A Meta-Analytic Review of College Students' Gains in General Education

Dena A. Pastor & Pamela K. Kaliski

James Madison University

Brandi A.Weiss

University of Maryland

Paper presented at the 2006 meeting of the
American Educational Research Association, San Francisco, CA.

THE CENTER FOR
ASSESSMENT &
RESEARCH
STUDIES

MSC 6806
Harrisonburg, VA 22807
540.568.6706 Phone
540.568.7878 Fax
assessment@jmu.edu
www.jmu.edu/assessment

A Meta-Analytic Review of College Students' Gains in General Education

Many institutions of higher education require all students, regardless of major, to take a pre-specified set of courses during their first several years in college. These courses are typically called *general education* or *core education* courses with general education being defined by Gaff (1991) as the common knowledge, skills, and developmental characteristics of the college educated person. Proponents of general education argue that these courses serve not only as a fundamental basis for a liberal arts education, but also ensure that students are exposed to material that will enable them to be educated citizens. A survey of a national sample of colleges and universities in 2000 indicated that the median general education requirement is 40% of the typical baccalaureate degree (Ratcliff, Johnson, La Nasa, & Gaff, 2001). Since a large proportion of a students' undergraduate education at institutions with this requirement is composed of general education courses, it is important to understand what impact these courses have on a students' knowledge and skills.

Although the Ratcliff et al. (2001) survey indicated that only 32% of institutions assess the effectiveness of general education programs, the authors note that this percentage is likely to rise given the increasing demand for accountability from state legislatures and accrediting bodies. According to Banta, Lund, Black and Oblander (1996), institutions that do engage in the assessment use a variety of different methods to evaluate gains in students' general education skills and knowledge. Some institutions simply ask students what kind of skills and knowledge they feel they have gained, whereas other institutions rely on more direct measures of student learning, such as course-embedded tests or portfolios. Regardless of which type of assessment is used, it is a good idea to acquire some baseline measure of what students know and are able to do prior to any college coursework. In contrast to only collecting information from students after completing the general education curriculum, collecting measures on students before (pretest) and after (posttest) their completion of the curriculum allows greater confidence in claiming that the change in scores is attributable to the program (Erwin, 1990). In other words, obtaining pretest and posttest measure provides more meaning to scores in that it allows one to quantify the value added by the general education curriculum.

If such a repeated measures design is used, there are a variety of different ways that the results can be conveyed. The most straightforward approach would be to report the pretest and posttest average scores, with the difference between the averages representing the typical change in raw scores over time. A disadvantage of this approach is its dependency on the particular score scale being employed. For instance, a typical gain of 5 points appears large on a 20-point scale, but negligible on a 50-point scale. For this reason, it is desirable to report scale-free measures of change. Scale-free measures of an effect are often conveyed using effect sizes, which are typically used to capture the practical significance of an effect. Readers may be familiar with the effect size known as Cohen's *d*, which provides a scale-free measure of the typical difference between groups. When the two means being compared are from the same group of people at different time points, different definitions of Cohen's *d* can be used to capture the difference between means or the change over time (Morris & DeShon, 2002). We elaborate more fully on these different definitions later in the paper.

There are several benefits associated with the use of effect sizes to represent the typical change over time in students' general education skills and knowledge. For instance, consider a mature assessment program, in which the same assessment data has been collected for several years. To obtain as accurate an estimate as possible of the program's effectiveness, (i.e., one that is based on a large sample) one may consider combining the data across years. This approach, however, requires that the data was properly archived and is accessible. It is more likely that only descriptive statistics of the results are available, perhaps in assessment reports. If this is the case, effect sizes can be calculated and a statistical procedure known as meta-analysis can be used to not only average the effect sizes, but to determine whether the effect sizes significantly differ from one another. If significant variation exists, meta-analysis can then be used to explore the extent to which certain variables are related to the effect size estimates. For instance, if effect sizes were collected before and after substantial program improvements, meta-analyses can be used to determine if effect sizes were significantly larger after the changes were made.

Purposes of the Current Study

Because the repeated measures design is encouraged in general education assessment (Erwin, 1990) and used extensively at our university, the first purpose of the present paper is to inform readers about the different definitions of Cohen's *d* when using repeated measures designs. We illustrate how to calculate, interpret, and decide among the various definitions, relying heavily on the information and suggestions provided by Morris and DeShon (2002). The second purpose of our paper is to illustrate the various ways in which effect sizes from a mature general education assessment program can be utilized. To this end, we use assessment data that has been collected at our university from five different cohorts to examine the effectiveness of the American Experience general education program. We decided to focus on this program since the learning objectives, courses and assessment instrument have remained relatively the same for the previous five cohorts.

To assess the objectives of the American Experience program (listed in Appendix A), the same faculty created multiple-choice instrument, the American Experience Test (AMEX), has been administered since August of 2000 to students as incoming freshmen (pretest) and again as second semester sophomores (45-70 credit hours; posttest). To provide information about student cohorts, an assessment report is provided each year to the program faculty. This report contains the descriptive statistics of the pretest and posttest scores for groups of students having different requirement completion statuses at the time of posttest. Specifically, each year descriptive statistics are reported for six groups: 1) students who have not completed the requirement (None) and five groups of students who have completed the requirement by 2) using AP credit (AP), 3) using transfer credit (TR), 4) completing GPOSC 225 (POSC), 5) completing GHIST 225 (HIST), or 6) completing both GPOSC 225 and GHIST 225 (Both). From each cohort report, the descriptive statistics for each of these six groups was collected resulting in a total of 30 effect sizes that were used to answer the following sets of research questions:

Set 1: Estimating the typical change over time and determining if there are significant differences in change over time.

1a. What is the average effect size? That is, what is the average gain made in American Experience knowledge and skills during students' first year and a half in college?

1b. Are there significant differences among the effect sizes? That is, do the effect sizes differ greatly from one another or are they all fairly similar?

Given significant variation among effect sizes, a second set of research questions was pursued.

Set 2: Exploring factors that may explain differences in change over time.

2a. Are there significant differences among the effect sizes associated with the various cohorts? (e.g., are students who were freshmen in 2003 showing larger gains in knowledge and skills than students who were freshmen in 2000?)

2b. Are there significant differences among the effect sizes associated with the various requirement completion statuses of students?

Research question 2b is important in that it allows us to capture the extent to which gains in American Experience knowledge result from maturation alone. This can be accomplished by comparing the effect size of the "control" group, which consists of students who had not yet completed their American Experience requirement by the time of posttest, to the effect sizes associated with other "treatment" groups. We are also able to compare the efficacy of different treatments by comparing the effect sizes associated with groups completing different courses at our university to one another and to those associated with groups of students who used AP or TR credit to fulfill this requirement.

Methods

We first describe how and from whom assessment data is typically collected at our university followed by a description of the assessment reports from which the information used in the meta-analysis was obtained. Second, we describe the two different effect sizes used in repeated measures designs and provide various ways to calculate, interpret, and decide among the two definitions. Third, the hierarchical linear modeling (HLM) approach to meta-analysis used in the present study is described with particular attention paid to the various model specifications that were used to answer each of the above research questions.

Procedure & Samples

James Madison University is a 4-year public university in the mid-Atlantic with ~ 15,000 undergraduate students and ~1,000 graduate students. All undergraduate students, regardless of major or professional program, are required to complete the general education program. The purpose of the American Experience program is to provide students with "…an understanding of the major themes and concepts that structure American life today". The learning objectives for this program are shown in Appendix A. One of two courses can be used to fulfill the requirement for the American Experience program: GHIST 225: U.S. History or GPOSC 225: U. S. Government. A description of these two courses can be found at: http://www.jmu.edu/gened/cluster4.html.

It is hoped that all students have their American Experience general education requirement completed by the end of their sophomore year and about 70% of students actually have this requirement completed prior to the second semester of their sophomore year. Also, students have the option of fulfilling this requirement if they scored above a 4 on one of two Advanced Placement exams (United States History or Government and Politics: United States); or, if allowed by the program coordinator, they can fulfill this requirement by transferring credit from completion of similar courses at another university. Table 1 shows the percentage of students in each status completion group by cohort. Completion status percentages were fairly similar across cohorts with the majority of students at the time of posttest having completed the requirement at our university by taking GHIST 225 (40%), GPOSC 225 (14%) or both courses (2%). A sizeable percentage of students had not completed the requirement (29%) at the time of posttest. About 10% and 5% of students fulfilled the requirement through AP and transfer credit, respectively.

Our university assesses the impact of the general education curriculum by using a repeated measures approach to assess the gains that are made in students' general education knowledge and skills. Two days are set aside each year, in mid-August and mid-February, for assessment and students' course registration is blocked if they fail to participate. To obtain a sense of what students' knowledge and skills are coming into college, a representative sample of incoming freshman are administered assessment instruments before classes start during an institution-wide Fall Assessment Day in August (pretest). In the Spring Assessment Day (posttest), which takes place in mid-February, a large sample of students having 45-70 credit hours (typically 2nd semester sophomores) is tested. Assessment Days are coordinated by the Center for Assessment and Research Studies (CARS), the assessment office at our university.

Because students are assigned to complete a set of assessment instruments using the last two digits of their student identification number, the sample administered any particular instrument is a random sample from the particular student cohort being tested. Thus, incoming freshmen completing the AMEX in Fall of 2000 are representative of the entire incoming freshmen class for that year. The pretest and posttest data collection dates for the five cohorts used in the present study are shown in Table 2.

Measure

The American Experience test (AMEX) is an 81-item multiple-choice test developed in 1998 to assess the objectives of the American Experience domain (see Appendix A). The AMEX was created in 1998 by faculty teaching the courses needed to fulfill the American Experience general education requirement. The items on the AMEX are scored right (1) or wrong (0) with the sum of the item scores representing the total score. It takes students, on average, 60 minutes to complete the AMEX. Coefficient alpha for the AMEX scores has been consistently high across cohorts and data collection dates (see Table 2).

Assessment Reports

Staff at CARS analyze the data after each Assessment Day and provide a report of the results to general education faculty. In order to provide scores by requirement completion status, student assessment data is linked to university records containing such information. Assessment reports for the American Experience domain, which were kept on file at CARS, were gathered for the five cohorts used in the present study. Although there is a wealth of information provided in the report, we only acquired for each status group the sample size, pretest and posttest means and standard deviations, as well as the correlations between pretest and posttest scores. The resulting data set is shown in Table 1.

Analysis

Effect Size

Definitions. The effect size of interest in the current study is the standardized mean difference, which is commonly used as a measure of practical significance when comparing two means. The standardized mean difference is often denoted as *d* to represent the sample statistic and δ to represent the population parameter. Although there tends to be agreement in the current literature as to how to define δ when using averages from between-subjects or independent groups (IG) designs, there is little agreement as to how to define δ when using averages from within-subjects or repeated measures (RM) designs. Regardless of which design is employed, the numerator of the standardized mean difference is the difference between means. In IG designs, the difference is between group averages (e.g., $\mu_{treatment} - \mu_{control}$) and in RM designs the difference is between pretest and posttest averages (e.g, $\mu_{post} - \mu_{pre}$). The denominator of δ differs for the two designs with IG designs using the pooled within-group standard deviation and RM designs using either the standard deviation of the gain scores (a.k.a. change or difference scores) or the standard deviation of the pretest or posttest scores. According to Gibbons, Hedeker, and Davis (1993), if the standard deviation of the gain scores (σ_{gain}) is used as the denominator, the resulting definition for δ is

$$\delta_{gain} = \frac{\mu_{post} - \mu_{pre}}{\sigma_{gain}}. \quad (1)$$

If the standard deviation of the pretest or posttest scores is used in the denominator, Dunlap, Cortina, Vaslow and Burke (1996) define δ as

$$\delta_{raw} = \frac{\mu_{post} - \mu_{pre}}{\sigma_{raw}}. \quad (2)$$

For the moment, consider σ_{raw} as the pooled pretest and posttest standard deviation. Although δ_{gain} and δ_{raw} are both appropriate effects sizes to use in RM designs, they are not on the same scale and therefore can neither be meaningfully compared nor combined. Specifically, δ_{gain} is on the change score metric, while δ_{raw} is on a raw score metric (Morris & DeShon, 2002). The definition and thus the metric of the standardized mean difference in RM designs has important implications not only for how the effect size is interpreted and estimated, but also how the sampling variance of the effect size is computed.

Interpretations. To illustrate the differences in interpretation of δ_{gain} versus δ_{raw}, consider a situation where they both equal 0.5. A value of 0.5 for δ_{gain} implies that the typical change in scores is half a standard deviation above zero. Morris and DeShon (2002) add that if a normal distribution of the change scores is assumed, a δ_{gain} of 0.5 would indicate that 69% of the change scores from pretest to posttest in the population are positive. Alternatively, when δ_{raw} is employed the interpretation is more familiar to those using the standardized mean difference effect size; a value of 0.5 for δ_{raw} implies that pretest and posttest means differ by half of a standard deviation unit.

If the numerators of Equations 1 and 2 are equal, the difference between the effect sizes will depend on the value of the correlation (ρ) between pretest and posttest scores. To illustrate, consider the following formula to compute σ_{gain}

$$\sigma_{gain} = \sigma_{raw}\sqrt{2(1-\rho)}. \quad (3)$$

As noted by Morris and DeShon (2002), only when $\rho = 0.5$ will σ_{gain} equal σ_{raw}, which results in equal values of the effect sizes on the various metrics. However, when $\rho > 0.5$, σ_{gain} will be larger than σ_{raw} and when $\rho < .5$, σ_{gain} will be smaller than σ_{raw}. Thus, δ_{gain} and δ_{raw} often do not provide the same values. Even when the two definitions yield the same value (e.g., $\rho = 0.5$) their interpretations differ.

Because δ_{gain} and δ_{raw} are not interchangeable, researchers must decide to use one definition over the other. In making this decision, Morris and DeShon (2002) recommend taking into consideration the research question being posed. If the research question focuses on comparing various treatments, then the raw score metric should be used; if the research question focuses instead on individual change over time, then the change score metric is preferred. However, Morris and DeShon (2002) also state that oftentimes "the same research question could be framed in terms of either metric" (p.111) and thus encourage researchers to also consider the interpretability of the effect size and the extent to which ρ varies across studies. If ρ varies across studies, Morris and DeShon (2002) suggest either using: (a) δ_{raw} or (b) δ_{gain} with subsets of effect sizes having similar ρ.

Although the type of research questions being asked in the current study supports the use of the change-score metric, we decided to use δ_{raw} as opposed to δ_{gain} for two reasons. First, we favored δ_{raw} since we believe our stakeholders are more likely to be familiar with its interpretation and second, we wanted to use an effect size that would not require us to split our data into subsets having homogeneous ρ.

Estimators. Having made our decision as to which definition to use, we then had to decide upon which estimator of δ_{raw} to employ. If we assume that pretest and posttest standard deviations in the population are equal to a common value (σ_{raw}) and therefore can be pooled, then we could choose from either d_{raw} computed using sample descriptive statistics or from the dependent samples *t* statistic (see Column A of Table 3). A simulation study comparing these two estimators favored the latter of the two estimators, particularly with large values of *r* (Dunlap, Cortina, Vaslow & Burke, 1996). The authors of this study indicate, however, that there is little difference between the two estimators when the sample size is large (n>50).

Although the estimators in Column A of Table 3 are used to directly estimate δ_{raw}, the estimators of δ_{gain} in Column C can also be used after a simple conversion to the raw score metric. The estimators of δ_{gain} result in effect sizes on the change score metric and can be converted to the raw score metric using a formula similar in form to Equation 3

$$\delta_{raw} = \delta_{gain}\sqrt{2(1-\rho)}\,. \tag{4}$$

In the current study, we decided to use d_{raw} computed using descriptive statistics because: (a) the resulting effect size is on the raw score metric (thus, the Equation 4 conversion is not required) and (b) only descriptive statistics for the pretest and posttest scores were available in the reports (the reports do not provide either the dependent samples *t* statistic or the standard deviation of the gain scores). Because the samples in our study are so large, d_{raw} computed using descriptive statistics should also be very similar to the estimator favored by Dunlap et al. (1996), which is d_{raw} computed using the *t*-statistic.

Becker (1998) suggests using the pretest standard deviation (SD_{pre}**)** in the denominator of the d_{raw} estimator since it is not affected by the treatment and thus more likely to be similar across studies. For this reason we chose to use the d_{raw} estimator based on descriptive statistics in Column B of Table 3. For the remainder of the paper we simplify our notation by referring to this estimator as d_j for effect size *j* and the corresponding population parameter as δ_j.

Sampling Variance. The sampling variance of d_j captures the accuracy with which d_j estimates δ_j and can also be thought of as the extent to which d_j varies due to sampling error. A linear model can be used to represent the relationship between d_j and δ_j

$$d_j = \delta_j + e_j \,, \tag{5}$$

where e_j represents sampling error, which is the discrepancy between the population parameter and the sample estimate. The variance of e_j (σ^2_{ej}) is the sampling variance of the estimator.

Another purpose for using the pretest standard deviation (SD_{pre}**)** in the denominator of d_j is because a precise estimator for the sampling variance exists for this particular formulation. When the pooled standard deviation is used in the denominator, as in Column A of Table 3, the estimator has an unknown sampling distribution and accordingly, an unknown sampling variance (Morris & DeShon, 2002). A general form of the sampling variance for the estimator used in the present study is provided by Morris and DeShon (2002)

$$\sigma^2_{e_j} = \left[\frac{2(1-\rho)}{n_j}\right]\left(\frac{n_j-1}{n_j-3}\right)\left[1+\frac{n_j}{2(1-\rho)}\delta^2\right]-\frac{\delta^2}{c_j^{\,2}} \,, \tag{6}$$

where n_j refers to the sample size (because we are using a repeated measures design, sample size is equal to the same value at pretest and posttest) and c_j refers to the bias function equal in this design to

$$c_j = 1-\frac{3}{(4n_j-4)-1} \,. \tag{7}$$

The bias function is used to correct for the bias associated with the use of small samples, which tend to overestimate the population effect size.

Researchers can choose to use estimated values of ρ and δ particular to each effect size j in Equation 6, or use common estimated values of these parameters, calculated by pooling estimates of ρ_j and δ_j across effect sizes. In the section following the description of our approach to meta-analysis, we describe our process for deciding between these two alternatives.

Meta-Analysis

In meta-analysis, effect sizes from various studies all examining the same research question are analyzed with the intent of estimating the average effect size across studies, the extent to which effect sizes differ from one another and to explore variables that can be used to explain why they differ. A hierarchical linear modeling (HLM) approach to meta-analysis was used in the current study with d_j serving as the dependent variable in all models. Readers interested in the use of HLM for meta-analysis should consult Raudenbush and Bryk (1985; 2002).

A series of hierarchical linear models having two levels were used in the present study. For all specifications, the Level 1 model is equal to Equation 5, which captures the relationship between the sample estimate of a particular effect size j and its corresponding population parameter. Because the Level 1 variance (σ^2_{ej}) is already known (calculated in this study using Equation 6), this Level 1 model is unlike most other Level 1 models in HLM applications. The population parameter for each effect size j in the Level 1 model is then used as the dependent variable in the second level of the model. In the following section we describe the various Level 2 model specifications (all using δ_j as the dependent variable) that were used to answer the research questions.

Unconditional Model. To answer the first set of research questions, an unconditional model was utilized to estimate the typical effect size and to capture the extent to which effect sizes vary. The population parameter for each effect size *j*

$$\delta_j = \gamma_0 + u_j \,, \tag{8}$$

is modeled at Level 2 as being a function of the grand mean (γ_0) and error (u_j), which captures the deviation of δ_j from γ_0. The variance of u_j (σ^2_{uj}) represents the extent to which population effect sizes differ from the grand mean. The significance of σ^2_{uj} (that is, whether the Level 2 error variance is significantly different from zero) was used to answer research question 1b, which asks whether there are significant differences among the *J* effect sizes, with *J* equaling the number of effect sizes in the study.

To assess the significance of σ^2_{uj}, we compared the fit of the model in Equation 8 against a model imposing the restriction that σ^2_{uj} equal zero. For the remainder of the paper, the former model is referred to as the random-effects unconditional model and the latter model as the fixed-effects unconditional model. The fixed-effects model is a more parsimonious model that is nested within the random effects model. The fit of the models can be compared by taking the difference between their deviance statistics, with the deviance statistic being a function of the log-likelihood that results from maximum likelihood estimation of the model parameters. To ascertain the significance of the difference between the deviance statistics, the difference is compared to a χ^2 distribution with degrees of freedom equal to the difference in the number of parameters being estimated (in this case, $df = 1$). A statistically significant difference implies that significant variability exists among the population effect sizes and that a random-effects model should be used with the data.

Because the value of γ_0 represents the average gain made in American Experience knowledge and skills during students' first year and a half in college, it was used to answer research question 1a. Depending on which model yielded superior fit, either the γ_0 from the random-effects or fixed-effects model was interpreted.

Cohort Model. Given significant variation in the population effect sizes, answers to the second set of research questions were pursued by adding predictor variables to Level 2 of the random-effects model in Equation 8. To explore if there were significant differences among the effect sizes associated with students in different cohorts, we added as predictors four dummy-coded variables to represent the cohort variable, with the first cohort serving as the reference group. The cohort model was therefore specified as

$$\delta_j = \gamma_0 + \gamma_1(Cohort2) + \gamma_2(Cohort3) + \gamma_3(Cohort4) + \gamma_4(Cohort5) + u_j \,. \tag{9}$$

In this model γ_0 represents the average effect size for Cohort 1 and γ_1 through γ_4 represent the differences of each cohort from Cohort 1. To determine if this more complex model fits the data significantly better than the unconditional random-effects model and thus to answer research question 2a, the difference between the deviance statistics associated with each model were computed. This difference was compared to a χ^2 distribution with degrees of freedom equal to four. If the difference was statistically significant, it was concluded that there were significant differences among the effect sizes associated with the various cohorts.

If this model fit significantly better than the unconditional random-effects model, the significance of coefficients γ_1 through γ_4 (where H_0: $\gamma_g = 0$) were examined to determine which cohorts differed significantly from Cohort 1. Pair-wise comparisons among all other cohorts (e.g., Cohort 2 vs. Cohort 3) were pursued by testing the null hypotheses H_0: $\gamma_g = \gamma_h$, with $g \neq h$. To decrease our chances of making a Type I error, significance tests associated with pair-wise comparisons of cohort effects were evaluated using $\alpha = .01$.

Status Model. A similar approach was taken to answer research question 2b, which was used to examine the association between effect sizes and the requirement completion statuses of students. Dummy-coded variables representing the status variable were used as predictors at Level 2, with the group not having yet completed their requirement at the time of posttest (None) serving as the reference group. The status model was therefore specified as

$$\delta_j = \gamma_0 + \gamma_1(AP) + \gamma_2(TR) + \gamma_3(HIST) + \gamma_4(POSC) + \gamma_5(Both) + u_j . \quad (10)$$

The difference in the deviances of this model and the unconditional random-effects model was again used to determine if gains made in AMEX knowledge differed by status. If the status model in Equation 10 fit significantly better than the unconditional model, we determined how the effect sizes associated with the various status groups differed from one another using the same approach outlined above for the cohort effect.

Deciding between values of ρ and δ to use in Equation 6.

Values of ρ. Even though the correlation between the pretest and posttest scores is not used in the computation of d_j, it is used in the computation of its sampling variance in Equation 6. It would only be appropriate to use a single value of this correlation for all effect sizes if the population correlations (ρ_j) associated with each effect size were homogeneous. To determine if there was significant variation among the population correlations (ρ_j), a meta-analysis using the sample correlations (r_js) as estimates of ρ_js was performed prior to calculating the sampling variance. Because the sampling distribution of r tends to be skewed, sample correlations were first transformed to z-scores using Fisher's r-to-z transformation (Fisher, 1928)

$$z_j = \left(\frac{1}{2}\right)\ln\left(\frac{1+r_j}{1-r_j}\right), \quad (11)$$

which has a sampling variance equal to

$$\sigma_{z_j} = \frac{1}{n_j - 3}. \quad (12)$$

An unconditional model similar to that used above (Equations 5 & 8) was specified, using z_j as the dependent variable at Level 1 as opposed to d_j. Specifically, fixed and random-effects unconditional models were fit to the data to determine if the transformed population correlations were homogeneous. If the random-effects model fit the data significantly better than the fixed-effects model, it was concluded that the correlations were heterogeneous. If this conclusion was made, we examined whether cohort or status could be used to explain the variation in correlations by including these variables as predictors in two separate models similar to Equations 9 and 10 respectively, but using the transformed population correlation associated with each effect size j as the dependent variable. If either of the two models fit the data significantly better than the unconditional model, then the coefficients associated with each level of the variable were transformed back to the correlation metric[1] and used in Equation 6. In this situation, different correlations would be used for the sampling variances of effect sizes associated with different cohorts or statuses.

Alternatively, it was concluded that the correlations were homogeneous if the random-effects model did not fit significantly better than the fixed-effects model. If this conclusion was made, the overall correlation (ρ_z) from the fixed effects model was transformed back to the correlation metric and used for all effect sizes in Equation 6.

V*alue of δ.* We used the average of the sample effect sizes as our estimate of δ when calculating the sampling variance of d_j for each study. This same approach was taken in the example provided by Morris and DeShon (2002).

Software

The PROC MIXED application in the software program SAS (version 9.1) was used for all analyses. Because comparisons of the deviance statistics of models that differed in both their random and fixed parts was utilized in this study, full as opposed to restricted maximum likelihood was used for estimation (Hox, 2002). A primer for how to use PROC MIXED with HLM in general is available from Singer (1998) and when performing meta-analysis in particular by Sheu and Suzuki (2001).

Results

The descriptive statistics of the pretest and posttest scores and the correlation between such scores collected from the assessment reports are shown in Table 1. To understand the levels at which the cohort and status groups are scoring on the AMEX at pretest and posttest, the average pretest and posttest means (weighted by sample size) were calculated across the various cohort and status groups and are shown in Figures 1 and 2 respectively. Figure 1 shows little variability of the pretest averages across cohorts, with the lowest pretest average being 40.85 (Cohort 5) and highest being 42.44 (Cohort 1). These averages correspond to percent correct scores of 50% and 52% respectively. Across cohorts, students on average are obtaining a percent correct score of about ~51% on the AMEX upon entry to college. There is also little variability among the posttest averages for the various cohorts, with the lowest average being 43.39 (Cohort 2) and highest being 45.98 (Cohort 1). These averages correspond to percent correct scores of 54% and 57% respectively. Across cohorts, students on average are obtaining a percent correct score of about ~56% on the AMEX in the second semester of the sophomore year. Although Cohort 2 stands out in Figure 2 as being the cohort with the lowest gain from pretest to posttest, the typical increase in points on the raw score scale (2-4) does not seem to vary substantially across cohorts. Also, it should be kept in mind that this gain of 2 to 4 points is not impressive when considering that the raw score scale is comprised of 81 points.

The pretest averages in Figure 2 were fairly similar for the status groups of None, TR, POSC and HIST, and equaled a value of about ~40 (49% on the percent correct scale). The pretest average for the Both status group was somewhat higher (43.87) implying that students who take both POSC and HIST within their first year and a half are coming into college with slightly more knowledge about the American Experience (compared to None, TR, POSC, & HIST). Perhaps these students have an increased interest in the content area thus resulting in their somewhat elevated pretest average and their completion of both courses soon after entry to college. The highest pretest average is for the AP status group (55.78). The elevated average for this group is quite possibly due to their recent study of the American Experience for the AP exam.

The gains made over time (~ 2 points) and the resulting posttest averages in Figure 2 are fairly similar for the None and TR group. As well, the gains made over time (~ 4.5 points) and the resulting posttest averages in Figure 2 are fairly similar for the HIST and POSC group, with the values for the HIST group being somewhat larger. The posttest average remains high and barely increases for the AP group. The largest increase is associated with the group having already completed both courses. Their average increases 7.5 points resulting in a posttest average score of 51.39. The large confidence intervals around the averages for this group reflect the relatively small sample size for this group; only 149 of 3134 total students used in this study had both courses completed at the time of posttest.

Figures 1 and 2 are useful in speculating as to what the typical gains may be and whether differ by cohort or status, but only a meta-analysis of the effect sizes capturing these gains will provide the statistical significance tests needed to answer the first and second set of research questions. To calculate the effect sizes shown in Table 1, the means of the pretest and posttest scores as well as the standard deviation of the pretest scores were used in the estimator formula for d_j shown in Column B of Table 3.

Deciding between values of ρ and δ to use in calculating d*'s sampling variance*

V*alues of ρ.* To determine the value (or values) to use for ρ in Equation 6, a meta-analysis of the correlations between pretest and posttest scores was executed. The results of this meta-analysis indicated that the random-effects unconditional model did not fit the data significantly better than the fixed-effects model ($\chi^2(1) = 2.2$, $p = .14$). It was therefore concluded that the correlations did not significantly differ from one another and the estimated population correlation ($\rho = .71945$) from the fixed-effects model was used in Equation 6.

V*alue of δ.* The average d_j, weighted by sample size, across the 30 effect sizes in Table 1 is .33769. This value was used as δ to calculate the sampling variance for each d_j using Equation 6. The resulting sampling variances using $\rho = .71945$ and $\delta = .33769$ for each d_j are shown in Table 1.

Meta-Analysis

Unconditional Model. The random-effects unconditional model fit the data significantly better than the fixed-effects model ($\chi^2(1) = 92.7$, $p = .0000$), indicating significant variability among the effect sizes. The parameter estimates for the unconditional random-effects model are shown in Table 4. The random-effects model yielded a value of γ_0 equal to .326 indicating that on average, the posttest average is about 1/3 of a standard deviation above the pretest average in the population. Multiplying the variance of the effect sizes in the population (σ^2_{uj}) by 1.96 captures the range for 95% of the effect sizes. In the current study, the variance was estimated as .039, implying that that 95% of the population effect sizes are between .25 and .40.

Cohort Model. To understand if cohort could explain variation in the population effect sizes, a model including the dummy-coded variables for cohort was fit to the data. The deviance statistics for this model and the unconditional random-effects model, as well as their chi-square difference test, are shown in Table 5. The cohort model did not fit significantly better than the unconditional random-effects model ($\chi^2(4) = 3.2$, $p = .53$), indicating that the effect sizes did not significantly vary by cohort.

Status Model. To understand if a status effect could explain variation in the population effect sizes, a model including the dummy-coded variables for status was fit to the data. As shown in Table 5, the model including the dummy-coded variables for status fit significantly better than the unconditional random-effects model ($\chi^2(5) = 55.3$, $p = .0000$). It was therefore concluded that the population effect sizes significantly varied by status. The parameter estimates for the status model are shown in Table 6. The estimated population effect size for each status group was computed by adding the coefficient for the group to the intercept, with the intercept itself serving as the population effect size for the None status group. Figure 3 shows the estimated population effect sizes for each status group. Ranging from lowest to highest, the number of standard deviation units difference between the posttest and pretest means was estimated as .04 for AP, .18 for TR, .28 for None, .42 for POSC, .54 for HIST and .90 for Both. Pair-wise comparisons among the status groups indicated that the effect size for the AP group did not significantly differ from that of the TR group, that the TR effect size did not significantly differ from that of the None group and that the HIST effect size did not significantly differ from the Both group. All remaining pair-wise comparisons among the status groups were significant. It may seem surprising that the HIST effect size was not significantly different than the Both effect size given the large difference in the point estimates of these two groups. However, because the confidence interval for the Both effect size is large due to the small sample size for this group, our pair-wise significance test did not find it significantly different than the HIST effect size.

After accounting for status, almost no variation remained among the effect sizes. In fact, by including the status variable in the model the variance in the population effect sizes decreased substantially, from a value of .039 in the unconditional model to less than .0001 in the status model. The small variance remaining after including status suggested that a simpler model including status as a predictor, but constraining the population effect size variance to zero could be used with the data. In fact, when this model was fit to the data the deviance statistic was still -47.9, implying that once accounting for status, no significant variance among the 30 population effect sizes remained. The coefficients and their standard errors for this model did not change from the previous model and are therefore not reported.

Discussion

Assessment data from the previous five cohorts of students at our university were used to explore the typical gains made during the first year and a half of college in students' American History knowledge and skills and variables thought to be related to such gains. In summarizing our results, the research questions are restated and the findings for each question described.

1a. What is the average effect size? That is, what is the average gain made in American Experience knowledge and skills during students' first year and a half in college?

The average effect size estimated from the unconditional random-effects models was .326 indicating that on average, the posttest average is about 1/3 of a standard deviation above the pretest average in the population. However, because effect sizes were found to significantly vary by requirement completion status, the effect size for each status group should be interpreted as opposed to the overall average.

1b. Are there significant differences among the effect sizes? That is, do the effect sizes differ greatly from one another or are they all fairly similar?

We did find that the effect sizes significantly differed from one another. The variation of the population effect sizes was calculated as .039, implying that that 95% of the effect sizes from this population are between .25 and .40. However, no variation among the effect sizes remained once controlling for requirement completion status.

2a. Are there significant differences among the effect sizes associated with the various cohorts?

Effect sizes did not significantly differ across cohorts implying that when it comes to the gains made in American Experience knowledge, the previous five cohorts of students at our university do not significantly differ from one another. Given that this general education program, the instrument used to assess this program, and the incoming student demographics (e.g., proportion female, SAT averages) have remained unchanged across years, the finding that the cohorts do not differ in their gains is not too terribly surprising.

2b. Are there significant differences among the effect sizes associated with the various AMEX completion statuses of students?

Students with various requirement completion statuses did significantly differ from one another in gains. In fact, status had such a strong relationship with the effect sizes that no variability among the population effect sizes remained once accounting for status.

Negligible gains were found for the AP ($\delta = .04$) and transfer students ($\delta = .18$), who did not significantly differ from one another in their effect sizes. It makes sense that we would see negligible gains for the AP group since it is likely that these students completed their AP test prior to the pretest and therefore were not exposed to any form of the “treatment” (e.g., university coursework in American Experience) during the time elapsing between pretest and posttest. The same reasoning applies to students transferring in credit to fulfill this requirement, although a subset of these student make have actually taken their transfer coursework elsewhere in the time between pretest and posttest. Regardless of when the transfer course was taken, we would not anticipate the effects sizes of the transfer group to be larger than those associated with students taking courses at our own university since the learning objectives for American Experience courses at other universities are likely to differ from our own.

The effect size for the group of students yet to fulfill the requirement was somewhat higher ($\delta = .28$), although not significantly different from that of transfer students. Because students in this group had not been exposed to any form of the “treatment”, negligible effect sizes were anticipated for this group. However, because assignment of students to various status groups is based on whether or not they had fulfilled the requirement prior to the semester in which the posttest was administered, it is possible that a subset of students in this group were actually enrolled in GHIST 225 or GPOSC 225 during the semester in which the posttest was administered. Perhaps this explains why the effect size of the None group is somewhat higher than the effect size of transfer students. It is suggested that future assessment reports provide the pretest and posttest statistics separately for students currently enrolled in American Experiences courses in order to explore this possibility.

Of the AP, TR and None status groups, the AP group is most likely to *not* have been exposed to any form of the “treatment” between pretest and posttest. Thus, the amount of gain seen in effect sizes due to maturation alone may be inferred from the AP group effect size. However, because the kind of student that takes the AP exam (and who scores high on the exam) is most likely to differ from other students at our university, the AP effect size cannot be used to infer what the gains due to maturation are for other students.

Students who completed either GHIST 225, GPOSC 225 or both courses prior to the posttest administration date had significantly larger effect sizes ($\delta = .42$, $\delta = .54$, and $\delta = .90$ respectively) than students who had either not completed the requirement or who had used AP or transfer credit to fulfill the requirement. We are pleased with this result since the students making the largest gains are the same students who were exposed to the most ideal form of "treatment", which are the courses that focus on the American Experience learning objectives. Students taking either course do not significantly differ from one another over time or in their resulting posttest scores on the AMEX.

While the effect size for the group of students completing both courses was quite a bit larger than for students completing only one of the courses, it did not significantly differ from the effect size associated with students completing only GHIST 225. As aforementioned, the lack of a significance between these two groups is probably attributable to the uncertainty regarding the population value of the Both group's effect size. After seeing the large effect size for the Both group, one may be quick to conclude that both courses should be taken to fulfill the requirement as opposed to only one. However, before this conclusion consider: 1) the uncertainty regarding the exact value of this gain and 2) the relatively high pretest average score for this group (see Figure 2). Because students in the Both status group may have a higher interest in the content area, they come into college knowing more about the American Experience and are also more sensitive to the "training", thus make larger gains over time. The interest that these students have in the content area may make them unique and therefore we caution against concluding that the same gains would be made for other students completing both courses.

There are several factors that need to be taken into consideration when interpreting the effect sizes found in our study. First, the fact the posttests were administered a year and half after the pretest should be taken into consideration. This is particularly important for the students taking their American Experience courses at our university since some may have completed their coursework their first semester and others the semester previous to the posttest. It would be interesting to report the pretest and posttest statistics for students in the HIST and POSC group by the semester in which the course was completed. This information could be used to assess the retention of knowledge and skills with time.

When interpreting the effect sizes it is also important to keep in mind the average pretest and posttest raw scores for the various status groups. While students completing either GHIST 225 or GPOSC 225 had pretest and posttest averages of about 40 and 44 respectively, the students completing both courses had corresponding averages of 44 and 51. It is disconcerting that the students taking only one course had posttest averages that were about equal to the Both group's *pretest* average, although we make this statement with caution given the small sample size upon which the Both group's averages were based. The pretest and posttest averages of the students taking only GHIST 225 or GPOSC 225 can also be compared to those of the AP group, who on average scored about a 56 on both the pretest and posttest. Students completing the minimal amount of coursework for the American Experience requirement are scoring, on average, about 12 points at posttest below the *pretest* score of students using AP credit to fulfill the requirement. It is for the American Experience faculty to decide how comfortable they are with this discrepancy, keeping in mind that students who score high on the AP exam tend to in general be stronger students.

Considering whether a posttest score of 56 is a reasonable expectation for students completing the American Experience requirement emphasizes one of the flaws in relying solely on the amount of gains made over time when assessing the effectiveness of a general education program. Students may be making gains over time, but are their resulting scores high enough? Ideally, the faculty of general education courses would use standard setting procedures to establish the minimum score on a general education assessment measure that would be expected from a typical student who has satisfactorily completed the general education curriculum. If a proficiency standard is set on the test, then more meaning can be gleaned from the test scores. For instance, with a repeated measures assessment design using a proficiency standard, not only can change over time be examined, but also whether or not the resulting test scores are acceptable.

With only about half of the items being answered correctly on the test by students who have already completed the minimal requirements for the program, the posttest scores in the current study may seem low. However, faculty may actually consider this level acceptable after a standard setting procedure, in which the difficulty of the content area and test but also the capabilities of the typical student are taken into consideration when setting the proficiency standard. The posttest scores may also be somewhat deflated due to the fact that there are no personal stakes associated with students performance on the test. However, despite the fact that there are no consequences associated with student's performance on the test, evidence exists to support the notion that students at this university put forth adequate effort and view the results of Assessment Day as important (Sundre & Wise, 2003; Wise & Kong, 2005).

Rules of thumb used to judge the magnitude of Cohen's *d* specify .2, .5, .8 as small, medium and large effects accordingly. Using this framework, the effect size found in our study is slightly larger than a small effect. However, even though this effect may be considered small, it is unusual in the social sciences to find medium or large effects. Rather than comparing this effect size to rules of thumb, in future research we hope to provide a more meaningful comparison by contrasting the effect sizes with other effect sizes calculated in the same way and used in educational repeated measures studies.

This study illustrates how the effect sizes from pretest and posttest scores can be used in general education program assessment. While the results are incredibly informative for faculty and administrators at our university, they are also informative for persons at other universities using a similar data collection scheme. Most introductory statistics textbooks suggest that instead of judging the magnitude of an effect size against rules of thumb, the effect size should be interpreted in the context of the research question being posed. What are typically considered "small" effect sizes by rules of thumb may actually be large when considering the research question. This study is a first step towards understanding the magnitude of effect sizes to anticipate from general education programs. We hope that others will refer to this study when judging the magnitude of general education gains at their own universities.

References

Banta, T. W., Lund, J. P., Black, K. E., & Oblander, F. W. (1996). *Assessment in practice: Putting principles to work on college campuses.* San Francisco, CA: Jossey-Bass.

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology, 41,* 257-278.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1,* 170-177.

Erwin, T. D. (1990). *Assessing student learning and development.* San Francisco, CA: Jossey-Bass.

Hox, J. (2002). Multilevel analysis techniques and applications. Mahwah: Lawrence Erlbaum.

Fisher, R. A. (1928). *Statistical methods for research workers* (2nd ed.). London: Oliver & Boyd.

Gaff, J. G. (1991). *New life for the college curriculum.* San Francisco, CA: Jossey-Bass.

Gibbons, R. D., Hedeker, D. R., & Davis, J. M. (1993). Estimation of effect sizes from a series of experiments involving paired comparisons. *Journal of Educational Statistics, 18,* 271-279.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7,* 105-125.

Ratcliff, J. L., Johnson, D. K., La Nasa, S. M., & Gaff, J. G. (2001).The status of general education in the year 2000: Summary of a national survey. Association of American Colleges & Universities.

Raudenbush, S., & Bryk, D. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics, 10,* 75-98.

Raudenbush, S., & Bryk, D. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks: Sage.

Sheu, C-F., & Suzuki, S. (2001). Meta-analysis using linear mixed models. *Behavior Research Methods, Instruments, & Computers, 33,* 102-107.

Sundre, D. L., & Wise, S. L. (2003, April). *'Motivation filtering': An exploration of the impact of low examinee motivation on the psychometric quality of tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*, 163-183.

Footnotes

1. Coefficients were transformed back to the original correlation metric using

$$\rho = \frac{\exp(2\rho_z) - 1}{\exp(2\rho_z) + 1}.$$

Appendix A
Learning Objectives of the American Experience General Education Domain

Students completing an American Experience course of Cluster Four will be able to identify, conceptualize and evaluate:

- Social and political processes and structures using quantitative and qualitative data
- Key primary sources relating to American history, political institutions and society
- The nature and development of the intellectual concepts that structure American political activity
- The history and operation of American democratic institutions
- The history and development of American society
- The history and development of American involvement in world affairs

Table 1
Pretest and Posttest Descriptive Statistics, Effect Sizes and Effect Size Sampling Variance by Cohort and Status

				Pretest		Postest				
Cohort	Status	*n*	%	*M*	*SD*	*M*	*SD*	*r*	*d*	σ^2_e
1	None	113	23%	39.60	9.51	42.31	9.72	0.72	0.28	0.006
	AP	47	10%	57.43	8.63	57.04	8.82	0.88	-0.04	0.014
	TR	23	5%	44.48	10.30	47.78	11.29	0.83	0.32	0.030
	HIST	180	37%	40.96	8.81	45.57	9.50	0.66	0.52	0.003
	POSC	104	22%	40.12	9.05	43.88	10.28	0.56	0.42	0.006
	Both	16	3%	47.19	8.03	55.25	7.15	0.81	1.00	0.045
2	None	136	37%	39.56	9.58	40.74	10.19	0.70	0.12	0.005
	AP	35	9%	53.89	7.89	53.71	11.46	0.85	-0.02	0.019
	TR	16	4%	40.69	8.58	37.81	8.09	0.80	-0.33	0.045
	HIST	145	39%	40.22	8.84	44.20	10.65	0.70	0.45	0.004
	POSC	33	9%	41.42	8.76	42.64	9.88	0.47	0.14	0.020
	Both	5	1%	34.40	5.64	42.60	6.99	0.91	1.45	0.274
3	None	216	27%	38.91	10.43	42.16	10.75	0.76	0.31	0.003
	AP	79	10%	56.34	7.91	56.70	8.16	0.83	0.04	0.008
	TR	41	5%	40.24	8.31	41.66	8.69	0.75	0.17	0.016
	HIST	351	43%	39.87	9.16	45.31	9.34	0.72	0.59	0.002
	POSC	113	14%	40.26	9.68	45.08	9.79	0.66	0.50	0.006
	Both	7	1%	41.57	9.02	48.71	7.91	0.07	0.79	0.140
4	None	194	24%	39.51	10.47	42.56	10.41	0.69	0.29	0.003
	AP	93	12%	54.79	8.35	55.76	9.70	0.61	0.12	0.007
	TR	39	5%	38.41	9.57	39.80	9.93	0.78	0.14	0.017
	HIST	339	43%	39.62	8.86	44.69	8.99	0.74	0.57	0.002
	POSC	118	15%	40.36	9.38	44.28	9.56	0.77	0.42	0.005
	Both	12	2%	43.42	9.71	51.67	8.03	0.58	0.85	0.065
5	None	226	33%	39.22	9.41	42.40	9.36	0.71	0.34	0.003
	AP	71	10%	56.30	8.51	56.34	8.68	0.76	0.00	0.009
	TR	30	4%	37.87	8.29	41.17	7.09	0.69	0.40	0.022
	HIST	258	38%	38.96	9.15	43.29	9.76	0.70	0.47	0.002
	POSC	88	13%	38.75	9.82	42.69	11.27	0.75	0.40	0.007
	Both	6	1%	46.50	10.91	51.00	11.75	0.83	0.41	0.185

Table 2
Sample size, pretest and posttest data collections dates and coefficient alpha by cohort

Cohort		Data Collection Dates		Coefficient Alpha	
Number	N	Pretest	Posttest	Pretest	Posttest
1	483	Fall 2000	Spring 2002	0.86	0.87
2	370	Fall 2001	Spring 2003	0.84	0.87
3	807	Fall 2002	Spring 2004	0.86	0.86
4	795	Fall 2003	Spring 2005	0.90	0.86
5	679	Fall 2004	Spring 2006	0.86	0.86

Table 3
Population parameters, estimators using descriptive statistics, estimators using dependent samples t *for repeated measures effect sizes on the raw score and change score metric*

	A. Raw Score Metric ($\sigma_{pre} = \sigma_{post} = \sigma_{raw}$)	B. Raw Score Metric ($\sigma_{pre} \neq \sigma_{post} \neq \sigma_{raw}$)	C. Change Score Metric
Population parameter	$\delta_{raw} = \frac{\mu_{post} - \mu_{pre}}{\sigma_{raw}}$	$\delta_{raw} = \frac{\mu_{post} - \mu_{pre}}{\sigma_{pre}}$	$\delta_{gain} = \frac{\mu_{post} - \mu_{pre}}{\sigma_{gain}}$
Estimator using descriptive statistics	$d_{raw} = \frac{M_{post} - M_{pre}}{SD_{raw}}$	$d_{raw} = \frac{M_{post} - M_{pre}}{SD_{pre}}$	$d_{gain} = \frac{M_{post} - M_{pre}}{SD_{gain}}$
Estimator using dependent samples t	$d_{raw} = t\sqrt{\frac{2(1-r)}{n}}$	Not applicable.	$d_{gain} = \frac{t}{\sqrt{n}}$

Table 4
Unconditional random-effects model parameter estimates

Unconditional Model	Value	SE	*t*	*p*
Fixed Effects				
γ_0	0.326	0.04	7.67	<.0001
Random Effects				
σ_{uj}	0.039	0.02		

Table 5
Deviance statistics and chi-square difference tests

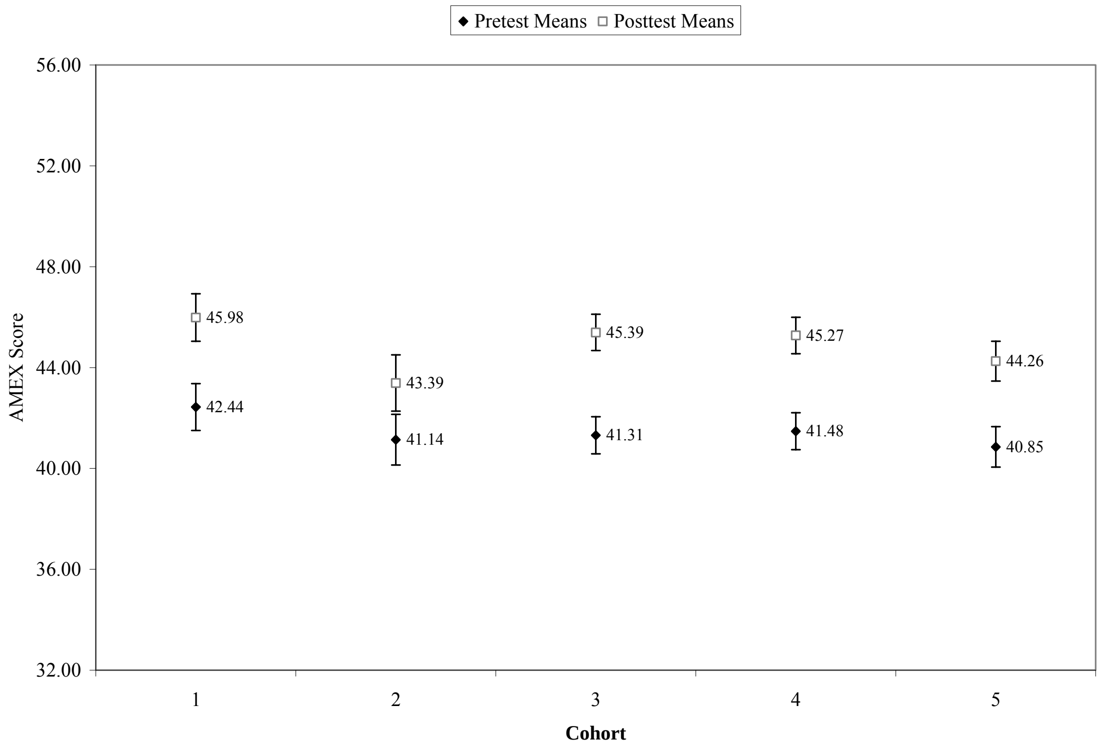
Model	Deviance	χ^2	*p*
Unconditional Model	7.4		
Cohort Model	4.2	3.2	0.5249
Status Model	-47.9	55.3	0.0000

Note. All models are random-effects models. χ^2 values computed comparing the model to the unconditional model.

Table 6
Status model parameter estimates

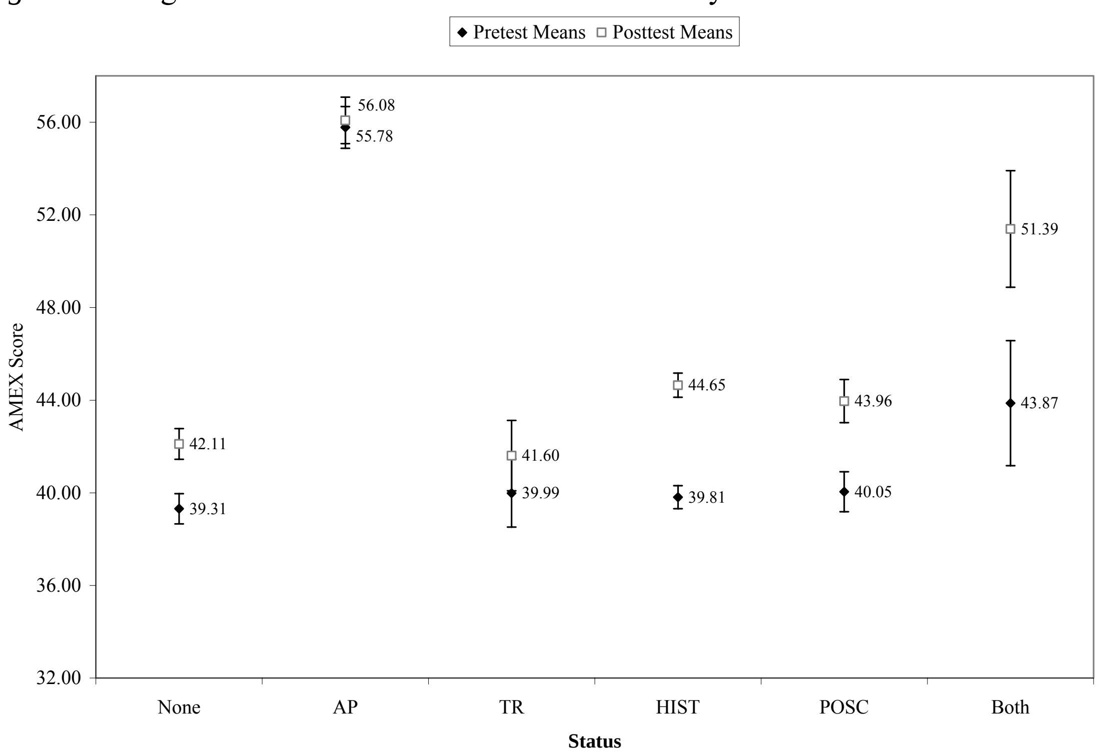
Unconditional Model	Value	SE	*t*	*p*
Fixed Effects				
γ_0	0.281	0.03	10.41	<.00001
γ_1	-0.244	0.05	-4.67	0.0001
γ_2	-0.099	0.07	-1.37	0.1832
γ_3	0.256	0.04	7.25	<.00001
γ_4	0.133	0.05	2.88	0.0082
γ_5	0.621	0.14	4.46	0.0002
Random Effects				
σ_{uj}	0.0001	0.001		

Figure 1. Weighted Pretest and Posttest AMEX Means by Cohort.



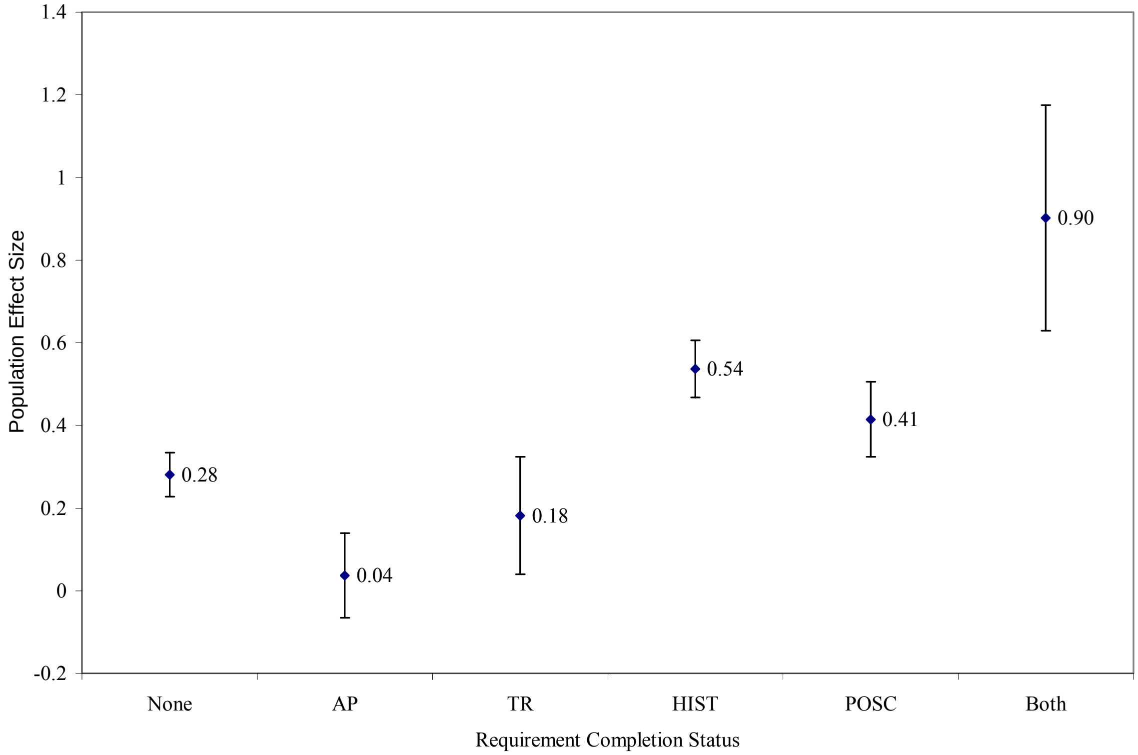
Note. Means are surrounded by 95% confidence intervals

Figure 2. Weighted Pretest and Posttest AMEX Means by Status.



Note. Means are surrounded by 95% confidence intervals.

Figure 3. Estimated Population Effect Sizes from Status Completion Model.



Note. Means are surrounded by 95% confidence intervals.