Impact of the Wisconsin Emerging Scholars First-Semester Calculus Program on Grades and Retention from Fall `93-`96

by
Steve Kosciuk

July 8, 1997

This research was funded by the Advanced Research Projects Agency-Technology Reinvestment Project grant “Diversity and Cultural change: Manufacturing Engineering Education for the Future” to the Engineering Research Center for Plasma-Aided Manufacturing, under grant #EEC8721545, administered by the NSF.

Executive Summary

During the Fall semesters from 1993 to 1996 the UW-Madison Department of Mathematics ran a total of 11 Wisconsin Emerging Scholars (WES) sections distributed over several first-semester calculus (Math 221) lectures. The students considered here were all first-semester freshman with no advanced standing, were 18 or 19 years old, and had enrolled in Math 221 in one of those four Fall semesters. Overall, we compared 169 WES students to 3,871 non-WES students.

We examined the WES program in terms of its impact on students:

1. success in calculus
2. retention in science, math, engineering, or technology (SMET) majors.

"Success" in calculus was quantified in terms of the proportions of students receiving a B or above in calculus. Specifically, for the various groups of students of interest, we analyzed the "odds of success," defined as the ratio of the number of students in the group with a B or above to the number with a BC or below. This measure was chosen because it concisely captures the most relevant part of the distribution of grades for our purposes. In contrast, differences in "mean" grades, for example, leave unanswered the question of whether one groups' higher average was due simply to more Cs in proportion to Ds, as opposed to more As and Bs in proportion to Cs and Ds. See (1,2) for some analyses in terms of mean grades.

Both of these factors were broken down by several other factors of interest. These include prior achievement or preparation (e.g., ACT, SAT math scores, UW-math placement scores, etc.), calculus lecture, gender, minority status, and whether the student was in the College of Engineering.

1. **Impact of WES on success in calculus**
   Roughly speaking, we can characterize the impact of the WES program on success rates in calculus by saying that no matter how we cut the data--by gender, minority status, engineering status, or prior achievement--the odds that WES students received a B or above in calculus were observed to be about twice that of their non-WES counterparts with a 95% confidence interval for this odds of success ratio of about (1.5, 3.0).

2. **Impact of WES on retention in SMET majors**
   The story here is also quite simple: there was no statistically significant association between persistence in a SMET major or more specifically persistence in engineering, and participation in the WES Math 221 program. That is, retention rates for the various groups were about the same for the WES participants as their non-WES counterparts. In fact, for some groups the retention percentages for WES were actually lower, although not statistically significantly so. Although another study of UW-Madison freshman enrolled in both math and chemistry in their first semester indicated that success in calculus is strongly correlated with persistence in SMET majors (6), if we use 1st semester enrollment in the College of Engineering as a proxy for enrollment in chemistry we find that our WES sample is too small (57 students) to say anything conclusive. In other words, even statistically "significant" (or insignificant) differences should not be interpreted as providing definitive conclusions when only a handful of students is in question. In fact, conclusions drawn from the total WES sample of 169 should also be treated with caution. See Appendix B.

Part A. Synopsis

During the Fall semesters from 1993 to 1996 the UW-Madison Department of Mathematics ran a total of 11 Wisconsin Emerging Scholars (WES) sections distributed over several first-semester calculus lectures. The students considered here were all first-semester freshman with no advanced standing; were 18 or 19 years old; and enrolled in Math 221 in one of those four Fall semesters. (We shall often refer to such cohorts as "entering freshman.") Overall, we compared 169 WES students to 3,871 non-WES students.

Methods

There are numerous ways to assess the impact of a program such as WES (1,2,3,4,5)[1]. Here we consider only the following two, as these are well suited to statistical analysis:

1. "Success" rates in Math 221 and subsequent courses, where we define success as the achievement of a B or above.

2. "Retention" in either the College of Engineering or in a science, math, or technology (SMET) major.

[1] As in other Emerging Scholars programs implemented nationally, WES students attend the same large calculus lecture, do regular homework problems, take the same exams, and are graded in the same fashion as students enrolled in traditional discussion sections. However, instead of enrolling in a discussion section led by a graduate teaching assistant that meets twice a week for 50 minutes, WES students enroll in a workshop that meets for two hours three times a week. Students receive an extra two credits for this time commitment. In each workshop session, the TA provides students with worksheets comprised of carefully designed difficult problems that can best be solved when students work collaboratively. The WES students work in heterogeneous groups of three or four to solve these worksheet problems, while the TA and an undergraduate student assistant circulate around the room asking strategic questions and offering hints when particular groups ask for help or are obviously frustrated. In order to foster a situation in which students turn to their peers rather than instructors for assistance in the problem solving process, the instructors avoid directly answering students' questions.

Workshops are typically comprised of 15 to 18 students of heterogeneous background, including about 50% students from diverse underrepresented minority groups, and 50% white students, largely from rural backgrounds. About half the students in each workshop are women. By contrast, traditional discussion sections enroll up to 25 students, have few if any underrepresented minority students, and have notably fewer women than men students.

The UW-Madison LEAD Center has, in addition to gathering and analyzing quantitative student outcomes data such as those presented here, conducted qualitative evaluation studies of the WES program. Major findings of these studies are that WES provides students a community of learners dedicated to majoring in SME disciplines, which also functions to mitigate the sense of feeling anonymous as a first-semester student at a huge university. That this community is comprised of students of heterogeneous backgrounds is valued by all students, particularly those from underrepresented groups. Compared with students in traditional discussion sections, the WES students showed higher levels of confidence in their mathematical ability, and greater comfort in performing calculus problems, learned to value multiple and creative ways of problem solving, and developed the interest and the ability to acquire a deeper, more conceptual understanding of calculus. Interviews and observations indicate that these student learning outcomes
are a function of primarily three factors: intensive group work experiences, carefully chosen and very difficult problems, and instructors who function as a guide on the side.

Both of these factors were broken-down by several other factors of interest: prior achievement or preparation (e.g., ACT, SAT math scores, UW math placement scores, etc.), calculus lecture, gender, minority status, and whether the student was in the College of Engineering.

The above formulation of the impact of WES in terms of calculus success rates and retention rates is particularly well suited to statistical analyses by the methods of Log-Linear Models. See Appendix A for a comparison of this approach with Multiple Regression and ANOVA methods[2].

Findings

1. Success Rate

The story here is fairly simple: no matter how you break down the student population, whether by prior achievement, calculus lecture, gender, minority status, or whether the student was in the College of Engineering, WES students have higher success rates than their non-WES counterparts. Thus, for example: among female students the overall success rate was 70% for WES versus 57% for non-WES; among under-represented minorities (URM) (URM includes African-, Native-, and Hispanic-American; 90 students all together; 45 in WES and 45 non-WES), the overall success rate was 55% for WES versus 36% for non-WES.

Moreover, the positive impact is similar across each of these categories. A convenient method of quantifying the data to show this similarity is by using the "odds of success." We define the odds of success, for a given group of students, as the ratio of the number of students in the group achieving a B or above to the number of students in the group achieving a BC or below. To compare two groups we simply examine the ratio of the odds of success for each group. We refer to this as the "odds of success ratio." We illustrate the method by comparing the success rates of WES vs. non-WES among the female students, and separately among the URM students.

The percentages of success are:

	Females	**URM**
WES	70%	55%
non-WES	57%	36%

The odds of success are:

	Females	**URM**
WES	2.33 =.70/.30	1.22 =.55/.45
non-WES	1.33 =.57/.43	0.56 =.36/.64

[2] Note that these latter methods construct p-values and confidence intervals by assuming, among other things, normal sampling distributions with equal variances. In contrast, Log-Linear methods are based simply on the distributions that were actually observed, i.e., the actual number of individuals observed in each category of the aforementioned factors. The net result is that p-values and confidence intervals more closely reflect the distributions that were observed, and at the same time, have simpler interpretations.

The odds of success ratios comparing the WES and non-WES students in each category are:

	Females	URM
odds of success ratio	1.76 = 2.33/1.33	2.17 = 1.22/0.56

We can assess the statistical variability in these ratios by hypothetically re-sampling from the observed distributions of success rates for the female and URM students a large number of times, computing these two odds of success ratios each time, and then examining the spread in values, i.e., distribution, of the two odds of success ratios. (This sort of re-sampling, which lies at the heart of the Log-Linear approach, can be referred to as bootstrapping.) For both ratios we estimate that about 95% of the hypothetical re-sampled values will lie between 1.5 and 3.0. We refer to the interval (1.5, 3.0) as the 95% confidence interval for these two odds of success ratios. We emphasize that this method of re-sampling uses exact observed proportions of successes for the various groups, as well as, the exact observed sample sizes, (WES/ non-WES/ female/ URM). By so doing, we take account of the higher statistical variability of smaller sample sizes with only minimal set of assumptions, (e.g., typical regression and ANOVA methods assume re-sampling distributions are normal with equal variances). Thus, although 1.76 is different from 2.17, when we examine the variability of the ratios in this way we find that they are similar in the sense of having similar 95% confidence intervals.

In fact, the interval (1.5, 3.0), or close to it, appeared as a 95% confidence interval for most of the odds of success ratios that we examined that compared WES to non-WES students. Thus, for example, when we break our population down by their ACT math scores we find that the odds of success ratio comparing WES to non-WES is about 2.1 in each ACT category with a 95% confidence interval of (1.5, 3.1).

This positive impact of WES persisted through 2nd and 3rd semester calculus. See Part B for details.

Roughly speaking, we can characterize the impact of the WES program on success rates in calculus by saying that no matter how we cut the data--by gender, minority status, engineering status, or prior achievement--the odds that WES students received a B or above in calculus were observed to be about twice that of their non-WES counterparts, with a 95% confidence interval for this odds of success ratio of about (1.5, 3.0).

2. Retention Rates

The story here is also quite simple: there was no statistically significant association between persistence in a SMET major or more specifically persistence in engineering, and participation in the WES Math 221 program. That is, retention rates for the various groups were about the same for the WES participants as their non-WES counterparts. In fact, for some groups the retention percentages for WES were actually lower, although not statistically significantly so.

This finding is perhaps not so surprising considering that most WES students participated for only one or two semesters of calculus, and that the innovation itself consisted in the replacement of the discussion section component of the calculus course with a cooperative learning approach that consisted in students working on challenging problems in small groups with graduate TAs and undergraduate peer leaders acting as facilitators and “guides on the side." This sort of impact must be compared to the myriad influences that may affect students' career choices over the subsequent four to eight semesters for which we obtained data on these cohorts.

Nevertheless, one might expect that greater success rates in calculus should correlate with greater persistence in science majors, at least among those students who are initially interested in pursuing such majors (cf. 6). If we use first-semester enrollment in the College of Engineering as a proxy for this initial interest we still find no greater persistence among the WES students. However, our WES sample of engineering students was extremely small (15 females, 42 males). Thus, even if the engineering WES students did show higher retention rates it would be ill-advised to draw definitive conclusions, since no matter what the p-values might be, only a handful of WES students are in question. In fact, conclusions drawn from the total WES sample of 169 should also be treated with caution. See Appendix B.

However, this analysis did lead to the remarkable observation that among the calculus students who were in the College of Engineering their first semester, retention rates for URM students were higher than that for non-URM. This was true, both for retention in engineering (60% vs. 56%), as well as for retention in SMET (69% vs. 63%). This occurred despite the fact the URM students tended to have lower ACT math scores than the non-URM students, and despite the fact that the 13 WES-URM students in this group had lower retention rates than the 22 non-WES-URM students. However, none of these differences were statistically significant, which in this case, is a significant finding!

For many retention questions it makes the most sense to look at only those cohorts which have at least two years at the UW. In addition one must also take account of whether a student was in the College of Engineering their first semester, since this is highly correlated to retention in engineering, SMET, and at the university. Unfortunately, examination of the `93 and `94 cohorts revealed that number of URM students in our population was too small to support statistical analyses that take all these factors into account. See Part C for details.

Since the numbers were too small for statistics, we examined the student records individual-by-individual and found no distinguishing features related to retention that separated the WES and non-WES URM students, whether it be cumulative grade point averages (GPAs), number of semesters taking SMET courses, grades in SMET courses, etc. However, this finding of parity between WES and non-WES is obsfucated by the fact that this study included GPAs only for those students who were enrolled in Spring `97, i.e., we were not able to compare GPAs for those who were not enrolled at UW-Madison in Spring `97. For the `93-`94 cohorts this amounted to 9 WES and 5 non-WES URM students.

If we examine retention among the female students in the `93-`94 cohorts, we are confronted with similar problems, as there were only 6 female WES students enrolled in engineering their first semester. However, again we find no distinguishing features related to retention separating these WES and non-WES female students. The 35 female WES students in the `93-`94 cohorts who were not in engineering their first semester comprise a population for which it begins to make sense to consider some percentages as measures of retention:

Table 1

Retention of Female Students in the `93-`94 cohorts who

did *not* begin in Engineering		
	non-WES (N=675)	**WES (N=35)**
not enrolled Spring ‘97	20%	14%
no major	12%	9%
non-SMET major	34%	46%
SMET major	34%	31%

Considering the small sample size (N=35), these retention rates are nearly identical, (e.g., the 12% difference in the non-SMET major category amounts to only 4 WES students). Likewise we find little difference between the cumulative GPAs for these groups.

For the male students we can look at both the populations who began in engineering as well as those who did not. Note the 39% WES-no major figure for males in Table 2 vs. 9% for females in Table 1.

Table 2

Retention of Male Students in the `93-`94 cohorts who did *not* begin in Engineering		
	non-WES (N=737)	**WES (N=28)**
not enrolled Spring ‘97	25%	20%
no major	16%	39%
non-SMET major	28%	11%
SMET major	30%	29%

Table 3

Retention of Male Students in the `93-`94 cohorts who *did* begin in Engineering		
	non-WES (N=518)	**WES (N=24)**
not enrolled Spring ‘97	9%	21%
no major	12%	8%
non-SMET major	7%	4%
SMET major	72%	67%

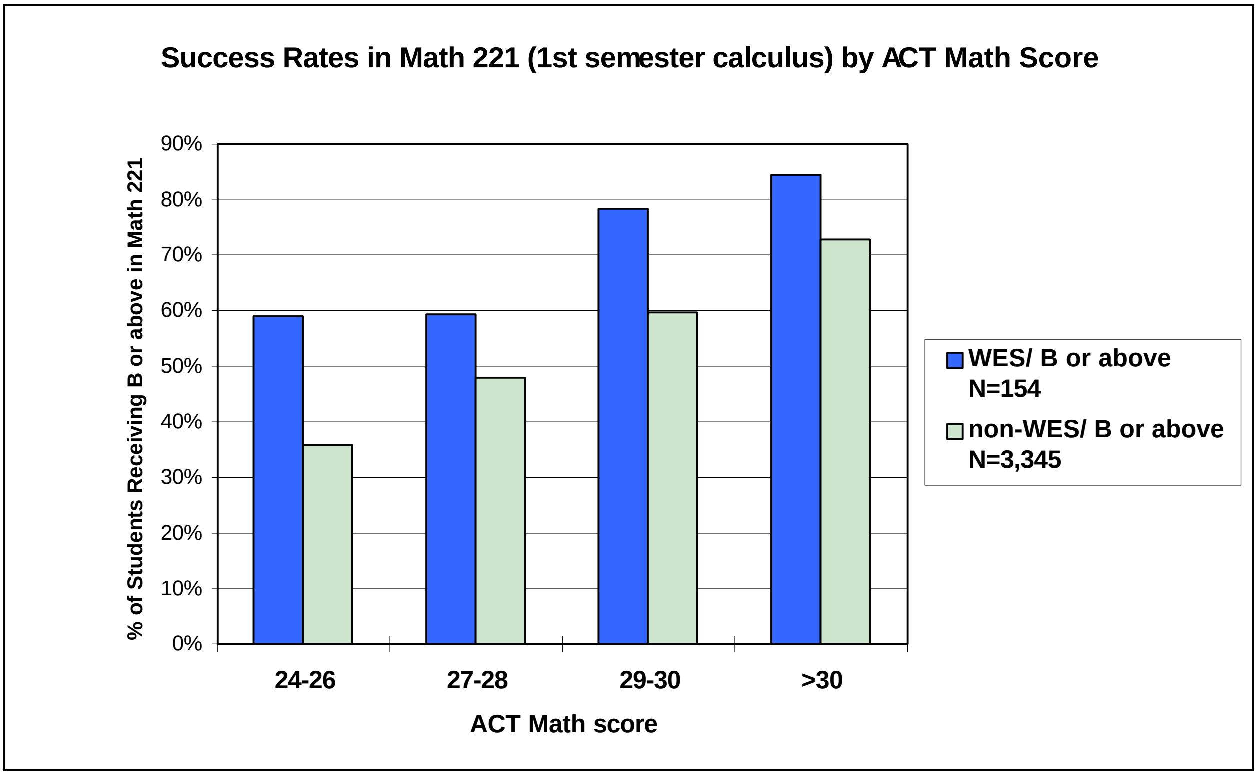
We remark that of the male students who began in engineering, the SMET major category breaks down as 63% engineering majors + 9% non-engineering majors for the non-WES students. By contrast, the 16 WES students, representing the 67% with SMET majors, were all in engineering majors. No remarkable differences between WES and non-WES GPAs were noted for any of these groups. For more details on retention of engineering students, see Part C.

Part B. Details of the Impact of WES on Calculus Success Rates

1. WES effects on calculus grades by prior achievement

We examined several measures of prior achievement including ACT math and science scores, SAT math scores, and UW-Madison math placement scores. In each case we found a similar pattern: WES students had greater success rates no matter the level of prior achievement. Chart 1 below presents the breakdown by ACT math scores. As mentioned in Part A, the students considered here were all "entering freshman," i.e., first-semester freshman with no advanced standing; were 18 or 19 years old; and enrolled in Math 221 in one of the four Fall semesters from 1993 to 1996. Note that ACT scores were not available for all students in the population.

Chart 1



This chart compares the success rates in Math 221 (1st semester calculus) for the four Fall semesters from 1993 to 1996. Entering freshman in WES sections are compared to all other entering freshman taking Math 221.

The chart makes it clear that the WES students had higher rates of success in each ACT math score category. To assess the magnitude of this contrast statistically we can use a generalization of the confidence interval method described in Part A.

As described in Part A, we take as our sampling-distribution (the distribution which generates our hypothetical re-sampling) the exact distribution that was observed, i.e., the exact proportion of the population observed in each ACT category broken-down by WES vs. non-WES and success vs. non-success. As our statistic, instead of the simple odds of success ratio as defined in Part A, we take a quantity (defined below) that measures the "average" contrast between WES and non-WES across all ACT categories. We find that this average contrast is significant at the p=0.0001 level, i.e., only 1 in 10,000 hypothetical re-samplings of this sampling-distribution would produce either a “no contrast” value, or one that favored the non-WES “sample.” Moreover, we also find that the positive effect of the WES program is statistically constant across all ACT categories. To understand this last statement and to understand the magnitude of the positive WES effect (i.e., the confidence interval for the contrast), we need to understand the statistic that measures the average contrast.

Our statistic is defined as follows: for each ACT category we compute the odds of success ratio (or “odds-ratio”) as described in Part A (WES odds of success over non-WES odds of success). (Note: if the non-WES students were just as likely to achieve success as the WES students, then their odds of success would be about the same as the WES students’ odds, i.e., the ratio of the WES students’ odds to the non-WES students’ odds would be about 1.) Rather than add the four odds of success ratios and divide by 4, for technical reasons the geometric mean is more convenient, i.e., we multiply the four odds-ratios and then take the 4th root. This latter quantity is our statistic. When its value is near 1, it indicates that the WES and non-WES students have about the same success rates, and when it is greater than 1 it indicates that the WES students have higher success rates.

The value of our statistic for the sample actually observed is 2.11. This indicates that overall, “averaged” across all ACT categories, the odds that a WES student received a B or above in 1st semester calculus is about twice that of a non-WES student. The 95% confidence interval computed using the observed distribution as the sampling-distribution (i.e., by bootstrapping) is: $1.44 < 2.11<3.09$, i.e., 95% percent of the values of the statistic computed by re-sampling from the observed distribution (a large number of times) lie between 1.44 and 3.09.

Now since our statistic is an average across all ACT categories it does not reveal whether the positive WES effect was the same in each category. This issue can be addressed by modifying the sampling distribution. Namely, we use a sampling distribution as close to the observed distribution as possible, but for which the odds-ratios (WES to non-WES) in each ACT category are equal. We call such sampling distributions “null-distributions” since they are used, in this case, to test the “Null Hypothesis” that odds-ratios in each ACT category are equal.

We now use the so-called “likelihood ratio” statistic to test whether our observed distribution was a likely sample from the null-distribution. If the null-distribution “fits” the observed distribution well, as measured by the p-value of the likelihood-ratio statistic, then the corresponding confidence interval on our odds-ratio statistic will assess the magnitude of the common positive WES effect across all ACT categories.

The results show that the null-distribution with equal odds-ratios in each ACT category approximates, or fits the observed distribution quite well with $p=0.76$, ($p=1.00$ is a perfect fit). This means that 76% of a large number of hypothetical samples from the null-distribution will approximate the null-distribution no better than our observed distribution. In this sense, our observed distribution is a “very typical” sample from the null-distribution. Moreover, our odds-ratio statistic has the value 2.13 for the null-distribution and a 95% confidence interval: $1.48<2.13<3.07$.

In light of the bootstrap confidence interval above we might conservatively estimate that the odds that a WES student receives a B or above in 1st semester calculus is at least 1.5 times that of a non-WES student--no matter their ACT math score.

This was quite a bit of work, but we've learned a lot. Namely, ***the observed contrast between the WES and non-WES success rates was by statistical standards huge—only 1 in 10,000 re-samplings of the observed distribution would have produced an "averaged" odds of success ratio less than or equal to 1. Moreover, the students with lower ACT math scores were helped as much as the students with higher scores. In fact, we can estimate that the odds that a WES student received a B or above was about 1.5 to 3.0 times that of a non-WES student.***

That said, we add one note of caution regarding the assertion of a constant effect across ACT category: since the size of our WES sample ranged from 32 to 46 in each ACT category, the results of the above test should be interpreted as asserting that the variation of the odds of success ratio across ACT category was not more than would be expected with samples of the given size from the null-distribution with equal odds of success ratios in each ACT category, i.e., we flip the same two coins--one for WES, one for non-WES--for each ACT category. (The probabilities of success on the WES and non-WES coins are chosen so as to maximize the likelihood of the observed sample while maintaining a constant odds of success ratio (WES vs. non-WES) across ACT categories.)

We will apply similar analyses to see whether WES had a similar effect on female students as on males, and likewise for minority students, and students in engineering. But first, it makes sense to explore other measures of the positive effects of the WES program irrespective of these other factors.

2. Comparisons within the same calculus lecture

Since our measure of success for the WES program is based the grades received by students in all lectures of calculus it makes sense to check whether the higher success rates of WES students could be accounted for simply by the lecture. In other words, did everyone in those lectures with WES sections receive higher grades? Even better, we can compare the WES students to the students in the same lecture and in this sense control for possible "lecture effects." Charts 2 and 3, below, show two different ways of comparing the students in WES sections to the rest of the students in the same calculus lectures. Recall that our population consists of "entering freshman" as described above from the Fall `93-`96 cohorts.

Chart 2

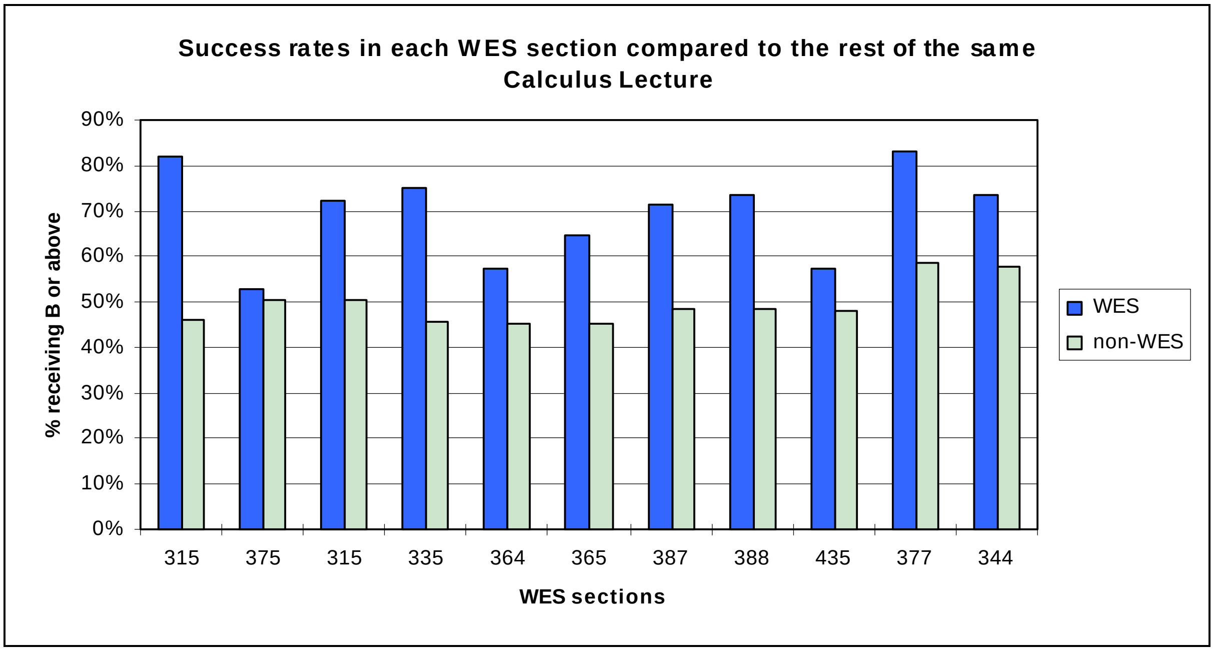
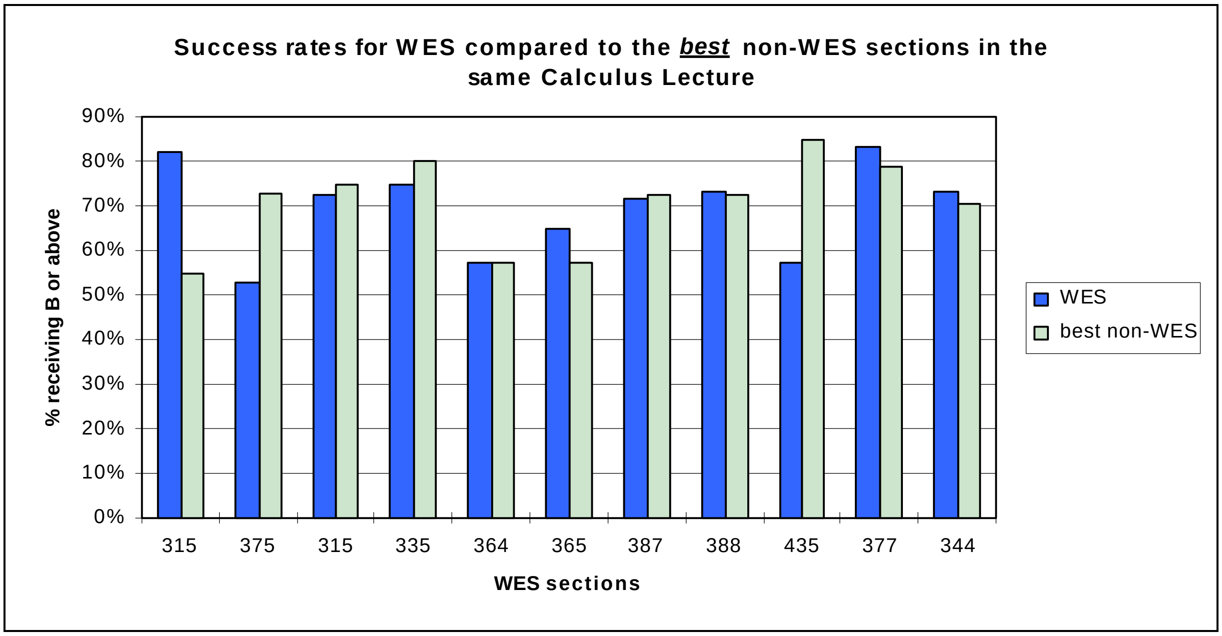


Chart 3



Together, these charts provide strong evidence that ***the students in most of the WES sections were consistently among the best students in the lecture***. In fact, only 2 of the 11 WES sections had success rates noticeably lower than the non-WES section of the same lecture with the highest success rate, and in all cases the WES sections had higher success rates than the rest of the lecture taken as a whole.

Chart 3 suggests that there is likely considerable variation in student success by discussion section. Since, when successfully implemented, the WES method relies less on the “teaching prowess” of individual instructors, the WES program might be considered as a relatively low-cost method of bringing the learning in all discussion sections close to the levels achieved by the “best” TAs.

3. Persistent effects in subsequent math courses

Charts 4 and 5 suggest that the positive effects of the WES program persist. ***The WES students maintain higher success rates in 2nd and 3rd semester calculus (Math 222 and 223)***. Even the WES students who received a BC or below in Math 221 had greater success than their non-WES counterparts. (Note: most of the WES Math 221 students who took Math 222/223 participated in WES sections in Math 222/223.) (The p-values in the charts are analogous to Chi-square p-values, but are computed using Fisher's Exact Test.)

Chart 4

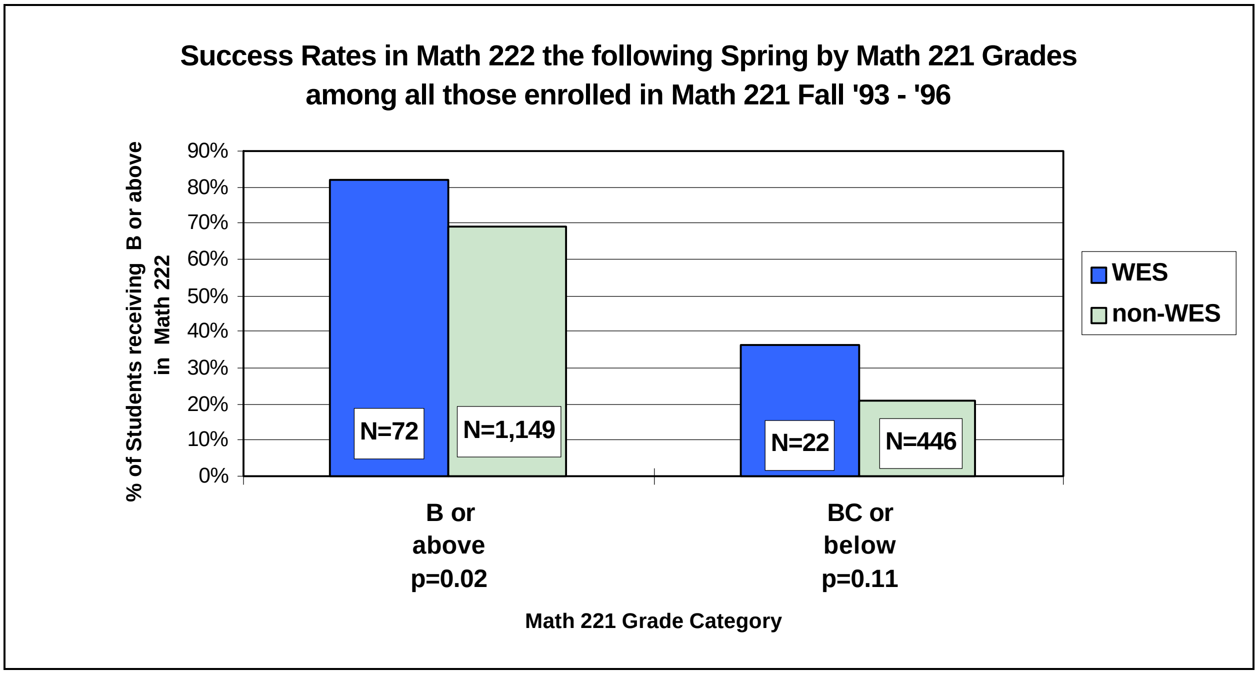
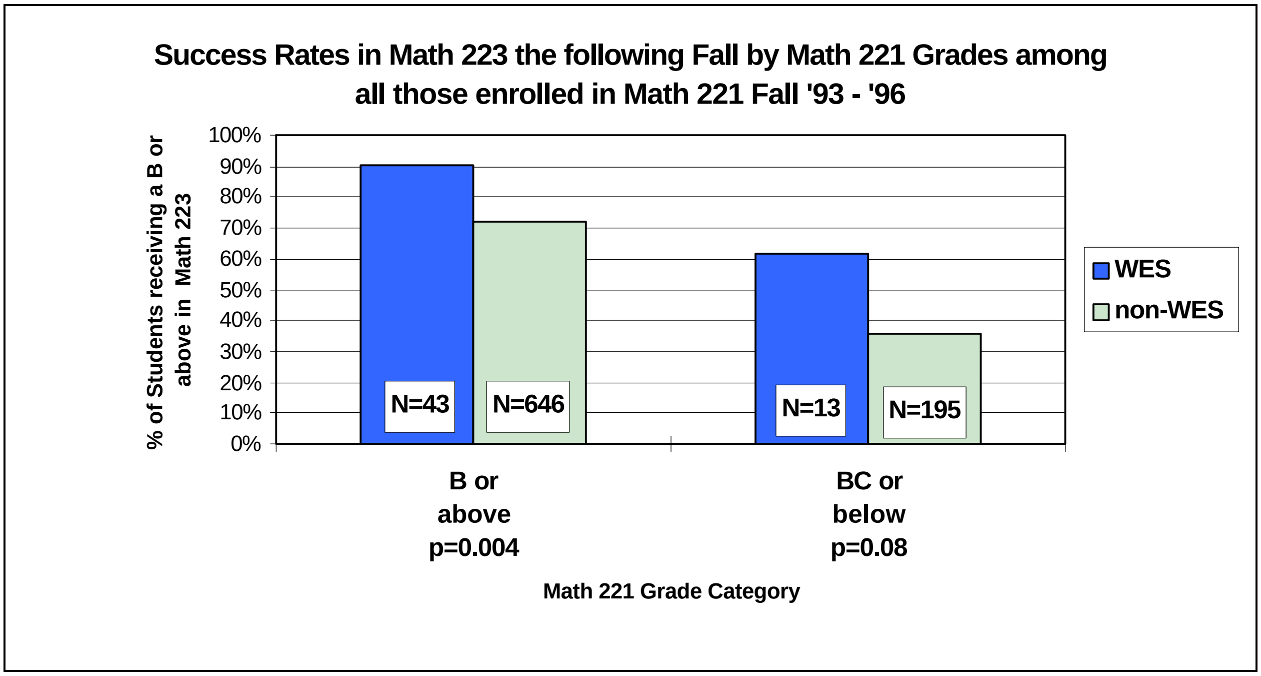


Chart 5



4. Gender effects

Charts 6 and 7 show that the positive effect of WES is about the same for female students as male students. It is interesting to note that overall, female students did slightly better than males in calculus. (The p-values in the charts are analogous to Chi-square p-values, but are computed using Fisher's Exact Test.)

Chart 6

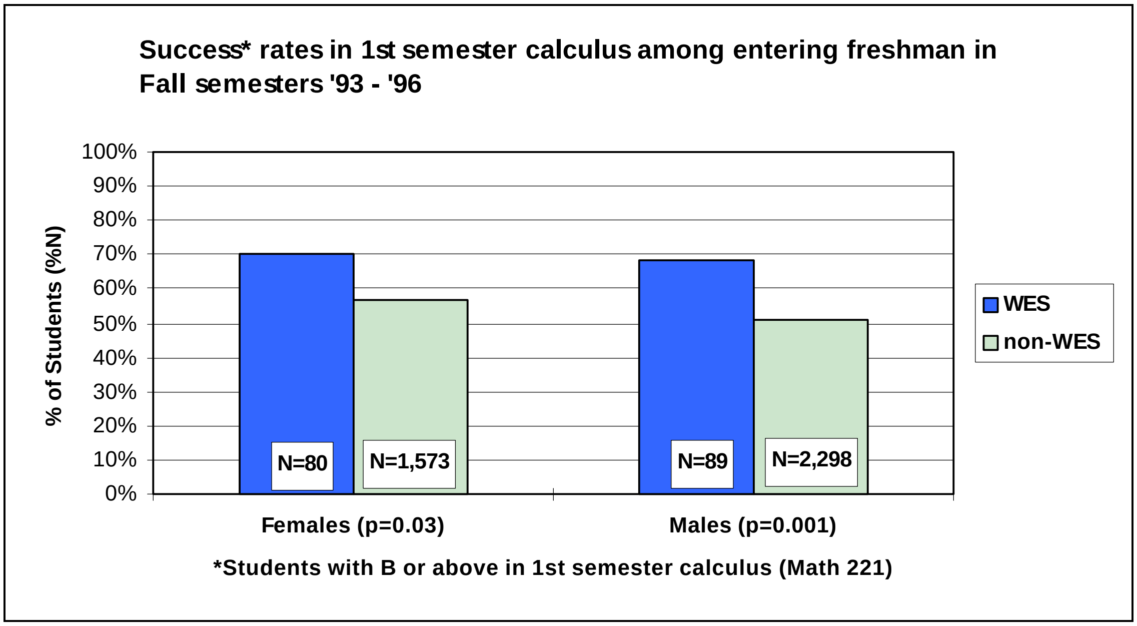
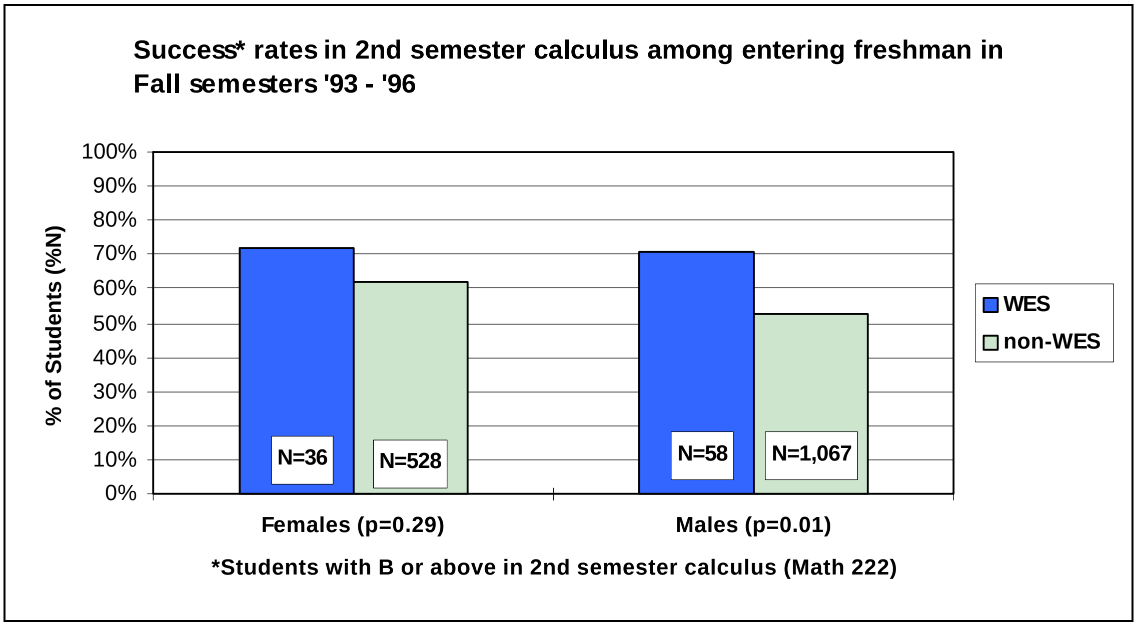


Chart 7



As in section 1, where we determined that the effects of the WES program were "constant" across ACT math scores, we would like to see whether the same might be true across the two gender categories.

The results show that the odds of success ratio measuring the contrast in WES vs. non-WES success rates between females and males is 1.11. That is:

(odds of WES female success)/(odds of non-WES female success)

is 1.11 times that for the males.

Using the observed distribution as our sampling distribution, this is not significantly different from 1.00 at the p=0.56 level, i.e., 56% of a large number of hypothetical samples from the observed distribution would have this ratio less than or equal to 1.00.

As a check, we can change our sampling distribution to the null-distribution defined as being as close to the observed distribution as possible, but also satisfying the constraint that the above odds of success ratios for males and females are equal. We find that this null-distribution approximates the observed distribution quite well, as the p-value of the associated likelihood-ratio statistic is p=0.56,

On the other hand, the contrast between the WES vs. non-WES success rates is virtually identical to what we saw when we broke down the observations by ACT math score. Namely, for the observed distribution our WES vs. non-WES odds-ratio statistic has the value 2.03 and a 95% confidence interval: 1.46<2.03<2.85. These values remain almost identical if, instead, we substitute the null-distribution described above with equal WES vs. non-WES odds-ratios for males and females. All these results are nearly identical to the ACT math score results, as they should be, given that we have determined that the positive WES effect is constant across ACT math score category and across gender.

Thus, as mentioned in Part A, ***the odds that a WES student--male or female-- achieved a B or above in calculus was about twice that of their non-WES counterpart, with a 95% confidence interval for this odds of success ratio of about (1.5, 3).***

5. WES effects on under-represented minorities

Taking account of the ACT math score, the success rate in first-semester calculus for non-under represented minority (non-URM) students is about 1.5 times than that of URM students, but this contrast is significant only at the level p=0.12. With respect to WES participation, both URM and non-URM WES students do better than their non-WES counterparts. Moreover, Charts 8 and 9 show that ***the WES URM students have success rates greater than those of the non-WES, non-URM students***. (The p-values in the charts are analogous to Chi-square p-values, but are computed using Fisher's Exact Test.)

Chart 8

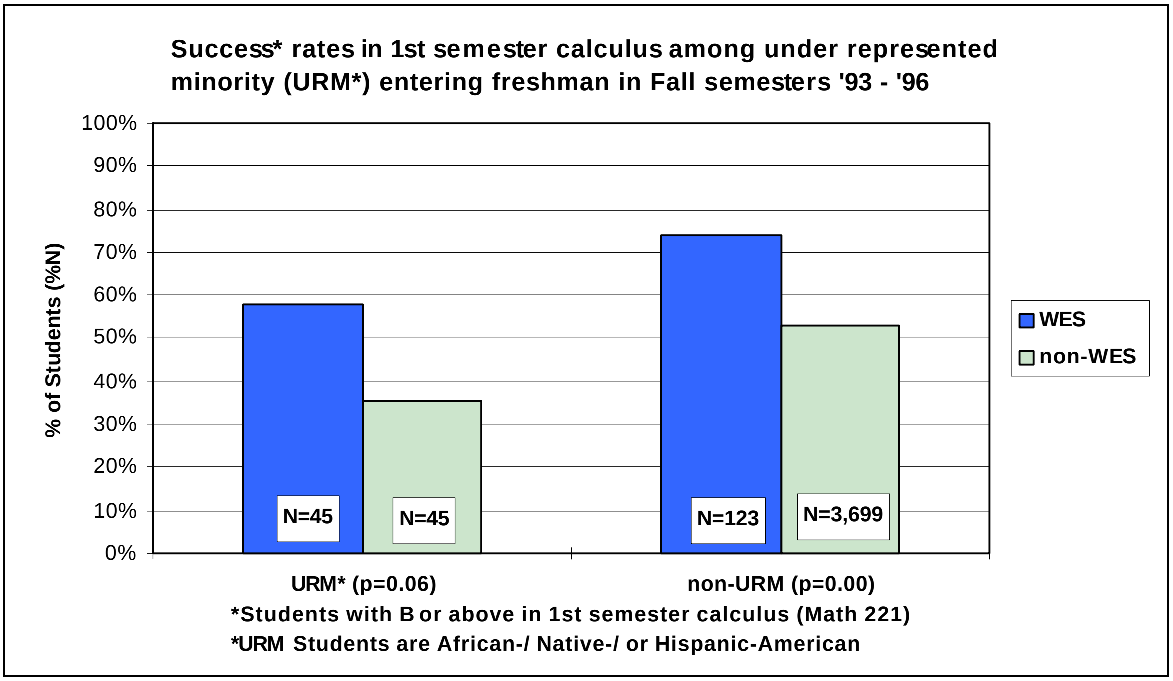
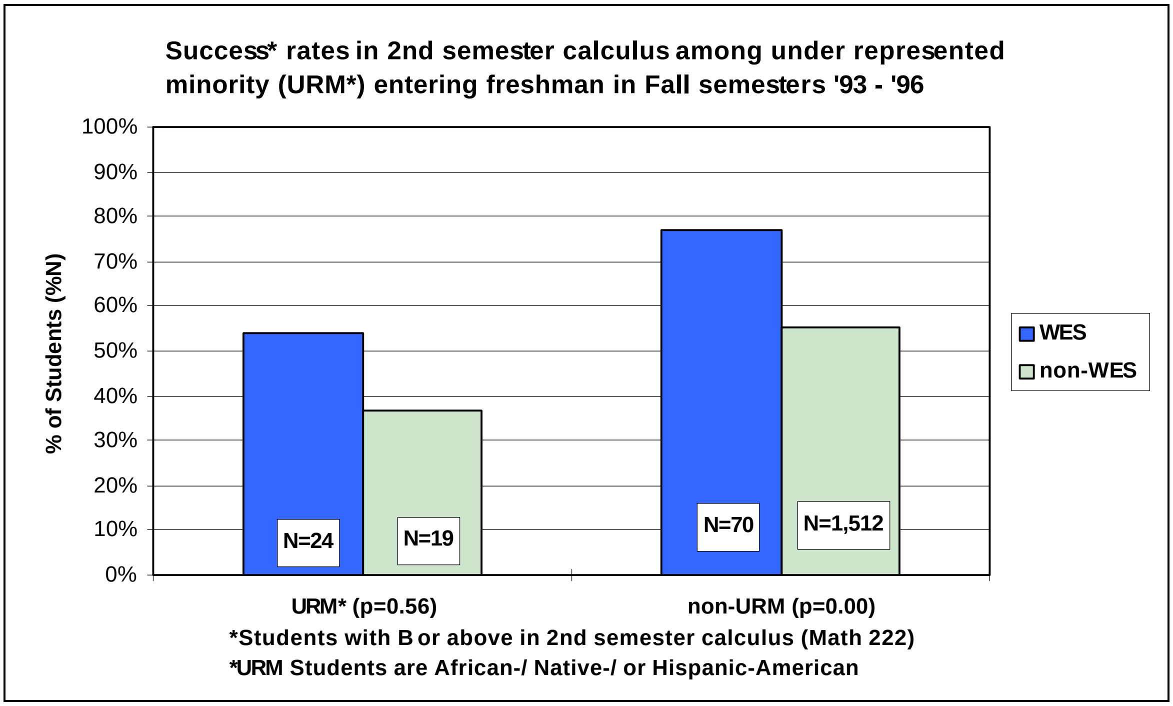


Chart 9



As with gender and ACT math scores, we find that the positive WES effect is constant across these two ethnic categories. The results show that the odds-ratio measuring the contrast in WES vs. non-WES success rates between URM and non-URM students is .90, which is not significantly different from 1.00 at the p=0.84 level. In addition, for the observed distribution, our WES vs. non-WES odds-ratio statistic has the value 2.40 and a 95% confidence interval: 1.42<2.40<4.07. As with the gender categories, these values remain almost identical if we substitute the null-distribution defined by equal WES vs. non-WES odds of success ratios for URM and non-URM students. In fact, the observed distribution fits this null-distribution with a likelihood-ratio p-value of p=0.95.

6. WES effects for engineering students

a. Engineering students overall

Charts 10 and 11 show the contrast in success rates between WES and non-WES-male and female students, who were in the College of Engineering their first semester. (The p-values in the charts are analogous to Chi-square p-values, but are computed using Fisher's Exact Test.)

Chart 10

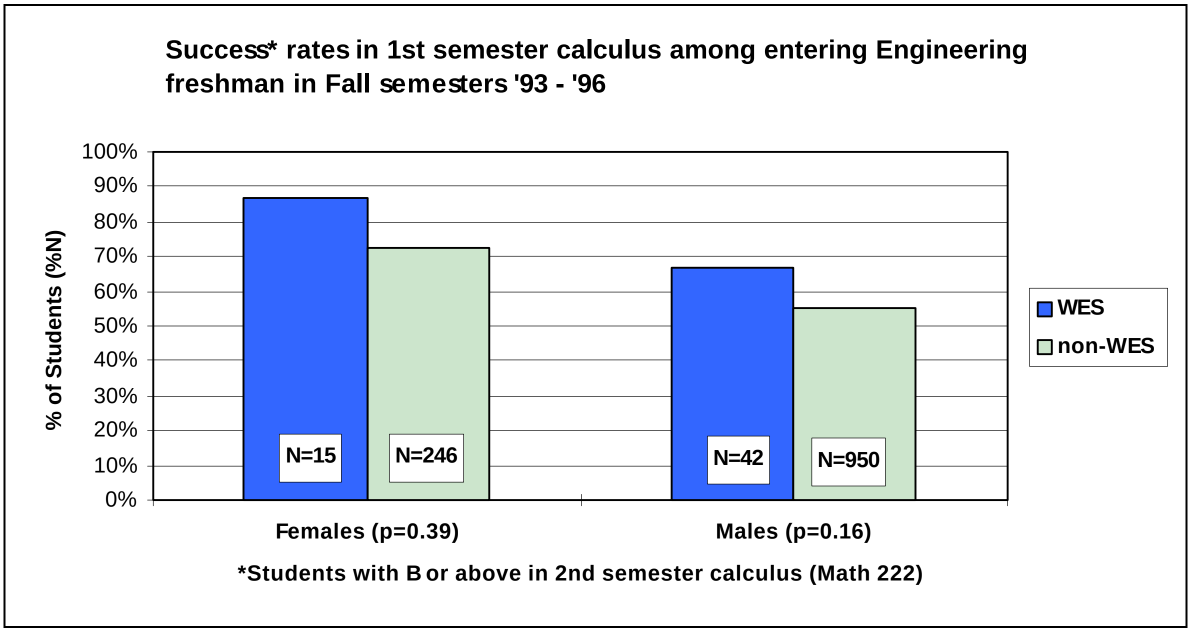
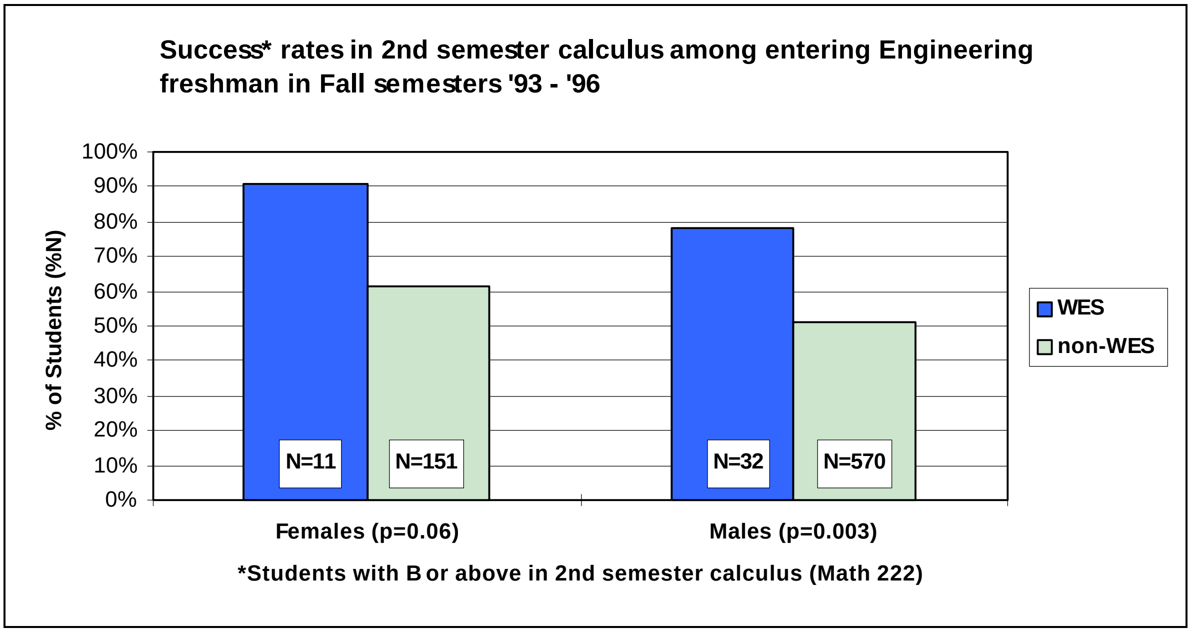


Chart 11



Proceeding as above, we examine the magnitude of the WES effect, and whether the effects were about the same for engineering students as for others. We begin by considering calculus success rates broken down only by a students' WES status and engineering status. For the observed distribution, our WES vs. non-WES odds of success ratio has the value 2.03 and a 95% confidence interval: 1.42<2.03<2.93.

In addition, the odds-ratio measuring the contrast in WES vs. non-WES success rates between engineering and non-engineering students is 0.93, which is not significantly different from 1.00 at the p=0.84 level. As before, these values remain almost identical if we substitute the null-distribution defined by forcing equality of the WES vs. non-WES odds-ratios for engineering and non-engineering students. In fact, the observed distribution fits this null-distribution with a likelihood-ratio p-value of p=0.84. Thus, as with the break down by gender, ***the odds that a WES student--engineer or not--achieved a B or above in calculus was about twice that of their non-WES counterpart, with a 95% confidence interval for this odds of success ratio of about (1.5, 3).***

At this point a word of caution is in order. Dividing our population both by: a) *gender*, WES status, and success status; and by b) *engineering* status, WES status, and success status; and then, in each case, finding a similar odds of success ratios (WES vs. non-WES) is quite different than dividing the population simultaneously by gender, engineering, WES, and success status. It doesn't necessarily follow that the WES effect will be constant across all combinations of the gender and engineering considered simultaneously. Of course, with a sample of only 169, the more we divide the population into smaller groups the less power we have for discerning contrasts specific to these groups. This is the reason for treating the factors of ACT math score, gender, minority, and engineering status separately in the first place, namely, we simply do not have a large enough sample to make statistical inferences about all these factors simultaneously. In the case at hand, we find that our sample size is right at the boundary of being able to resolve some contrasts in success rates specific to the gender and engineering status. The details are considered in the next section.

b. Contrasts between special groups of engineering students and their non-engineering counterparts.

Because the retention of women and ethnic minority students in engineering is of particular interest, we use these methods to see whether WES had any particular effects on these groups among the engineering students in comparison to their non-engineering counterparts.

To this end we consider success rates in calculus broken down simultaneously by gender, engineering status, and WES status. We find that the odds-ratios measuring the contrasts in success rates based on gender or engineering status are 1.01 for both. This value is not significantly different from 1.00 at the p=0.98 level, (using the observed distribution as our sampling model). Thus, ***even when we consider gender and engineering status simultaneously, WES had similar positive effects no matter the combination of students' gender or engineering status.***

Interestingly, there were some contrasts of interest unrelated to students' participation in WES. For these analyses it was deemed prudent to exclude the WES students, as they comprised a comparatively small population with a markedly different calculus experience. Nevertheless, the patterns described below for the non-WES students were also seen among the WES students, although the confidence intervals were quite different due to the small sample size.

In general, we find that engineering students tended do better in calculus than non-engineering students. The odds of success in calculus for an engineering student was 1.7 times that of non-engineering students. However, most but not all of this contrast can be accounted for by the fact that engineering students tend to have higher ACT math scores. When we "average" the contrast across ACT math score categories, we find that the contrast reduces to 1.28 with a bootstrapped 95% confidence interval: 1.09<1.28<1.50. The contrast is significantly different from 1.00 at p=0.002. This mild tendency for engineering students towards higher success rates is statistically constant across ACT math score categories, as the observed distribution fits the corresponding null-distribution with a p-value p=0.84.

However, the picture is not quite so simple. We would like to check whether this positive "engineering effect" is the same for males as females. Thus, we average the contrast over all combinations of ACT category, engineering status, and gender. We find that the observed contrast increases to 1.59 with a bootstrapped 95% confidence interval: 1.31<1.59<1.93 (the p-value is off the scale at p<0.000). Moreover, the contrast for the females is 2.02 times that for the males. (Evidently, this effect is washed-out when the females are lumped in with the males.) Nevertheless, this 2.02 is only significantly different from 1.00 at p=0.23. We suspect that the high p-value may be due to the small number of females in engineering in some of the ACT categories. We can make more efficient use of our observations by using a null-distribution which sets the positive engineering effects to be equal across the ACT math score categories, while allowing them to be different for males and females. For this sampling distribution the engineering effect for females is 2.51 times that for males with a p-value of p=0.10. This null-distribution fits the observed distribution fairly well with a likelihood-ratio statistic of p=0.56. In addition, the corresponding 95% confidence interval for the overall engineering-effect is virtually identical to the bootstrapped interval mentioned above.

Thus, if we restrict the variability of our sampling distribution across ACT categories, but not across the engineering status, and gender categories, ***the observed contrast in the engineering-effect between males and females appears to be more likely the result of a greater engineering effect for females than males***. Nevertheless, any conclusions one might wish to draw regarding this contrast must remain tentative as the null-distribution which sets this contrast between males and females equal fits the observed distribution with p=0.38, showing that a "null contrast" should not be ruled out.[3]

We can proceed analogously to check for contrasts between the engineering minority students and their non-engineering counterparts. Unfortunately, the small sample size prevents all but the most tentative suggestions. For example, one observed contrast shows that the positive WES effect was 3.81 for non-engineering URM students, while only 1.23 for engineering URM students. This disparity was due to both a lower odds of success for URM engineering WES students compared to URM non-engineering WES (0.86 vs.1.67), as well as a higher odds of success for the URM engineering non-WES students compared to the URM non-engineering non-WES (0.69 vs. 0.44). The ratio 3.81/1.23=3.10 measures the size of this contrast, indicating that the size of the WES effect for non-engineering URM students was three times that for engineering URM students. Nevertheless, due to the low number of minority students involved, this contrast was significant only at the level p=0.22. Unlike in the case above where we could improve our ability

[3] This is a good example of nature of the "quasi-duality" between the Confidence Interval and the Hypothesis Testing methods. Namely, the notable difference between p=0.10 and p=0.38 reminds us that these two p-values estimate two distinct probabilities. Typically, the duality becomes sharper the smaller the p-value.

to detect contrasts between males and females by choosing a sampling distribution which nullified other contrasts (i.e., the “ACT-effect” on success rates), the contrast 3.10 above considers how all of the factors under consideration (engineering status, URM status, and WES status) correlate with success in calculus.   Thus, we do not have the option of nullifying the contrasts in one factor so as to enhance our ability to detect a contrast in another. One way to assess our inability to detect contrasts in success rates which are defined in terms of two or all three of these factors is to choose a null-distribution which nullifies all such two- and three-way contrasts. We find that the observed distribution approximates this null-distribution at the level p=0.60.

Part C. Details of the impact of WES on Retention rates in SMET

The lack of remarkable contrasts between WES and non-WES retention rates obviate the need for some of the more sophisticated statistical analyses that were applied to quantify the rather notable contrasts in calculus success rates above. Instead, we simply provide several different “cuts” of the data as regards some of the issues related to retention of students in science and engineering.

Assessing retention in SMET fields is a tricky business for students who enter in colleges or schools which do not consist entirely of SMET departments, since it is difficult to impossible to determine whether a student was interested in pursuing a SMET major in the first place. This difficulty occurs certainly for the College of Letters and Sciences (L&S), as well as the College of Agricultural and Life Sciences (ALS), (e.g., departments such as Family Resources, Agricultural Economics, and Agricultural Journalism offer some majors that probably should not be considered SMET majors).

We first look at retention among students who began their first semester in the College of engineering (COE). For our WES sample, 34% began in COE compared to 31% of the non-WES sample. We then modify our approach to consider the non-engineering students. The bulk of the remainder of the WES sample began in the Colleges of L&S and ALS (47% and 15%, respectively, vs. 52% and 8%, respectively, for the non-WES sample).

1. Retention among engineering students

Charts 12-14 below show the distribution of majors as of Spring `97 of all students in the study whose first semester at the UW was in COE. In the cases where the number of WES students in some categories was too small we show only the non-WES population and simply comment that the distribution of the WES students was similar. Also note that the "drop*" category refers to students not enrolled at the UW-Madison Spring `97. (Note: the reader is advised against drawing *any* definitive conclusions about the impact of WES on retention, since no matter what the differences are and what the p-values might be, only a handful of WES students are in question.)

Chart 12

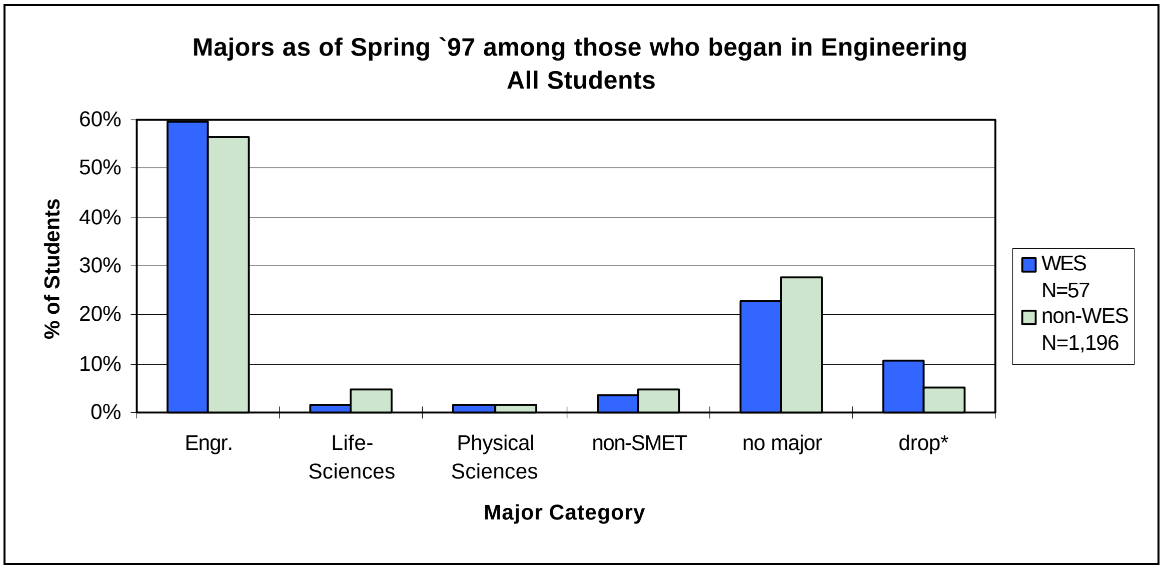


Chart 13

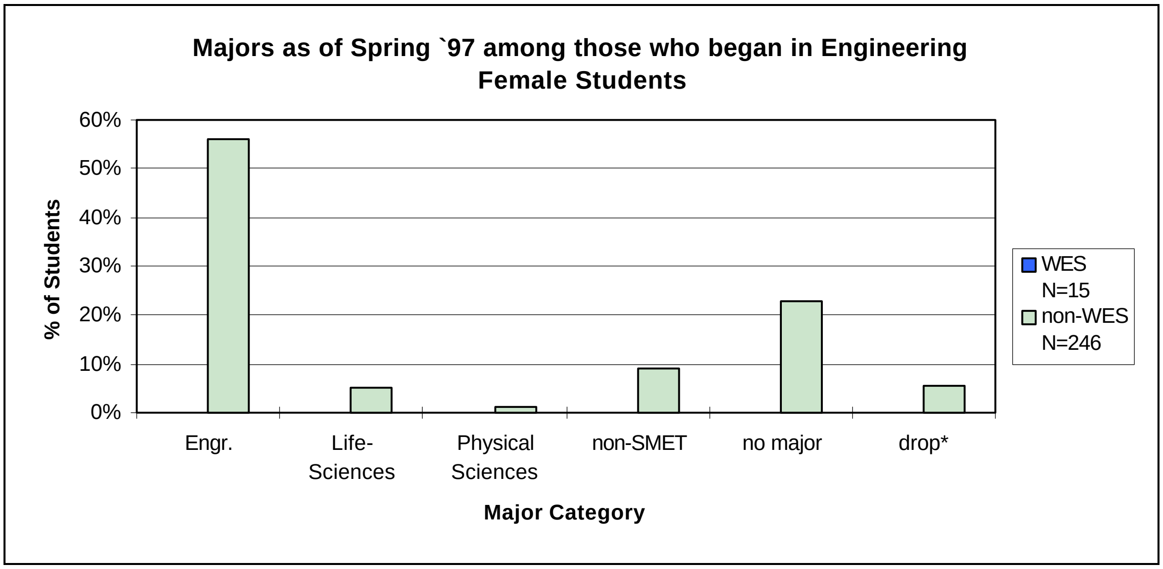
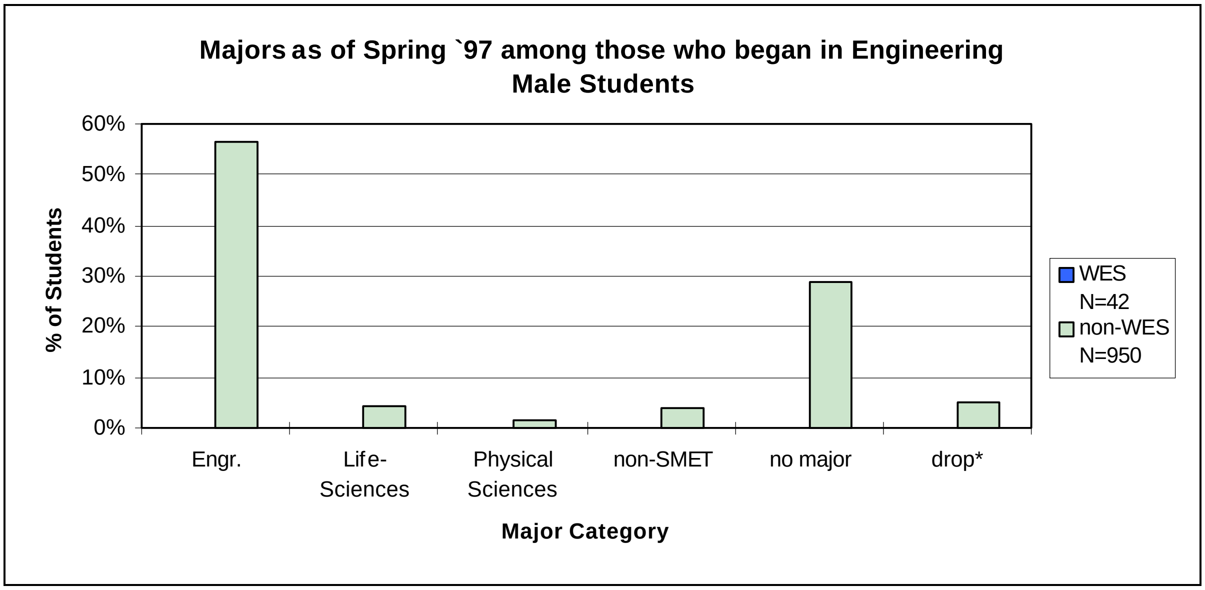


Chart 14



By combining these major categories into “engineering vs. others,” and “SMET vs. others,” we can at least consider some Chi-square statistics. These are provided in Charts 15 and 16 below. Note the male-female parity in retention irrespective of participation in WES.

Chart 15

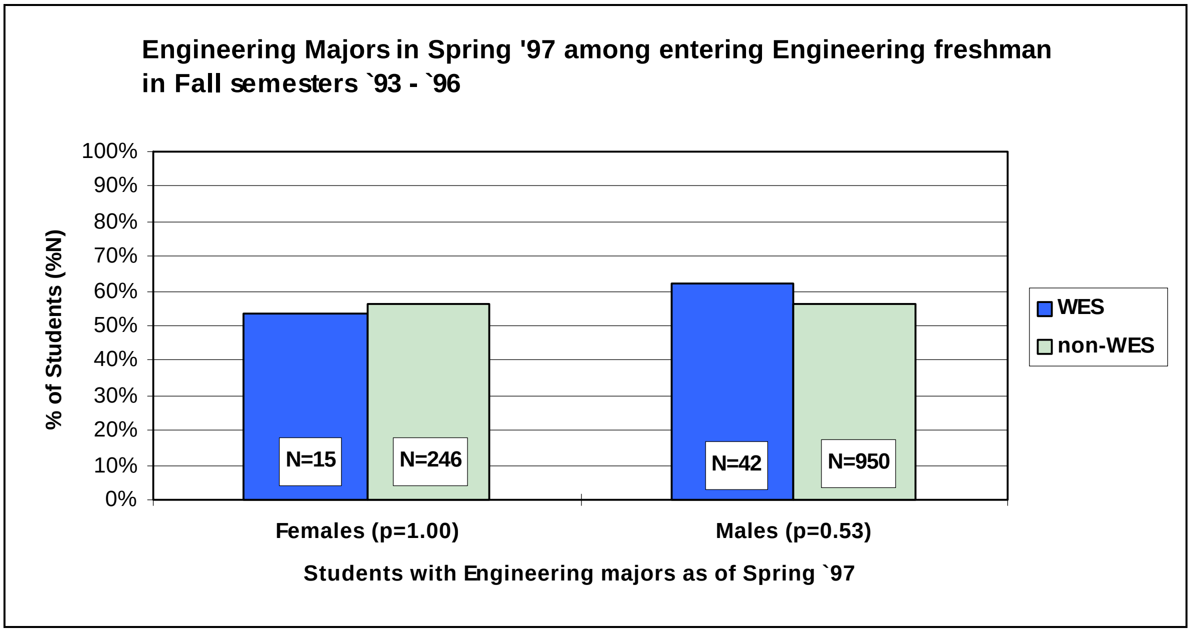
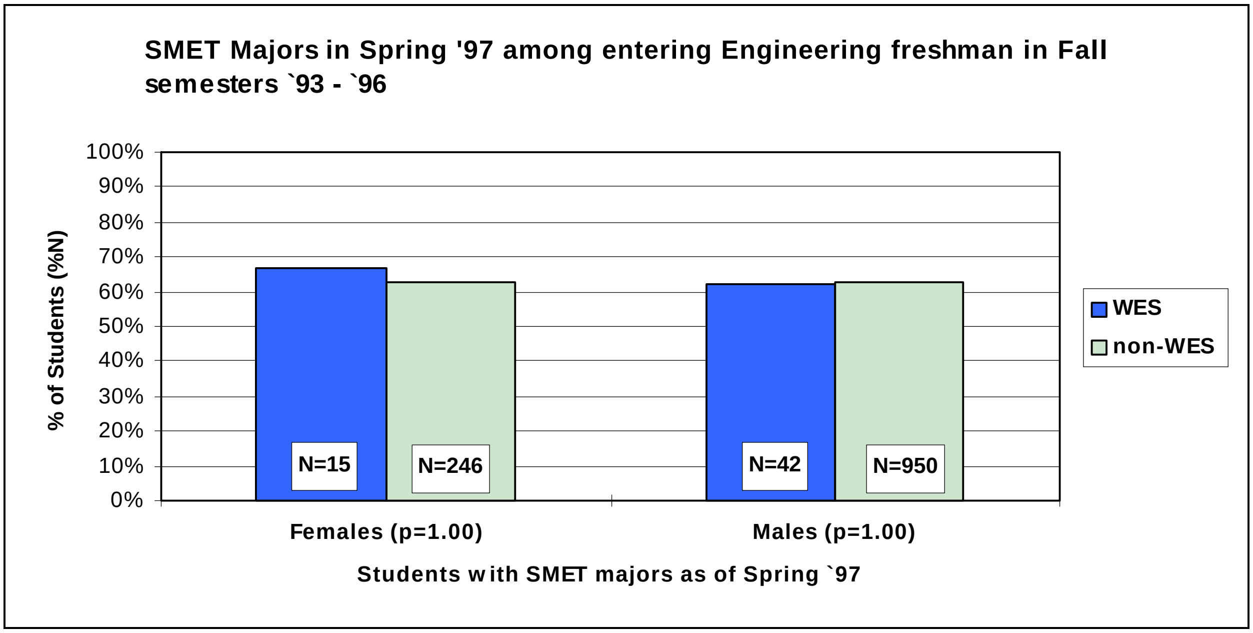


Chart 16



Charts 17 and 18 provide the analogous information for the URM students in engineering. Note that the sometimes large percentage differences do not give rise to significant p-values, since these differences reflect only a handful of students. Thus, the lower retention rates for the WES students should not be construed as a negative impact of the WES program on retention.

Chart 17

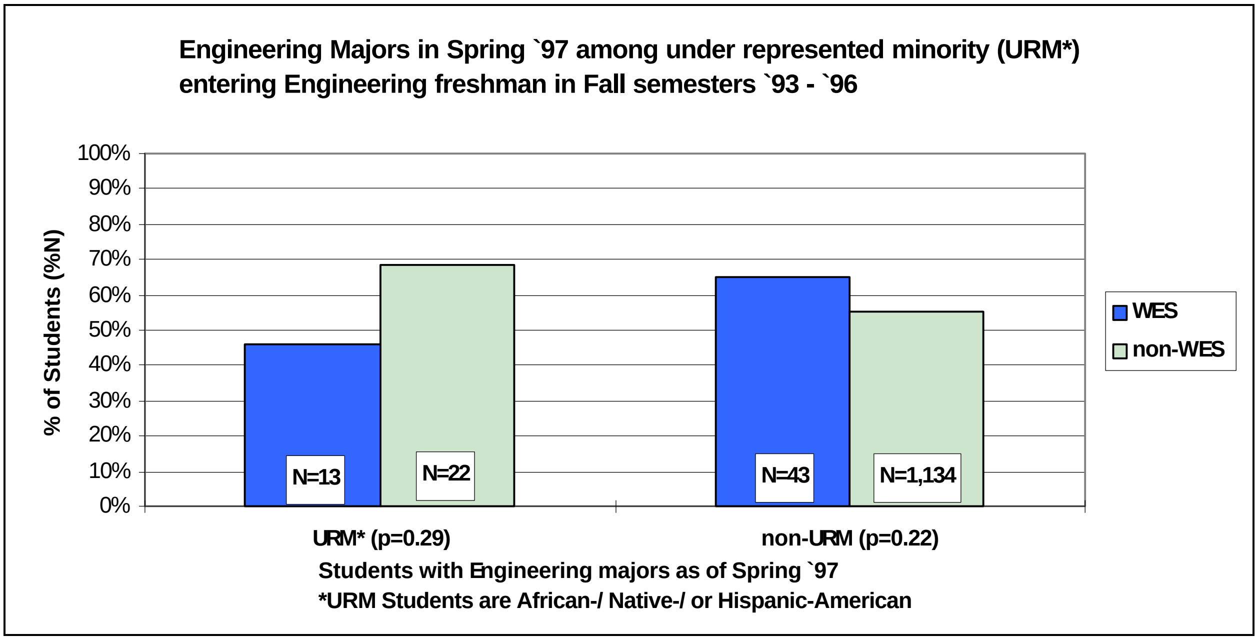


Chart 18

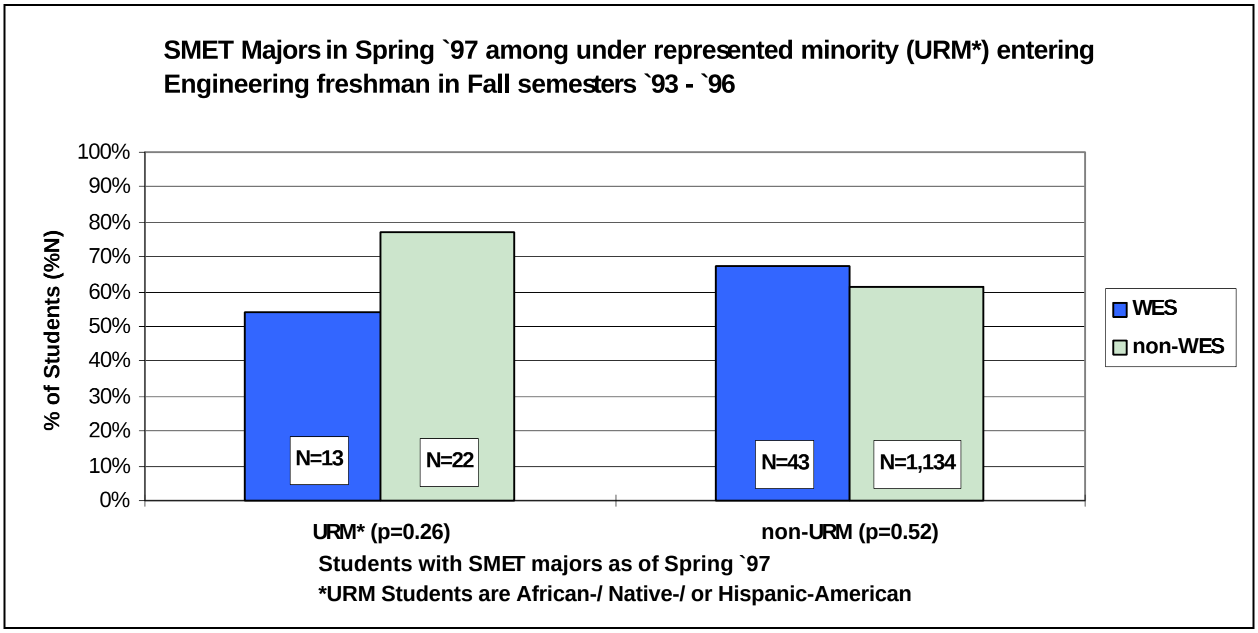
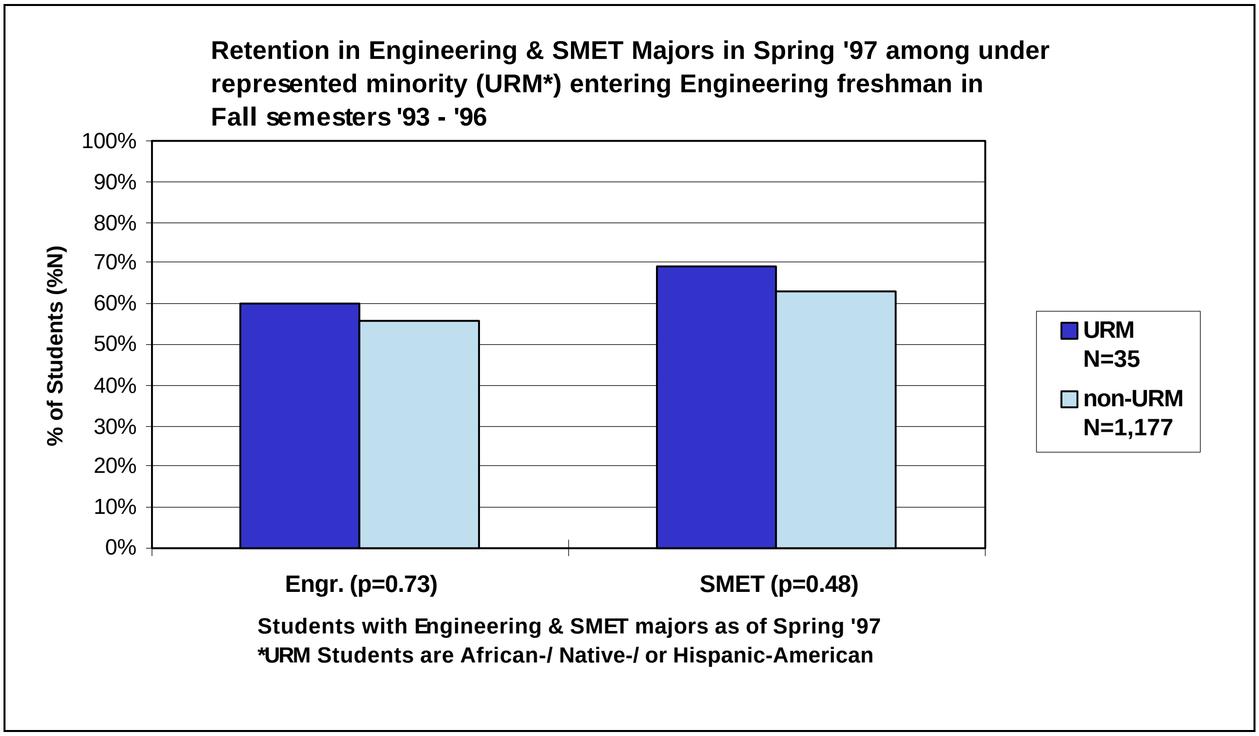


Chart 19 shows that irrespective of participation in the WES program the URM students in our population have as high a retention rate in engineering and SMET majors as do the non-URM students. Note that, in this case, the fact that the p-values show no “significant” difference is quite significant!

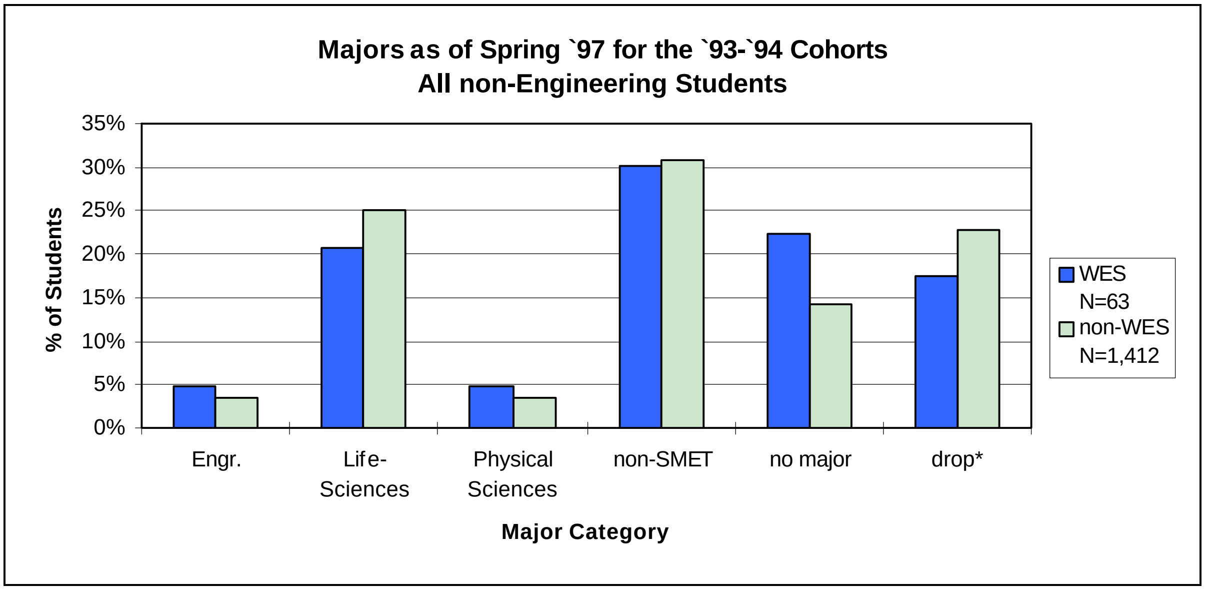
Chart 19



2. Retention among non-engineering students

For the non-engineering population we look at only the `93-`94 cohorts, as the more recent cohorts still have at least 50% of the students with no major declared. We present only one chart depicting the break down analogous to Chart 12 for all engineering students. The male-female break downs are shown in the Tables 1-3 in Part A. As there were only 18 WES and 11 non-WES URM students we present no chart, but comment only that the two distributions were similar. Also note that, as above, the "drop*" category refers to students not enrolled at the UW-Madison Spring `97. (Note: the reader is advised against drawing *any* definitive conclusions about the impact of WES on retention, since no matter what the differences are and what the p-values might be, only a handful of WES students are in question.)

Chart 20



Appendix A. Comparison of Statistical Methods

When assessing the impact of a program on student performance in a course one naturally wants to take some account of the incoming abilities, and prior achievement. This is especially true when the number of participants in the program (sample size) is small. Thus, for example, it would hardly due to compare the WES students to the rest of the students taking calculus if the vast majority of the WES students had exceptionally high ACT math scores. The simplest approach would be to compare WES and non-WES students with similar ACT math scores or placement scores. This is the approach we follow here. The advantages are obvious, we compare apples to apples, and oranges to oranges. The disadvantages are equally obvious: when comparing two baskets of fruit the sample size of the apples and oranges will be smaller than that of the total. Another disadvantage, not so obvious, is a fact of life known as the Bonferroni inequality which implies that the more statistical tests you do the more likely you are to find a difference that is "significant." This phenomenon implies that if you toss a coin with a 5% chance of "heads" four times the chance of seeing at least one head is estimated by Bonferroni to be no more than 20% (4 x 5%)[4]. In the case of two baskets of four different kinds of fruit, it says that even if both baskets of fruit were sampled from the same population we would have a 20% chance to see at least one of the types of fruit showing a significant difference at the 5% level. Nonetheless, both of these difficulties can be accommodated by the use of efficient statistical methods, (in this case Log-Linear Models), without sacrificing the sound idea of actually seeing the comparison of "apples to apples."

This approach lies in stark contrast to the commonly employed method of multiple regression which seeks to maintain a large sample size for the comparison of, say, the grades of the experimental and control groups by "adjusting" them for ACT math scores etc. Here, the "adjusted" grade is supposed to be interpreted as the grade that the student would have received had they had, say, the average ACT math score, i.e., the idea being that we can think of the entire population as having the same ACT math score, (by "sliding" each observation parallel to the regression line to some fixed ACT score), so we haven't reduced our sample size. Of course, by another way of thinking we have actually reduced the sample size to zero since no one actually received an adjusted grade. This is complicated by the fact that the adjusted grade is a highly non-linear function of the entire set of grades. In addition, in the case of linear regression, one assumes that the grade depends linearly on ACT math score for both experimental and control groups, whereas, in the case of the WES program, we have particular interest in whether the method has the same effect on "weaker" students' performance as it does for "stronger" students.

All these problems arise before the stage of estimating and interpreting "p-values," for which the method has a host of deficiencies, not the least of which is the assumption of "Normal" sampling distributions. Namely, one more accurately considers grades as a giving a discrete distribution on an ordinal scale (i.e., A is better than B etc.), which distinguishes them significantly from any continuous distribution on a linear scale (i.e. 4.0 is 33% better than 3.0), a fortiori, from a Normal distribution. The interpretation of p-values from Normal theory tests depends crucially on the fact that after any "normalizing" or "variance stabilizing" transformations are made: 4.0 = 1.33 x 3.0, as well as on the "fiction" that we will observe grades in every interval between 3.0 and 4.0 (e.g., from pi =3.14 to 3.45) with their appropriate Normal probability.

[4] Of course, the exact value is $1-(0.95)^4 = 0.185 < 0.20 = 4 \times 0.05$. The inequality itself: $1 - (1-p)^n < np$, $0<p<1$, with approximate equality for p small, stems from the days before calculators when it was a time-saver.

The manner in which Normal Theory p-values depend on the specific characteristics of the Normal distribution is complex: effectively it involves how convergence rates for the Law of Large Numbers and the Central Limit Theorem are affected if one assumes that the independent observations all have normal distributions. Generally speaking, these convergence rates are faster for Normal distributions than others. This implies that if you assume that all your samples have normal distributions with the same variance, "significant" p-values can be expected from smaller sample sizes, in comparison to dropping these assumptions. Thus, one can expect that the p-values obtained from Normal Theory tests will overstate observed differences. While this feature of Normal Theory tests can be useful for detecting marginal differences, by the same token it can be misleading when trying obtain conservative estimates of confidence intervals and p-values. Thus, the Normal Theory p-value presents a quandary: it may be too small, but by how much?

Although use of Analysis of Variance (ANOVA) methods address some of the deficiencies of multiple regression ANOVA still assumes an underlying Normal distribution of grades (with equal variances) and so the aforementioned problems with interpreting p-values and confidence intervals remain.

Of course, Log-Linear Models do not eliminate the problems associated with interpreting p-values and confidence intervals, but when the statistical problem can be prudently stated in terms of mult-nomial distributions, the relevant interpretations are usually easier to understand. As an example, let's consider a regression approach to assessing the impact of the WES program while controlling for ACT math scores. Ignoring that grades are more of an ordinal than a numerical measure, we could have regressed (the numerical version of) calculus grades on ACT math scores for the WES and non-WES groups and asked whether the regression lines were parallel and/or had different intercepts. Often, some sort of transformation of the data is required to obtain something that looks like normal residuals with equal variance. However, note that the assumption that *some* normal distribution "fits" (some transformation of) your data, and the assumption that you re-sample from this distribution are quite different, the latter being a much more stringent condition. In addition, one also needs to address the common occurrence of correlation among the residuals, which can substantially overstate significance levels.

To interpret the relevant p-values and confidence intervals we consider: re-sampling the (transformed) WES and non-WES samples from the fitted regression lines with residuals having normal distributions with equal variances; computing the highly non-linear slopes and intercepts each time; and then examining their distributions. In the end, we might be able to say that the distributions of regression slopes were roughly the same for WES and non-WES, but that the distribution of intercepts (adjusted grades) was greater for the WES than the non-WES. In any case, one is still inclined to ask: "What does it all mean? … what exactly is an adjusted grade, and what about that correlation in the residuals?"

In contrast, the Log-Linear approach described in Part B, section 2 shows that we need only consider re-sampling from either the observed distribution, or from a nearby null-distribution where contrasts in success rates are set equal across ACT categories. In either case, the p-values and confidence intervals involve only the ratios of successes to non-successes, for WES vs. non-WES, across the ACT categories. For example, the p-value for the test of equality of the odds of success ratio (WES over non-WES) across ACT categories says, quite simply, that 76% of the samples from the aforementioned null-distribution will show variation across ACT categories at least as great as the observed sample. Thus, the Log-Linear approach provides a clear and succinct response to the question of whether the observed variation across ACT categories is more than one would “expect by chance,” given the sample sizes. The reason for this is precisely because the phrase “expect by chance” has been given an easy to understand interpretation; namely, re-sample from a distribution close to the observed but with equal odds of success ratios across ACT categories.

By comparison, the regression approach leaves one wondering what “expects by chance” means at all, i.e., one expects some residuals, (all assumed normal with equal variances), to be larger, some smaller, but how a given residual affects the regression slope or intercept is complicated by the highly non-linear dependence of these statistics on the entire data set. Thus, a question such as whether the observed variation in adjusted grade is more than one would “expect by chance” has a complex meaning.

In contrast, with the Log-Linear approach, the question of whether the variation in our observed odds of success ratio across ACT categories is more than one would “expect by chance” means precisely that we compare our observed odds with what we would see if we flipped the same two coins (one for WES; one for non-WES) for each ACT category. (The probabilities of success on the WES and non-WES coins are chosen so as to maximize the likelihood of the observed sample while maintaining a constant odds of success ratio (WES vs. non-WES) across ACT categories.)

Finally, note that with neither approach is it advisable to imagine that the re-sampling scheme approximates actually repeating the experiment (with real students). This is a major reason why there are lies, damn lies, and statistics. For more on this point see Appendix B. below.

Appendix B. A word of caution on p-values and sample sizes

To inspire the reader not to over interpret “significant” p-values in the presence of moderate sample sizes we consider two hypothetical populations of students one of size 170 and the other of size 4,000. We suppose that both populations have been spilt into two groups (successes and non-successes), and that the larger group has a 50% success rate (2,000 successes). Obviously, if the smaller group is split 85 to 85 the Chi-square p-value will be p=1.00, but how many additional successes would we need to get a “significant” p-value? It turns out that a 100 to 70 split gives a Fisher’s Exact two-sided p-value of p=0.03, whereas a 105 to 65 split yields a p-value of p=0.003. Thus, the difference between “highly significant” p-values and “highly insignificant” p-values amounts to the successful or non-successful endeavors of only a handful of students (15 to 20). We are reminded that, in this case (roughly speaking), the p-value is telling us only that if we consider a large number of repetitions of the experiment of flipping a fair coin 170 times: only 3% of these repetitions will lead to 100 to 70 split or worse, and only 0.3% of the repetitions will yield a 105 to 65 split or worse. The point is that p-values tell us about the known range of behavior of flipping coins (or sampling from Normal distributions). This may or may not tell us about the range of behavior of students taking calculus. Put another way, if we are measuring body temperatures a sample size of 170 students would be huge, because we have a good idea of the known range of body temperatures. Thus, for example we would likely find it quite significant if we found 15 of the 170 students with body temperatures below 96 degrees Fahrenheit. In contrast, we should be much more cautious about the significance of an additional 15 to 20 successful calculus students above our hypothetical “norm” of a 50% success rate.

References

1. Millar, S. B., Alexander, B. B., and Lewis, H. A., *Final Evaluation Report on the Pilot Wisconsin Emerging Scholars Program, 1993-94*, LEAD Center Report, 1995.

2. Alexander, B. B., Burda, A. C., and Hwang, Y., *Final Evaluation Report of the Math 222 Wisconsin Emerging Scholars (WES) Pilot Program, Spring 1995,* LEAD Center Report, 1995.

3. Alexander, B.B., Millar, S.B., and Lewis, H.A., *Evaluation of the Pilot Wisconsin Emerging Scholars Program, 1993-94, an Audiocassette Program*, LEAD Center Product, 1995, 100 minutes.

4. *Script of the Audiocassette Program, Evaluation of the Pilot Wisconsin Emerging Scholars Program, 1993-94*, LEAD Center Product, 1995.

5. Alexander, B. B., Burda, A. C., and Millar, S. B., "A Community Approach to Learning Calculus: Fostering Success for Under represented Ethnic Minorities in an Emerging Scholars Program," *Journal of Women and Minorities in Science and Engineering,* vol. 3, no. 3, 1996.

6. Kosciuk, S. A., “Initial Observations on ‘Gateway’ factors for Science Students at UW-Madison Based on Longitudinal Student Records,” Draft Manuscript available from the LEAD Center 1996.