ethnicity. Students were instructed to complete

two tasks in their essays: (a) to provide a general description of the information provided in the table and (b) to summarize trends in the data. The prompts were furnished by the director of the freshman English program.[4] These prompts were of the type typically employed in standard freshman English composition; thus, the writing sample tapped those skills that students completing standard freshman composition should have acquired. The three prompts were used in equal proportions at all three test administrations. In addition, each student received a different prompt at each administration.

Indirect assessment. The Test of Standard Written English (TSWE, College Entrance Examination Board) is a 50-item multiple choice test that assesses ability to use conventional standard written English as found in college text books. The range of skills assessed in the TSWE ranges from subject-verb agreement to "the logic of comparisons and the appropriate subordination or coordination of ideas within a sentence" (Breland and Gaynor 1979, 122). Only a few items test punctuation, and none tests knowledge of grammar terminology, spelling, or capitalization.

Testing Schedule and Location

The assessment schedule is summarized in Table 1. All students enrolled in remedial courses (remediation, accretion) in the fall semester, and the 12 randomly selected sections of standard freshman composition (standard) received the TSWE and writing sample as part of in-class testing at the outset and end of the fall semester. Nonremediation students were administered the fall semester pretests outside class. Posttest writing samples were administered in class, because the writing prompts were incorporated into the final examination. Remediation and accretion students were also assessed outside class at the end of the spring semester following completion of the standard freshman English composition course. Finally, the writing samples and TSWEs were administered in class at the end of the spring semester to the five randomly selected sections of standard freshman English composition (spring group). Students tested outside class were offered monetary incentives to minimize attrition. In all instances, students were given 45 minutes to complete the writing sample. The TSWE administration followed standard procedures of the College Entrance Examination Board.

Scoring of Writing Samples

Judges. Seven doctoral students in English, all of whom had experience in the teaching of writing, served as primary judges for scoring the writing

samples. Scoring was administered by the director of the freshman writing program.

Scoring guide. All writing samples were scored according to a standardized scoring guide developed specifically for the writing assignments used in the evaluation. The writing samples were rated according to the extent to which they accurately accomplished the two tasks posed by the prompts (i.e., description of information and summarization of trends) while observing conventions of organization, syntax, usage, and mechanics. Accuracy of portrayal of information was treated as the "primary trait" on which scoring was based (White 1994). A 6-point scale was employed for rating writing samples. On this scale, a score of 6 represented essays that were both highly accurate and syntactically error free. In contrast, a score of 1 was assigned to essays that made a recognizable attempt to address the prompt but that wandered off the task, failed to provide an accurate summary, or failed to establish a discernible pattern of organization. Although accuracy of portrayal of information was the primary focus of the scoring, the form of language was also considered; an essay was also scored 1 if the writer was unable to use the vocabulary and idiom required by the writing task or if the writing was marked by frequent, serious errors in grammar, mechanics, and usage. The upper half of the scale (scores 4 through 6) required that the essay address both writing tasks (explaining and summarizing) and have an identifiable structure and command of word choice, syntax, usage, and mechanics. Essays that contained significant factual inaccuracies were rated in the lower half of the scale, as were papers that were holistically judged to have been poorly written in terms of word choice, syntax, usage, and mechanics.

Training materials and procedures. Training materials for judges were prepared for each prompt and included a copy of the prompt, a complete scoring guide, and writing samples illustrating each score on the scale. Writing samples were drawn from those produced by study participants. Training materials also included a section that emphasized differences between adjacent scores on the 6-point scale. Training for each prompt occurred on a different day. Judges reviewed the training materials and then scored a series of writing samples until they reached agreement with the trainer.

Scoring procedure. All writing samples from the evaluation (992 in all) were scored within a single week. The procedures were developed to minimize rater drift and to keep judges blind to students' identity, experimental group within the design, and measurement point (pretest, posttest at end of fall semester, posttest at end of spring semester).

The themes generated from a single prompt were scored on 1 day, which began with training on that prompt. Scoring procedures followed those recommended by White (1994).[5] In addition to the seven judges, there were

two table monitors who spot checked scoring to ensure proper use of the scoring system.

Each writing sample was rated independently by two judges. If these ratings were within one point, the mean of the two ratings was given. In cases in which rating discrepancies were larger than one point, the final score was assigned by one of two highly experienced adjudicators who independently rated the essay.[6]

On each day, each judge scored a balanced set of themes from all three measurement points and all groups in the study. Each judge was paired with each other judge an equal number of times. Each theme was judged once in the morning and once in the afternoon to account for both practice and fatigue effects.

Interjudge reliabilities were determined for each pair of judges. For each judge, the mean correlation (using Fisher's z transformations) of that judge with the other six judges ranged from .53 to .65, with a mean overall interjudge correlation of .61. These reliabilities are comparable to those reported by Breland et al. (1987).

RESULTS

OVERVIEW AND PRELIMINARY ANALYSES

Effect Size Estimates

The focus of our analyses was the estimation of standardized effect sizes for common treatments, but with estimates based on different design components of the evaluation. Our primary interest was in the direct comparison of effect sizes among the three designs: Design A, the randomized tie-breaking experiment; Design B, the nonequivalent control group design; and Design C, the regression discontinuity design.

The standardized effect size estimates reported are based on Hedges and Olkin (1985) and were computed using DSTAT software (Johnson 1989). The familiar standardized effect size estimate, $g = (M_E - M_C)/SD$, where M_E and M_C are the means of the experimental and control group, respectively, and SD is the pooled within-class standard deviation, exhibits small sample bias. Hedges and Olkin have proposed a modified estimate d, which corrects this bias; the estimate d is reported below. (See Hedges and Olkin 1985, 78-81, for a discussion). Standardized effect sizes were estimated based on three

types of data: (a) comparison of outcomes in two groups with pretest measures controlled, (b) partial regression coefficients representing the effect of remediation in the regression discontinuity design, and (c) a comparison of pretest and posttest within individual groups. All the measures were converted to *d* and are thus directly comparable.

Case Retention

All analyses were confined to those students who passed the courses relevant to the particular analysis. Having passed a course was taken to imply that the student had received the full treatment that was to have been provided by the course. Students who withdrew from a course, obtained a grade of "incomplete" in a course, or failed a course typically did not attend the final portion of the course, yielding sparse posttest data.[7] Writing samples were available on only 13% of students who failed, withdrew, or took incomplete grades in standard composition; TSWEs were available on only 22% of these students.

Table 2 summarizes the sample sizes, available demographic characteristics, and standardized test scores of each of the groups at the point of recruitment. The remediation and accretion groups by design took remedial English followed by standard freshman composition. Of the 53 students who were assigned by lottery to the remediation group and who actually enrolled in the university, 48 entered remedial composition in the fall; 39 passed the course (81%). All 39 followed the protocol and entered standard freshman composition in the spring, and 34 (87%) passed standard composition. Of the 76 students recruited into nonremediation, 75 entered the standard freshman composition course in the fall and 68 passed (91%). In the accretion group, 75 entered remedial composition and 60 passed (80%). Of these 60, 44 followed the protocol and entered standard composition, and 36 passed (82%).

Attrition Analyses

Attrition of participants potentially limits the external validity of our evaluation; differential attrition of participants from treatment and control groups potentially compromises the internal validity of our evaluation (Cook and Campbell 1979). Students drop out throughout the academic year in university settings for a variety of reasons.

In our evaluation, there were two general sources of loss during the fall semester: (a) students who were recruited into the randomized experiment

during the summer but who failed to enter the appropriate class in the fall and (b) students who entered the appropriate fall class but who failed to complete the course successfully (i.e., dropped the course, failed the course, or received a grade of incomplete). Sources of loss for the spring semester included failure of students in the remediation and accretion groups to enter standard freshman English composition and again the failure of students to complete the course successfully. To test for differential attrition across groups, we used procedures modeled after those originally suggested by Jurs and Glass (1971).[8] In these procedures, a Treatment Group × Attrition Status analysis of variance (ANOVA) is performed on the available pretest data, here TSWE and writing sample scores at the beginning of the fall semester. Significant Treatment Group × Attrition Status interactions indicate sources of differential attrition.

Two tests of attrition were performed. The first focused on attrition during the fall semester among students whose admissions scores identified them as needing remediation. Three groups were included: remediation and nonremediation from the randomized experiment, plus the accretion comparison group from the nonequivalent control group design. For the pretest writing sample, a 3 × 2 ANOVA with group and retention (retained, not retained) as the variables yielded only a trend to a main effect of retention, $F(1, 176) = 3.25$, $p = .07$. Retained students ($M = 2.43$) had higher pretest scores than students who were not retained ($M = 2.00$) The Group × Retention interaction did not approach significance, $F(2, 176) = 1.94$, *ns*. For the pretest TWSE, there was neither an effect of retention, $F(1, 179) = 1.42$, *ns*, nor an interaction of group by retention, $F(2, 179) = 1.08$, *ns*. The absence of both group by retention interactions indicated that there was no evidence of differential attrition across groups during the course of first semester.

The three designs (randomized experiment, nonequivalent control group design, and regression discontinuity design) also required the comparison of performance across groups at the end of standard freshman English composition. The second attrition analysis focused on attrition up to the time students completed this course, and it included four groups: remediation and nonremediation from the randomized experiment, accretion from the nonequivalent control group design, and standard English group from the regression discontinuity design. The pretest TSWE and writing sample of those students who successfully completed versus failed to complete standard freshman English composition were compared across groups. For the writing sample, a 4 (groups) × 2 (retention) ANOVA showed a significant effect of group, $F(3, 304) = 6.63$, $p < .001$, due to the expected superior pretest scores

of the standard students over the other three groups that were deemed in need of remediation. The Group × Retention interaction did not approach significance, $F(3, 304) = .32$, *ns*. For the TSWE, the group effect reached significance, $F(3, 304) = 33.50$, $p < .001$, again due to the superior performance of the standard students over the other three groups. The Group × Retention interaction again did not approach statistical significance, $F(3, 304) = .17$, *ns*. Once again, there was no evidence of differential attrition across the groups assigned to take standard freshman composition during the course of the evaluation. These analyses rule out the possibility that differences on measured variables at pretest could plausibly account for the obtained results; however, they do not rule out the possibility that differential attrition could be associated with other unmeasured pretest variables (West, Biesanz, and Pitts in press).[9]

Pretest Comparison of Groups Eligible for Remediation

Table 3 provides mean pretest performance on writing measures for the remediation, nonremediation, and accretion groups for students who passed the fall course. All three groups were composed of students whose admissions scores fell either between 12 and 16 on the ACT or between 360 and 380 on the SAT. The remediation group had been randomly assigned by lottery to take the remedial course in the fall semester, whereas the accretion group had self-selected into the course. These two groups did not differ significantly at pretest on either the TSWE, $F(1, 91) = 1.26$, $p > .20$, $d = .24$, or the writing sample, $F(1, 92) = .34$, *ns*, $d = .12$.

The pretest performance of each remediated group (remediation, accretion) was compared to that of the nonremediation group to provide estimates of pretest effect size (Heinsman and Shadish 1996). The remediation and nonremediation groups differed neither on the pretest TSWE, $F(1, 100) = .56$, *ns*, $d = .15$, nor on pretest writing sample, $F(1, 101) = .69$, *ns*, $d = .17$. However, the accretion group had lower pretest TSWE scores that the nonremediation group, $F(1, 116) = 4.38$, $p < .05$, $d = .38$; no corresponding difference was found on the writing sample, $F(1, 116) = .06$, *ns*, $d = .05$.

Given the small to moderate differences observed on pretest writing measures, all between-group comparisons that follow were accomplished with analyses of covariance (ANCOVAs) on posttest scores, with corresponding pretest scores as the covariate.

TABLE 3: Comparison of Performance at Outset (pretest) Versus End (posttest) of Fall Semester for All Groups Falling Below the Cut Points for Remediation

	Test of Standard Written English				*Writing Sample*			
Group (Course)	*Fall Pretest*	*Fall Posttest*	*Test of Change*	*Effect Size*[a]	*Fall Pretest*	*Fall Posttest*	*Test of Change*	*Effect Size*[a]
Remediation (remedial)								
M	36.08	38.44	$t(38) = 2.16$	.34	2.35	2.35	$t(38) = 0.00$	.00
SD	6.32	8.53	$p = .037$	[.30]	.72	.90		[.00]
n	39	39			39	39		
Accretion (remedial)								
M	34.48	36.78	$t(52) = 3.79$	.52	2.44	2.56	$t(45) = .73$	.11
SD	6.65	7.27	$p < .001$	[.39]	.76	1.08	$p = .47$	[.15]
n	54	58			55	51		
Nonremediation (standard)								
M	37.30	38.78	$t(61) = 1.96$	.25	2.48	2.89	$t(61) = 2.64$	.33
SD	7.77	8.29	$p = .055$	[.20]	.93	.98	$p = .01$	[.39]
n	64	64			63	66		

NOTE: This table confined to students who passed the fall course.
a. Effect size estimates taking into account the pre-post correlation (no brackets), and correcting for pre-post correlation (in brackets) according to Dunlap et al. (1996, Equation 3; see Footnote 9).

MAIN ANALYSES ADDRESSING CENTRAL EVALUATION QUESTIONS

We consider each of the central evaluation questions below. The impact of the courses is considered from several statistical perspectives including conventional significance testing, confidence intervals, and standardized effect size measures. For questions on which multiple designs can be brought to bear, separate effect size estimates are presented for each design.

1. Do Students Show Gains From the First Semester Course?

The three experimental groups portrayed in Table 3 had placement test scores that led them to be characterized as needing remediation. Of these groups, the remediation and accretion groups took the remedial course in the fall semester; the nonremediation group took standard freshman composition. Posttest writing measures at the end of the fall semester are given in Table 3 for both the TSWE and writing sample, along with tests of significance of change from pretest to posttest within each group. All three groups showed gains on the TSWE. However, only the nonremediation group exhibited gains on the writing sample.

2. How Do Students Receiving Remediation Versus No Remediation Compare at the End of the Fall Semester?

Design A. Table 4 presents pairwise comparisons of the remediation group with the nonremediation group at the end the fall semester in the randomized experiment. An ANCOVA of posttest TSWE with the pretest TSWE as the covariate showed no difference on the TSWE, $d = -.08$. For the writing sample, the nonremediation group, which received standard freshman English composition, exhibited significantly higher writing scores than the remediation group, $d = -.51$, $p = .01$. (The negative effect sizes result from the order of means, i.e., remedial course-standard course).

Design B. Parallel analyses to those reported for Design A are given for Design B, the nonequivalent control group design, in Table 4. For the TSWE, there was again no difference between the accretion group, which received remedial English, and the nonremediation group, which received standard Freshman English, $d = -.07$. For the writing sample, the comparison paralleled that of Design A, (i.e., higher performance in nonremediation), $d = -.27$, though the comparison did not reach conventional significance levels.

TABLE 4: Outcomes at the End of Fall Semester Among Students Requiring Remediation Who Took Remedial Composition Versus Standard Freshman Composition

Design	Groups	F Test[a]	d[b]	95% Confidence Interval[c]
	Test of standard written English			
A	Remediation vs. nonremediation[d]	$F(1, 98) = .14$	−.08	(−.48, .33)
B	Accretion vs. nonremediation[d]	$F(1, 112) = .15$	−.07	(−.44, .29)
	Remediation vs. accretion[e]	$F(1, 89) = .00$	.00	(−.41, .41)
	Writing sample			
A	Remediation vs. nonremediation[d]	$F(1, 98) = 6.23^*$	−.51	(−.91, −.10)
B	Accretion vs. nonremediation[d]	$F(1, 105) = 1.94$	−.27	(−.65, .11)
	Remediation vs. accretion[e]	$F(1, 82) = .96$	−.21	(−.64, .22)

a. Analysis of covariance with corresponding pretest measure as covariate.
b. Sign of effect size reflects discrepancy of first group listed minus second group listed.
c. 95% confidence interval for effect size estimate.
d. Comparison of outcomes following remedial course (first group) versus standard course (second group).
e. Comparison of randomly assigned (remediation) versus self-selected (accretion) students who had equivalent admissions test scores and who took remedial composition during the fall semester.
*$p < .05$.

Summary. All groups gained on the TSWE during the first semester, but only the group receiving standard freshman composition gained on the writing sample. For the writing sample outcome, effect sizes were −.51 and −.27 from Designs A and B, respectively, for the disadvantage of having taken remedial writing rather than standard freshman composition.

3. Do Students Show Gains at the End of Standard Freshman Composition?

Table 5 presents the performance of the remediation, accretion, and regular groups at the outset and end of the standard freshman composition course. The corresponding data for the nonremediation group was presented in Table 3. On the writing sample, the nonremediation, the remediation, the accretion, and the standard group showed significant gain during the standard freshman English composition course. The accretion group's gain did not reach conventional significance levels.

TABLE 5: Comparison of Performance at Outset Versus End of Standard Freshman Composition

Group (Course)	*Test of Standard Written English*				*Writing Sample*			
	Before Course	*After Course*	*Test of Change*	*Effect Size*[a]	*Before Course*	*After Course*	*Test of Change*	*Effect Size*[a]
Remediation (standard)								
M	38.74	41.48	$t(26) = 1.60$	.30	2.31	2.97	$t(32) = 3.04$	.52
SD	8.85	7.00	$p = .12$	[.26]	.95	.90	$p = .005$	[.68]
n	34	27			34	33		
Accretion (standard)								
M	37.92	40.39	$t(30) = 1.61$	.28	2.56	2.89	$t(31) = 1.54$	.27
SD	6.61	7.47	$p = .12$	[.30]	1.22	.77	$p = .13$	[.35]
n	38	31			33	37		
Regular (standard)								
M	46.45	47.07	$t(104) = .48$	.05	3.10	3.37	$t(101) = 2.21$	.21
SD	7.28	9.95	$p = .63$	[.04]	1.04	1.08	$p = .03$	[.25]
n	115	109			111	112		
Spring (standard)								
M	—	45.91			—	3.48		
SD	—	10.32			—	.89		
n	—	22			—	26		

NOTE: Performance of nonremediation group in standard composition is given in Table 3.
a. Effect size estimates taking into account the pre-post correlation (no brackets), and correcting for pre-post correlation (in brackets) according to Dunlap et al. (1996, Equation 3; see Footnote 9).

4. How Do Students Receiving Remediation Versus No Remediation Compare at the End of Standard Freshman Composition?

Design A. We used ANCOVA to compare the outcomes at the end of the standard English composition course for the remediation versus the nonremediation group in the randomized experiment (see Table 6). Initial pretest scores from the beginning of fall semester served as the covariate. For the TSWE, the remediation group exhibited an advantage of moderate effect size ($d = .59$) over the nonremediation group. For the writing sample, the remediation group showed little, if any, advantage over the nonremediation group ($d = .16$).

Design B. Similar estimates of effect size were found in the nonequivalent control group design. The accretion group had an advantage over the nonremediation group on the TSWE ($d = .57$), but not on the writing sample ($d = .06$).

Design C. The regression discontinuity design included the data from the remediation, accretion, and standard groups. Each writing measure at the end of the standard freshman composition course was predicted from three measures: (a) the admissions test score (ACT English or SAT Verbal), which served as the quantitative measure of need; (b) the score on the same writing measure at the outset of the evaluation; and (c) a binary treatment variable, coded 1 if the student had taken the remedial English course, 0 otherwise. Students had taken the SAT Verbal, the ACT English, or both admissions tests; consequently, we ran separate sets of regression analyses for each test, including in each analysis all students who had taken the test in question. In all, 55% of students in the three groups included in the regression discontinuity design had taken the ACT English; 76%, the SAT Verbal. Listwise deletion was used for missing data.

Table 7 presents the results of the regression discontinuity analyses. Our interest is in effect size and significance of the regression coefficient for the binary treatment predictor, which estimates the effect of having taken the remedial course on the writing performance. The test of the effect of treatment (remedial writing) reached conventional levels of statistical significance for the TSWE with the ACT as the admissions test score, but not with SAT as the admissions test score. The effect sizes from both TSWE analyses were small to moderate, $d = .32$ for SAT sample and $d = .49$ for ACT sample. Analyses of the writing samples did not support a substantial impact of treatment with either admissions test. Effect size measures from both writing sample analyses were small: $d = .02$ for SAT sample; $d = .22$, for ACT sample.

The basic regression discontinuity design assumes that the form of the regression is linear and that there is no Treatment × Selection variable (SAT,

TABLE 6: Outcomes at the End of Standard Freshman Composition Students Requiring Remediation Who Took Remedial Composition Plus Standard Freshman Composition Versus Only Standard Freshman Composition

Design	*Groups*	*F Test*[a]	d[b]	*95% Confidence Interval*[c]
	Test of standard written English			
A	Remediation vs. nonremediation[d]	$F(1, 86) = 6.68^*$	.59	(.13, 1.05)
B	Accretion vs. nonremediation[d]	$F(1, 88) = 6.57^*$	.57	(.12, 1.02)
	Remediation vs. accretion[e]	$F(1, 53) = .01$	.02	(–.49, .55)
	Writing sample			
A	Remediation vs. nonremediation[d]	$F(1, 92) = .54$	.16	(–.27, .58)
B	Accretion vs. nonremediation[d]	$F(1, 95) = .07$	.06	(–.35, .47)
	Remediation vs. accretion[e]	$F(1, 66) = .17$	.10	(–.38, .57)

a. Analysis of covariance with corresponding pretest measure at outset of fall semester as covariate.
b. Sign of effect size reflects discrepancy of first group listed minus second group listed.
c. 95% confidence interval for effect size estimate.
d. Comparison of outcomes following both courses (first group) versus only standard course (second group).
e. Comparison of randomly assigned (remediation) versus self-selected (accretion) students who had equivalent admissions test scores and who took both remedial and standard freshman composition.
$^*p < .05$.

ACT) interaction (Trochim 1984). We performed a sensitivity analysis of the effect size estimates under changes in the model by testing two additional regression discontinuity models, one containing in addition a quadratic selection term, and one containing an interaction between the selection variable and the binary treatment predictor. There was no appreciable change in effect size estimates.

The cut scores for the ACT and SAT had been determined by the institution. In all, 414 students were enrolled in all sections of remedial English in the fall semester. Of these, 236 took the SAT; 223 the ACT. On a post hoc basis, we observed that the ACT range of 12 to 16 covered the 34th to 92nd percentile of the 223 students. In contrast, the SAT range of 360 to 380 covered the 73rd to 91st percentiles of the 236 students. We recomputed the regression discontinuity design with the ACT sample including only students with ACT scores of 15 and 16 (the 70th to 92nd percentiles); there was again no appreciable change in effect size estimates.

TABLE 7: Outcomes and Effect Size Estimates From the Regression Discontinuity Design

Outcome Variable	*Admissions Test*	*Standardized Regression Coefficients*			t *Test for Treatment*	d	*95% Confidence Interval*[b]
		Admissions	*Pretest*	*Treatment*[a]			
TSWE	SAT	.22*	.55***	.16	$t(118) = 1.61$	.32	(–.07, .72)
TSWE	ACT	.34^{+}	.54***	.34*	$t(90) = 2.33$*	.49	(.06, .91)
Writing	SAT	.20*	.22**	.01	$t(143) = .12$	.02	(–.33, .37)
Writing	ACT	.38*	.26**	.17	$t(114) = 1.19$	.22	(–.15, .60)

$^{+}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

a. Binary measure = 1 if student had taken remedial composition, 0 otherwise.

b. Confidence interval on effect size estimate.

Summary. All three designs provide evidence of the positive impact of remedial English composition on the TSWE but not on the writing sample. For the TSWE effect sizes, estimates were highly similar across designs A and B, with slightly smaller estimates from Design C. For the writing sample, estimates from Designs A, B, and C for the writing sample were highly similar. Calculation of 95% confidence intervals suggested that there was no meaningful variation in these estimates as a function of design type.

5. What Is the Pattern of Gain of Students Taking Remedial Writing During the Two-Course Sequence?

The remediation and the accretion groups had taken remedial English composition followed by standard freshman English composition. Table 8 summarizes the gains in the two measures of writing during the year, providing a preexperimental estimate (Campbell and Stanley 1963) of effect size for the impact of remedial English.[10] These estimates should be compared with the estimate of the gain from the standard English composition course alone, presented in Table 3 for the nonremediation group and in Table 5 for the regular group that was not in need of remediation. Over the year, large effect size gains were noted for the TSWE, gains that had accrued over both semesters. For the writing sample, moderate gains were noted; these had accrued during the spring semester, in standard freshman composition.

PROBING A THREAT TO VALIDITY: CYCLIC INSTITUTIONAL VARIATION

Our evaluation involved comparisons of standard freshman English composition across groups that had completed this course during the fall versus the spring semesters. Consequently, we compared random samples of students who had self-selected into standard freshman composition during the fall semester (regular group) versus the spring semester (spring group). These groups did not differ on admissions test scores: for SAT Verbal, $F(1, 116) = .79$, *ns*; for ACT English, $F(1, 97) = 1.21$, *ns* (see Table 2). They also did not differ on writing outcomes at the end of the standard English composition course: for writing sample, $F(1, 136) = .25$, *ns*; for TSWE, $F(1, 129) = .25$, *ns*. Thus, there was no evidence of difference in quality of students in the standard English composition classes during the fall versus spring semesters.

TABLE 8: Comparison of Performance at Outset of Remedial Composition Versus at End of Standard Composition

	Test of Standard Written English				*Writing Sample*			
Group (Design)	*Before Remedial Course*	*After Standard Course*	*Test of Change*	*Effect Size (Confidence Interval)*[a]	*Before Remedial Course*	*After Standard Course*	*Test of Change*	*Effect Size (Confidence Interval)*[a]
Remediation (design A)								
M	36.00	41.48	$t(26) = 5.15^{***}$	.96	2.35	2.97	$t(32) = 3.34^{**}$	.57
SD	6.31	7.00		(.39, 1.52)	.73	.90		(.08, 1.06)
n	27	27			33	33		
Accretion (design B)								
M	34.03	40.28	$t(28) = 4.85^{***}$	.87	2.44	2.90	$t(36) = 2.74^{**}$	.45
SD	6.19	7.72		(.34, 1.41)	.75	.78		(–.02, .91)
n	29	29			36	35		

NOTE: Table based only on students who passed standard composition.
a. 95% confidence interval for effect size, see also note 10.
$^{**}p < .01$. $^{***}p < .001$.

DISCUSSION

CONVERGENCE OVER DESIGNS

The Randomized Experiment (Design A) and the Nonequivalent Control group Design (Design B) led to a remarkably consistent overall conclusion: Compared to remedial students who took the one-semester standard English composition course, those remedial students who took a remedial course followed by standard English composition exhibited superior performance of moderate effect size on the TSWE (d = .59, .57, respectively), with almost no benefit on the writing sample (d = .16, .06, respectively). The regression discontinuity design yielded a pattern of effect sizes for the TSWE and writing sample that was similar to that from Designs A and B: larger effect size estimates for the TSWE than for the writing sample. Specifically, for the TSWE, the effect sizes from the regression discontinuity design were .32 and .49, with SAT versus ACT as the selection criterion, respectively, compared to .59 and .57 from Designs A and B, respectively. Similarly, for the writing sample, the effect sizes from the regression discontinuity design were .02 and .22, with SAT versus ACT as the selection criterion, respectively, compared to .16 and .06 from Designs A and B, respectively.

Looked at from the perspective of change during the full academic year from entrance into the remedial course to completion of standard freshman composition, the remediation and accretion groups exhibited, during the two-semester sequence, large gains on the TSWE (d = .96, .87, respectively) and moderate gains on the writing sample (d= .57, .45, respectively).

Between- Versus Within-Subject Effect Sizes

The question may be raised as to why the between-subject effect size estimates obtained from comparisons of the remediation and accretion to the nonremediation group (Table 6) were substantially smaller than those from the pre-post comparisons within the experimental groups (remediation, accretion) (Table 8). The answer is simply that the nonremediation group, which served as the control, also received a treatment, that is, the semester of standard English composition. The nonremediation group exhibited gains on both writing measures during this one-semester course. Thus, the relative gain in the remediation and accretion groups, compared to that in the nonremediation group, was smaller than the absolute gain from initial pretest to final posttest during the two-semester period in the remediation and accretion groups. Shadish (Heinsman and Shadish 1996; Shadish and

Ragsdale 1996) reported larger effect sizes in between-study comparisons in which the control group was passive (received no treatment) than in which the control group was active (received some form of treatment).

The strong convergence of effect size estimates across Designs A and B is in part due to the fact that a common control group (the nonremediation group) was used in both designs. Further, Design C, the regression discontinuity design, used all the data from the nonremediation and accretion groups in effect size estimation. Thus, there is again overlap in the cases included in estimates of effect size from Design C with those derived in Designs A and B.

Beyond this overlap of designs, a substantial number of design-related factors also contributed to the convergence of outcomes across designs. The same population was used in Designs A an B: entering freshmen, who by their placement scores, were mandated to take remedial English (Designs A and B), and whose SAT or ACT scores fell within a small range below a cut score. Entering freshmen from the same recruitment pool for whom remediation was not required were added in Design C.

Heinsman and Shadish (1996) identified a substantial set of methodological factors that have contributed to discrepancies in effect size estimates from randomized experiments versus nonequivalent control group designs in existing literature. In our implementation of the true experiment (Design A) and the nonequivalent control group design (Design B), we held constant a number of these design features. These features, which are briefly considered below, apparently led to close correspondence in effect size estimates across the designs.

(a) Use of internal controls. Nonequivalent control group designs may use internal controls, drawn from the same population as the experimental group, or external controls drawn from a different population. Here, the controls were internal. Shadish and Ragsdale (1996) argue that the use of internal controls is critical for obtaining accurate effect sizes from quasi-experiments.

(b) Selection. With regard to selection, the pool of remedial students was mandated by the university to take remedial composition; this minimized differential selection into the remediation versus accretion groups on the basis of the verbal skills measured by the SAT and ACT. Identifiable potential sources of selection bias are the earlier matriculation of the remediation group than of the accretion group and the inability to reach some students in the accretion group during the recruitment. How these factors might correlate with posttest measures of writing competency is unclear. Shadish and

Ragsdale (1996) have argued that having control groups that do not self-select is also critical for obtaining accurate effect size estimates from quasi-experiments.

(c) Activity level in the control group. That the nonremediation group served as the comparison group for both Designs A and B meant that the critical factor of activity in the control group, that is, no treatment versus some standard treatment, was held constant across designs (Heinsman and Shadish 1996).

(d) Statistical controls for pretest differences. In all three designs, we used pretest scores on the writing measures as covariates in the computation of effect sizes for between-group comparisons. Adjusting for pretest differences leads to more similar effect size estimates (Heinsman and Shadish 1996; Ray and Shadish 1996; Shadish and Ragsdale 1996).

(e) Effect size calculations. All our standardized effect size estimates were direct estimates and were in the same metric, *d*. For the regression discontinuity design, the *t* test for the regression coefficient of the binary-coded treatment variable was converted to the effect size *d*, with sample sizes corresponding to the number of individuals in the analysis who did versus did not receive remediation. Use of direct versus indirect effect size estimates leads to different estimates of magnitude of effect (Heinsman and Shadish 1996; Ray and Shadish 1996).

A second factor that may have contributed to the convergence of effect size estimates, particularly in Designs A and B, is that we limited our analysis to those individuals who had passed their courses. Thus, we minimized variation in exposure to treatment across designs.

REGRESSION DISCONTINUITY DESIGN

The regression discontinuity design showed a significant effect of remediation on TSWE scores with ACT as the selection variable, whereas this effect failed to reach conventional levels of statistical significance when the SAT served as the selection variable. This discrepancy was not due to sample size, which was approximately equal for the SAT (n = 122) and the ACT analysis (n = 118). In addition, the ACT and SAT were equally correlated with the posttest TSWE (both rs = .49 in both cases). However, the marginal split of students who had taken versus not taken remediation was more equal in the ACT than the SAT regression discontinuity analysis of the TSWE: (.39/.61 versus .29/.71, respectively), a discrepancy that would be expected to decrease the power of the statistical test for the SAT analysis.

In general, the regression discontinuity design can be expected to have lower power than the randomized experiment to detect true effects of treatment. If the sizes of the treated and untreated groups are equal, Goldberger (1972) estimates that the regression discontinuity design may often require 2.5 times the number of cases as the randomized experiment to achieve an equal level of statistical power. Sample size demands to achieve equal power will be even greater in the regression discontinuity design because unequal proportions of the participants typically receive the treatment and control conditions (Higginbotham, West, and Forsyth 1988). Treatments allocated on the basis of need or merit are typically given to a relatively small proportion of the sample of participants.

TREATMENT AS DOSAGE

The remedial versus the standard English composition course was substantially different in the focus of the writing. The remedial course focused on writing from experience; the standard course, on writing from reading and observation. Yet, one might ask whether the effects of the combined year are a matter of dosage of exposure to any training in writing whatever (Sechrest et al. 1979). The TSWE scores improved in each of the courses; students who took both courses continued to show gains in their TSWE scores in the second course. The writing sample measured skill in writing from data, a skill taught in the standard but not the remedial course. It is not surprising then that on the writing sample, improvement was associated with the standard freshman composition course but not the remedial course. This latter finding had an important policy impact with regard to the remedial writing experience.

POLICY IMPACT OF THE EVALUATION

An early report of the present evaluation was presented to the Department of English, the Office of the Provost, and the State Board of Regents. Following consideration of the evaluation by these groups, the one-semester remedial writing program was judged to be ineffective in improving the types of writing deemed critical for success in college. Of critical importance, writing sample scores had not improved during the remedial course, and students who had had both remediation and standard freshman composition had no higher writing scores at the end of the year that did remedial students who had only taken the semester of standard English composition. More-

over, our evaluation showed that the rate at which remedial students passed standard freshman composition without prior remediation (91%) exceeded the rate of passing the remedial course (82%). These results led to a substantial redesign of the freshman year writing experience for students described as "beginning writers" (Glau 1996), that is, students with SAT scores (after upward rescaling of the SAT) of 460 or below or ACT scores of 18 or below. These students take a "stretch" standard English composition course, a 1-year course that covers the material of standard freshman composition (Glau 1996). No longer do they write from experience, as in the remedial course; rather, their full writing efforts are devoted to writing drawn from reading or observation, the sort of writing required for success in university-level courses.

INSTITUTIONAL RESOURCES REQUIRED FOR THE EVALUATION DESIGNS

This evaluation required substantial institutional monetary resources, as well as the cooperation of the Offices of Student Affairs, Admissions, the Registrar, Institutional Analysis, Student Housing, the Department of English, and the Freshman writing program. Of the three designs, the randomized experiment required the greatest amount of institutional resources and cooperation, due to the complex recruitment process. Creation of the accretion group that formed the basis of the quasi-experimental design was much less demanding on resources, in that it was composed of students who had already come to the university and had enrolled in remedial English. However, the creation of the nonremediation group required that entering students be contacted prior to registration and be placed into sections of standard Freshman composition, that is, the same process required for the true experiment. The regression discontinuity design posed only a relatively small incremental demand over those of the randomized experiment and nonequivalent control group designs, specifically the addition of students who had taken standard English composition to those who had taken the remedial course. Absent the other designs, the regression discontinuity design could have been implemented by sampling students from the remedial course versus standard English composition after the fall semester had begun, with no need for the complex recruitment process of the randomized experiment. Of course, the larger sample size required to achieve statistical power comparable to that of the randomized experiment would greatly increase the costs associated with scoring the writing samples.

CONCLUSION

Dennis (1990) has argued that randomized field experiments are subject to a variety of logistical problems that undermine their validity and that quasi-experiments are easier to manage (Dennis and Boruch 1989). Further, Dennis (1990) suggested that although the scientific ideal may be a sequence of studies, policy makers often require results in a short time frame that does not permit the sequential accumulation of research findings. Dennis recommended the simultaneous implementation of both true and quasi-experimental research. When the evaluation outcomes converge, there would be a strong empirical basis from which to inform policy.

Institutional settings provide an arena in which multiple designs can be implemented and evaluated. We strongly recommend the incremental effort to add design components to a core evaluation design that permit the contrasting of conclusions reached from multiple outcomes. This strategy offers two advantages. First, it provides multiple tests of outcomes that may rule out alternative explanations of outcomes and lead to stronger substantive conclusions (Cook 1985). Second, as called for by Dennis and Boruch (1989), the strategy contributes needed insights into the conditions under which quasi-experiments will yield approximately the same treatment effect estimates as randomized experiments. Finally, Heinsman and Shadish (1996) concluded that if important design features, such as those we have highlighted above, are held constant, then nonequivalent control group designs "can yield a reasonably accurate effect size in comparison with randomized designs" (p. 154). The present study provides strong support for this conclusion.

NOTES

1. The SAT was rescaled by Educational Testing Service after data collection for this study; thus, the SAT cut scores used in this study are equivalent to somewhat higher scores on the current SAT.

2. A small percentage (6%) of the total participants in the study were beyond their freshman year; they are excluded from all analyses.

3. Because of excess enrollment pressure, spaces in required English courses were not available for all students. Thus, the guarantee of a place in standard freshman English was a clear incentive for participation.

4. The third author was director of the freshman writing program of the University at the time of the evaluation. He and the fourth author, an evaluation specialist, collaborated in the direction of all components of the evaluation that involved scoring of writing samples, including the development of the scoring guide, the training of judges, and the management of the scoring

sessions. See White (1994) for extensive discussion for the need for writing professionals to have significant roles in the evaluation of writing programs.

5. The scoring procedures also reflect standard practice for scoring writing samples employed by the Educational Testing Service (ETS).

6. The adjudicators were the director of the freshman writing program and a member of the English faculty who had substantial experience in essay scoring at ETS.

7. We attempted to test students on whom data were missing outside class during examination week, before they left campus either at the end of the fall or spring semester. Our attempt was unsuccessful, even with the offer of monetary compensation.

8. The Jurs and Glass (1971) analysis assumes random assignment to treatment condition.

9. Angrist, Imbens, and Rubin (1996) have developed procedures for obtaining unbiased estimates of effect sizes in randomized experiments when there are noncompliers—participants who are randomly assigned to the treatment condition but who do not comply with the treatment assignment. These methods assume noncompliers do not receive *any* of the treatment protocol, but instead receive the control treatment (usually no treatment). They also assume that *all* participants have been measured at posttest (see West, Biesanz, and Pitts in press, for a discussion of this and other approaches to treatment noncompliance and attrition). Given that the necessary conditions for applying the Angrist, Imbens, and Rubin procedure were not met in the present study, we used the traditional Jurs and Glass (1971) procedure for identifying pretest variables associated with differential attrition from treatment and control conditions.

10. Dunlap et al. (1996) have argued that effect size d estimates computed from repeated measures t-tests are biased because the denominator of d is reduced by virtue of the correlation between pretest and posttest, that is, the standard deviation of the difference between correlated means and not that between uncorrelated means (the pooled within-class standard deviation) serves as the denominator of d. They point out that the Hedges and Olkin (1985) formulation does not address repeated measurements. Dunlap et al. (1996) derived the correct estimate of effect size d from the repeated measures t-test t_c, that is, $d = t_c[2(1 - r)/n]^{1/2}$, where r is the pre-post correlation (their Equation 3, p. 171). Effect sizes in Table 8 are based on Hedges and Olkin (1985). Corresponding effect sizes based on Dunlap et al. (1996) are .82 and .85, for the TSWE, in Designs A and B, respectively (in lieu of .96 and .87, respectively); for the writing sample, .76 and .60 (in lieu of .57 and .45).

REFERENCES

Angrist, J. D., Imbens, G. W., and Rubin, D. B. 1996. Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association* 91:444-89.

Becker, B. J. 1990. Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal. *Review of Educational Research* 60:373-417.

Berk, R. A., and D. Rauma. 1983. Capitalizing on nonrandom assignment to treatment: A regression discontinuity of a crime program. *Journal of the American Statistical Association* 78:21-8.

Boruch, R. F. 1975. Coupling randomized experiments and approximations to experiments in social program evaluation. *Sociological Methods and Research* 4:31-53.

———. 1997. *Randomized controlled experiments for planning and evaluation: A practical guide*. Thousand Oaks, CA: Sage.

Breland, H., R. Camp, R. J. Jones, M. Morris, and D. A. Rock. 1987. *Assessing writing skill.* New York: College Entrance Examination Board.

Breland, H. M., and J. L. Gaynor. 1979. A comparison of direct and indirect assessments of writing skill. *Journal of Educational Measurement* 16:119-28.

Campbell, D. T. 1984. Forward. In *Research design for program evaluation: The regression discontinuity design*, edited by W.M.K. Trochim. Beverly Hills, CA: Sage.

Campbell, D. T., and Stanley, J. C. 1963. *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Colditz, G. A., J. N. Miller, and F. Mosteller. 1988. The effect of study design on gain in evaluation of new treatments in medicine and surgery. *Drug Information Journal* 22:343-52.

Cook, T. D. 1985. Postpositivist critical multiplism. In *Social science and social policy*, edited by L. Shotland and M. Mark, 21-62. Beverly Hills, CA: Sage.

Cook, T. D., and D. T. Campbell. 1979. *Quasi-experimentation: Design and analysis issues for field settings.* Boston: Houghton Mifflin.

Cooper, W. H., and A. J. Richardson. 1986. Unfair comparisons. *Journal of Applied Psychology* 71:179-84.

Dennis, M. L. 1990. Assessing the validity of randomized field experiments: An example from drug abuse treatment research. *Evaluation Review* 14:347-73.

Dennis, M. L., and R. F. Boruch. 1989. Randomized experiments for planning and testing projects in developing countries. *Evaluation Review* 13:292-309.

Dunlap, W. P., J. M. Cortinal, J. B. Vaslow, and M. J. Burke. 1996. Meta-analysis of experiments with matched pairs or repeated measures designs. *Psychological Methods* 1:170-7.

Faigley, R,. R. Cherry, D. Jollisee, and A. Skinner. 1985. *Assessing writers' knowledge and processes of composing.* Norwood, NJ: Ablex.

Gilbert, J. P., B. McPeek, and F. Mosteller. 1978. Statistics and ethics in surgery and anesthesia. *Science* 78:684-9.

Glau, G. R. 1996. The "Stretch Program"; Arizona State University's new model of university-level basic writing instruction. *WPA: Writing Program Administration* 20:79-91.

Goldberger, A. S. 1972. *Selection bias in evaluating treatment effects: Some formal illustrations.* Discussion paper 123-72. Madison: University of Wisconsin, Institute for Research on Poverty.

Gomez, F. 1985. Community-based distribution—the case of Colombia. In *Health and family planning in community-based distribution programs*, edited by M. Warner, S. Huffman, D. Cebula, and R. Osborn, 203-13. Boulder, CO: Westview.

Hazelrigg, M. D., H. M. Cooper, and C. M. Borduin. 1987. Evaluating the effectiveness of family therapies: An integrative review and analysis. *Psychological Bulletin* 93:388-95.

Hedges, L. V., and I. Olkin. 1985. *Statistical methods for meta-analysis.* San Diego, CA: Academic Press.

Heinsman, D. T., and W. R. Shadish. 1996. Assignment methods in experimentation: When do nonrandomized experiments approximate the answers from randomized experiments? *Psychological Methods* 1:154-69.

Higginbotham, H. N., S. G. West, and D. R. Forsyth. 1988. *Psychotherapy and behavior change: Social, cultural and methodological perspectives.* New York: Pergamon.

Johnson, B. T. 1989. *DSTAT: Software for the meta-analytic review of research literature.* Hillsdale, NJ: Lawrence Erlbaum.

Judd, C. M., and D. A. Kenny. 1981. *Estimating the effects of social interventions.* Cambridge, UK: Cambridge University Press.

Jurs, S. G., and G. V. Glass. 1971. The effect of experimental mortality on the internal and external validity of the randomized comparative experiment. *Journal of Experimental Education* 40:62-6.

LaLonde, R. 1986. Evaluating the econometric evaluations of training programs. *American Economic Review* 76:604-20.

Lipsey, M. W., and D. B. Wilson. 1993. The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist* 48:1181-209.

Meier, P. 1985. The biggest ever public health experiment ever: The 1954 field trial of the Salk poliomyelitis vaccine. In *Statistics: A guide to the unknown*. 2d ed. Edited by J. M. Tanur, F. Mosteller, W. H. Kruskal, R. S. Pieters, G. R. Rising, and E. L. Lehmann, 3-15. Monterey, CA: Wadsworth & Brooks/Cole.

Miao, L. L. 1977. Gastric freezing: An example of the evaluation of medical therapy by randomized clinical trials. In *Costs, risks, and benefits of surgery*, edited by J. P. Bunker, B. A. Barnes, and F. Mosteller, 198-211. New York: Oxford.

Ray, J. W., and W. R. Shadish. 1996. How interchangeable are different estimators of effect size? *Journal of Consulting and Clinical Psychology* 64:1316-25.

Reichardt, C. S., and M. M. Mark. 1997. Quasi-experimentation. In *Handbook of applied research methods*, edited by L. Bickman and D. J. Rog, 193-228. Thousand Oaks, CA: Sage.

Sechrest, L., S. G. West, M. A. Phillips, R. Redner, and W. Yeaton. 1979. Some neglected problems in evaluation research: Strength and integrity of treatments. In *Evaluation studies review annual*. Vol. 4. Edited by L. Sechrest and Associates, 15-35. Beverly Hills, CA: Sage.

Shadish, W. R., T. D. Cook, and D. T. Campbell. 1997. *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Allyn & Bacon. In preparation.

Shadish, W. R., and D. T. Heinsman. 1997. Experiments versus quasi- experiments: Do you get the same answer? In *Meta-analysis of drug abuse prevention programs*. NIDA Research Monograph, DHHS Publication No. (ADM) 97-170. Edited by W. J. Bukoski, 147-64. Washington, DC: Superintendent of Documents.

Shadish, W. R., and K. Ragsdale. 1996. Random versus nonrandom assignment in controlled experiments: Do you get the same answer? *Journal of Consulting and Clinical Psychology* 64:1290-305.

Shapiro, D. A., and D. Shapiro. 1982. Meta-analysis of comparative therapy outcome studies: A replication and refinement. *Psychological Bulletin* 92:581-604.

Smith, M. L., G. V. Glass, and T. I. Miller. 1980. *The benefits of psychotherapy*. Baltimore: Johns Hopkins University Press.

Trochim, W.M.K. 1984. *Research design for program evaluation: The regression discontinuity design*. Beverly Hills, CA: Sage.

———. 1990. The regression-discontinuity design. In *Research methodology: Strengthening causal interpretations of nonexperimental data*, edited by L. Sechrest, E. Perrin, and J. Bunker. Washington, DC: U.S. Department of Health and Human Services, Agency for Health Care Policy and Research.

West, S. G., J. Biesanz, and S. C. Pitts. In press. Causal inference in field settings. In *Handbook of research methods in social psychology*, edited by H. T. Reis and C. M. Judd. New York: Cambridge University Press.

White, E. M. 1994. *Teaching and assessing writing*. 2d ed. San Francisco: Jossey-Bass.

———. 1995. The importance of placement and basic studies: Helping students succeed under the new elitism. *Journal of Basic Writing* 14 (2): 75-84.

White, E. M., and L. Polin. 1986. *Research in effective teaching of writing: Final Report*. NIE-G-81-0011 and NIE-G-82-0024. Washington, DC: National Institute of Education. (ERIC Document Reproduction Service No. ED 275 007)

Witte, S. P., and L. Faigley. 1983. *Evaluating college writing programs*. Carbondale: Southern Illinois University Press.

Leona S. Aiken is currently professor of psychology at Arizona State University and director of the Concentration in Quantitative Psychology. She is coauthor of Multiple Regression: Testing and Interpreting Interactions (1991) and Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences *(3d ed., forthcoming). Her primary research interests are in women's health, with specific emphasis on the design, implementation, and evaluation of interventions to increase preventive health behaviors.*

Stephen G. West is currently professor of psychology and coprincipal investigator of the National Institute of Mental Health funded Preventive Intervention Research Center at Arizona State University. His primary research interests are in the design and statistical analysis of field research and the development and evaluation of theory-based preventive interventions.

David E. Schwalm is associate professor of English at Arizona State University Main. He is currently serving as vice provost for Academic Affairs and dean of East College at Arizona State University East, a new campus of ASU. His current academic research interest is in determining the degree of difficulty of writing tasks.

James L. Carroll is currently president/CEO of VIA, Inc., developers of wearable computer systems and integrated network system solutions. His primary interests are in implementation and evaluation of mobile communication and computing systems for industry, business, and education.

Shenghwa Hsiung received his Ph.D. in cognitive psychology from Arizona State University in 1993. His dissertation addressed determinants of recognition of printed Chinese words. He currently works as a counselor for illegal immigrant youth who have been detained by immigration and naturalization officers.