Examining College Students' Gains in General Education

Dena A. Pastor and Pamela K. Kaliski
James Madison University

Brandi A. Weiss
University of Maryland

Abstract

Do students change as a result of completing their general education requirement? This question was examined by using a pretest/posttest design with five different cohorts of students required to complete a general education program in American history and politics. Differences among various groups in Cohen's *d* (the standardized difference between pretest and posttest means) were examined using hierarchical linear modeling. Results indicated that differences could be explained by how the requirement was fulfilled; negligible gains were found for students using advanced placement or transfer credit (.04 -.18), whereas moderate/large gains were found for students who had completed the university course(s) (.42 - .90). The gain found for students yet to fulfill the requirement (.28) was explained by the large presence in that group of students currently enrolled in the course. Different definitions of *d* used with the pretest/posttest design are described and implications of the results for assessment are discussed.

Examining 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 knowledge, skills, values and personal characteristics of the 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, lifelong learners and mindful servants to society (Fong, 2004). 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 student's undergraduate education at institutions with this requirement is composed of general education courses, it is important to understand what impact these courses have on 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 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 100-point scale. For this reason, it is desirable to report standardized measures of change. Standardized measures of an effect are often conveyed using effect sizes, which are typically used to capture practical significance. Readers may be familiar with the effect size known as Cohen's *d*, which provides a standardized measure of the difference between group means. By standardized we mean that the difference is reported in standard deviation units, not in the unit of the raw score scale. 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. One advantage in using effect sizes is that only descriptive statistics are required for their computation. To illustrate this advantage, consider a mature assessment program, in which the same assessment data have 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 were properly archived and are accessible. It may be relatively easier to obtain the descriptive statistics of the results, perhaps from assessment reports. Only the descriptive statistics in the reports are needed to compute the effect sizes.

Once effect sizes are collected, a statistical technique known as meta-analysis is often used to average effect sizes, determine the extent to which effect sizes vary, and examine the extent to which certain variables are related to effect size estimates. Returning to our example above, meta-analysis could be used to: (a) pool effect sizes across years to obtain an accurate estimate of the program's effectiveness, (b) examine the extent to which effect sizes vary across years and (c) explain why effect sizes vary. For instance, if our effect sizes were collected before and after substantial program improvements, meta-analysis could also be used to determine if larger effect sizes are associated with the program improvements.

Effect sizes could also be quite useful when wanting to compare program effectiveness across institutions. Whereas it may be nearly impossible to obtain the raw data from institutions, the acquisition of descriptive statistics is feasible, perhaps from required reports sent to state councils of higher education or accrediting bodies. Once the descriptive statistics are acquired, effect sizes from various institutions can be compared. The use of different instruments or assessment designs (e.g., different amounts of time elapsing between pretest and posttest) does not preclude the comparison of different institutions, but does necessitate the investigation of whether the differences in institution effect sizes are attributable to differences in the instruments or assessment designs that were employed. Effect sizes based on the measurement of different constructs (e.g., quantitative reasoning vs. writing ability) can also be compared, so long as "construct type" is formally examined as a source of variability among the effect sizes. A meta-analysis of effect sizes corresponding to different constructs may be a useful way to study if change over time differs across the various general education domains.

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.

Every year since the fall of 2000, the American Experience Test (AMEX) has been administered to a random sample of students. These students have completed the assessment on two occasions: once as incoming freshmen (pretest) and again as second semester sophomores (posttest). After the posttest administration, an assessment report has been created containing the descriptive statistics for groups of students having different requirement completion statuses. Specifically, descriptive statistics have been reported for six groups: (a) students who have not completed the requirement and five groups of students who have completed the requirement by (b) using advanced placement credit, (c) using transfer credit, (d) completing a political science course, (e) completing a history course, or (f) completing both the political science and history courses.

The descriptive statistics for each of these six groups were obtained from each cohort report resulting in a total of 30 effect sizes that were used to answer two sets of research questions. The first set of research questions deals with estimating the typical change over time and determining if there is significant variation in change over time. Specifically, the following questions were posed:

1a. What is the average effect size?

1b. Is there significant variation among the effect sizes?

If significant variation among the effect sizes was found, a second set of research questions was pursued to explore why the effect sizes vary. Specifically, the following questions were posed:

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

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, which consists of students who had completed the requirement by the time of posttest. 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 advanced placement or transfer credit to fulfill this requirement.

Methods

We first describe how and from whom assessment data are 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 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, its learning objectives, and a detailed description of its courses can be found at http://www.jmu.edu/gened/cluster4.html. One of two courses can be used to fulfill the requirement for the American Experience program: General Education History 225: U.S. History (GHIST 225) or General Education Political Science 225: U.S. Government (GPOSC 225). Students also have the option of fulfilling this requirement by scoring a four or above 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 other universities. The following labels will be used for the remainder of the paper to describe the requirement completion status of students: None–requirement not yet fulfilled, AP – requirement fulfilled through advanced placement credit, TR–requirement fulfilled through transfer credit, HIST-requirement fulfilled by completion of GHIST 225, POSC-requirement fulfilled by completion of GPOSC 225, and Both-requirement fulfilled by completion of both GHIST 225 and GPOSC 225.

About 70% of students fulfill their American Experience general education requirement prior to the second semester of their sophomore year. Table 1 shows the percentage of students in each status completion group by cohort. Requirement completion status percentages were fairly similar across cohorts. Across cohorts, the majority of students at the time of posttest completed the requirement at our university by taking GHIST 225 (40%), GPOSC 225 (14%) or both courses (2%). A sizeable percentage of students across cohorts had not completed the requirement (29%) at the time of posttest. About 10% and 5% of students across cohorts fulfilled the requirement through advanced placement 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 institution-wide assessment days are set aside each year for assessment purposes and students' course

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

Cohort	Status	n	$\%$	Pretest M	Pretest SD	Postest M	Postest 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

Note. AP=Advanced Placement, TR=transfer, HIST=completion only of GHIST 225, POSC=completion only of GPOSC 225, Both=completion of both GHIST 225 & GPOSC 225. r represents the correlation between pretest and posttest, d the standardized difference between pretest and posttest means, and σ^2_e the sampling variance of d.

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 the Fall Assessment Day in August (pretest). In the Spring Assessment Day (posttest), which takes place in mid-February, students with 45-70 credit hours upon completion of their first semester of their second year are tested. Because students are assigned assessment instruments using the last two digits of their student identification number, the instruments taken by a particular student at pretest are the same as those taken by that student at posttest. Assignment of instruments to students using their student identification number also ensures that the sample administered any given instrument is a random sample from the particular student cohort being tested. The pretest and posttest data collection dates for the five cohorts used in the present study are shown in Table 2.

Table 2
Sample Size, Pretest, and Posttest Data Collection Dates and Coefficient Alpha by Cohort

Cohort Number	N	Data Collection Dates		Coefficient Alpha	
		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

Measure

The American Experience test (AMEX), an 81-item multiple-choice test, was created in 1998 by faculty to assess the objectives of the American Experience domain. Coefficient alpha for the AMEX scores has been consistently high (0.84 - 0.90) across cohorts and data collection dates (see Table 2).

Assessment Reports

After each Assessment Day a report of the results is provided to general education faculty. In order to provide scores by requirement completion status, student assessment data are linked to university records. Assessment reports for the American Experience domain 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 cohort and status completion 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.

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 independent groups (IG) designs, there is little agreement as to how to define δ when using averages from 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 (either pooled or unpooled). 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

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

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

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

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. Alternatively, when δ_{raw} is employed the interpretation is more familiar to those using the standardized mean difference effect size with an IG design; a value of 0.5 for δ_{raw} implies

that pretest and posttest means differ by half of a standard deviation unit.

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. However, Morris and DeShon 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, the correlation between pretest and posttest scores, varies across studies. If ρ varies across studies, Morris and DeShon suggest either using: (a) δ_{raw} or (b) δ_{gain} with subsets of effect sizes having similar ρ.

In the current study 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 effect sizes 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. In the current study, we used the estimator because the

$$d_{raw} = \frac{M_{post} - M_{pre}}{SD_{pre}} \qquad (3)$$

resulting effect size is on the raw score metric and its calculation only requires the descriptive statistics available in the reports. As suggested by Becker (1988), we used 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 effect sizes. For the remainder of the paper we simplify our notation by referring to this estimator as d_j for effect size j (j=1, …, 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

$$d_j = \delta_j + e_j \qquad (4)$$

where e_j represents sampling error, which is the discrepancy between the population parameter and the sample estimate. The variance of e_j ($\sigma^2 e_j$) 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. A general form of the sampling variance for the estimator used in the present study is provided by Morris and DeShon (2002)

$$\sigma_{e_j}^2 = \left[\frac{2(1-\rho)}{n_j} \right] \left(\frac{n_j - 1}{n_j - 3} \right) \left[1 + \frac{n_j}{2(1-\rho)} \delta^2 \right] - \frac{\delta^2}{c_j^2} \qquad (5)$$

where n_j refers to the sample size and c_j refers to the bias function equal in this design to

$$c_j = 1 - \frac{3}{(4n_j - 4) - 1} \qquad (6)$$

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 (Hedges, 1981).

Meta-Analysis

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 4. 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 follow-

ing 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 Models. 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

$$\delta_j = \gamma_0 + u_j \qquad (7)$$

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. Because the value of γ_0 represents the average gain made in American Experience knowledge and skills, it was used to answer research question 1a. The variance of u_j ($\sigma^2 u_j$) represents the extent to which population effect sizes differ from the grand mean (γ_0). The significance of $\sigma^2 u_j$ was used to answer research question 1b, which asks whether there is significant variation among the population effect sizes.

To assess the significance of $\sigma^2 u_j$, we compared the fit of the model in Equation 7 against a model imposing the restriction that $\sigma^2 u_j$ equal zero. The fit of these nested models was compared by taking the difference between their deviance statistics and comparing it 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. Depending on which model yielded superior fit, either the γ_0 from the modeling estimating $\sigma^2 u_j$ or one constraining $\sigma^2 u_j$ to zero was interpreted when answering research question 1a.

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 model in Equation 7. 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

$$\delta_j = \gamma_0 + \gamma_1(Cohort2) + \gamma_2(Cohort3) + \gamma_3(Cohort4) + \gamma_4(Cohort5) + u_j \qquad (8)$$

In this model γ_0 represents the average effect size for Cohort 1 and γ_1 through γ_4 represent, respectively, the differences in average effect size of each cohort from Cohort 1. To determine if this more complex model fit the data significantly better than the unconditional model, and thus to answer research question 2a, the difference between the deviance statistics associated with each model were computed and 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 model, the significance of each of the coefficients γ_1 through γ_4 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 H0: $\gamma_g = \gamma_h$, with g≠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 requirement completion status. 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

$$\delta_j = \gamma_0 + \gamma_1(AP) + \gamma_2(TR) + \gamma_3(HIST) + \gamma_4(POSC) + \gamma_5(Both) + u_j \qquad (9)$$

The difference in the deviances of this model and the unconditional model was again used to determine if gains made in AMEX knowledge differed by status. If the status model in Equation 9 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 model.

Deciding between values of ρ and δ to use in Equation 5.

Researchers can choose to use estimated values of ρ and δ particular to each effect size j in the

formula for the sampling variance, or use common estimated values of these parameters, calculated by pooling estimates of ρ_j and δ_j across effect sizes. Our process for deciding between these two alternatives is described below.

Values of ρ. It would only be appropriate to use a single value of the correlation between the pretest and posttest scores in the sampling variances for all effect sizes if the population correlations (ρ_j) associated with the effect sizes were homogeneous. To determine if there was significant variation among the population correlations (ρ_j), a preliminary meta-analysis using the Fisher's (1928) r to z transformed sample correlations (r_js) as estimates of ρ_js was performed. The results of this meta-analysis indicated that population correlations did not significantly differ from one another ($\chi^2(1) = 2.2$, p=.14). The estimated population correlation (ρ=.71945) was therefore used when computing the sampling variances of all effect sizes.

Value of δ. We used the average of the sample effect sizes as our estimate of δ when calculating the sampling variance of dj for each study. This same approach was taken in the example provided by Morris and DeShon (2002). The average d_j, weighted by sample size, across the 30 effect sizes in Table 1 was calculated as .33769. This value was used as δ to compute the sampling variance (δ^2_{ej}) for each d_j using Equation 5. The resulting sampling variances using ρ=.71945 and δ=.33769 for each d_j are shown in Table 1.

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 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

Descriptive statistics of the pretest and posttest scores by cohort and requirement completion status are first described followed by the results of the meta-analytic models that were used to answer the first and second set of research questions.

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 1 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 courses within their first year and a half are coming into college with slightly more knowledge about the American Experience (compared to None, TR, POSC, and HIST). The highest pretest average is for the AP status group (55.78).

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.

Meta-Analysis

    Unconditional Model. Because significant variability existed among the effect sizes ($\chi^2(1)=92.7$, $p=.0000$), the parameter estimates for the model estimating δ^2_{uj} are shown in Table 3. The unconditional 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 (δ^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. The deviance statistics for the cohort model and the unconditional model, as well as their chi-square difference test, are shown in Table 4. The cohort model did not fit significantly better than the unconditional model ($\chi^2(4) = 3.2$, p = .53), indicating that the effect sizes do not significantly vary by cohort.

    Status Model. As shown in Table 4, the model including the dummy-coded variables for status fit significantly better than the unconditional 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 5. The estimated population effect size for each status group was computed by add-ing the coefficient for the group to the intercept, with the intercept itself serving as the population effect size estimate 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, .41 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
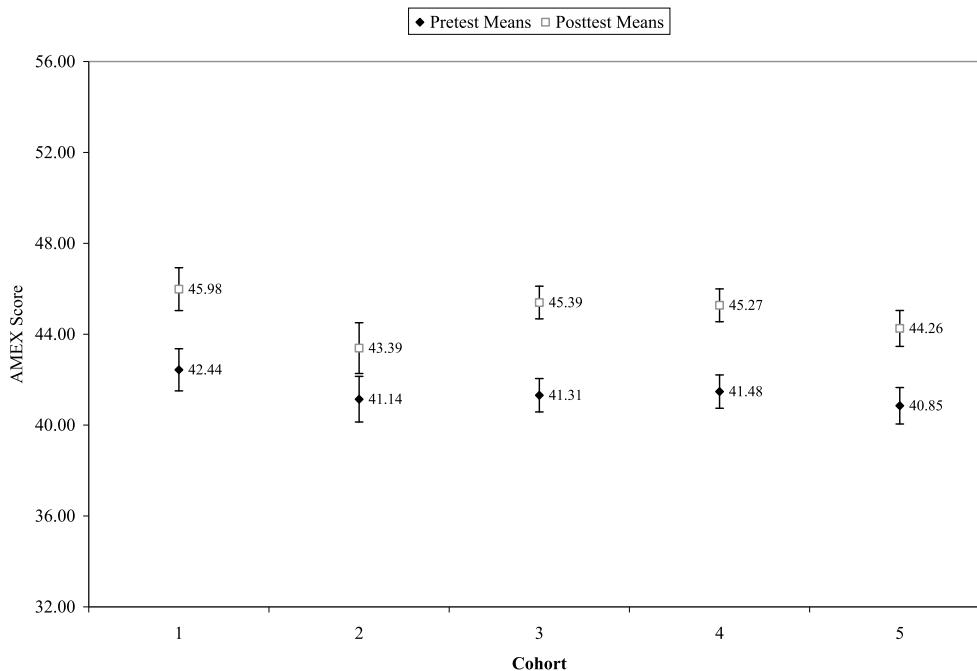
Figure 1. Weighted pretest and posttest AMEX means by cohort.
    Note. Means are surrounded by 95% confidence intervals

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 precision of the Both effect size is low 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.
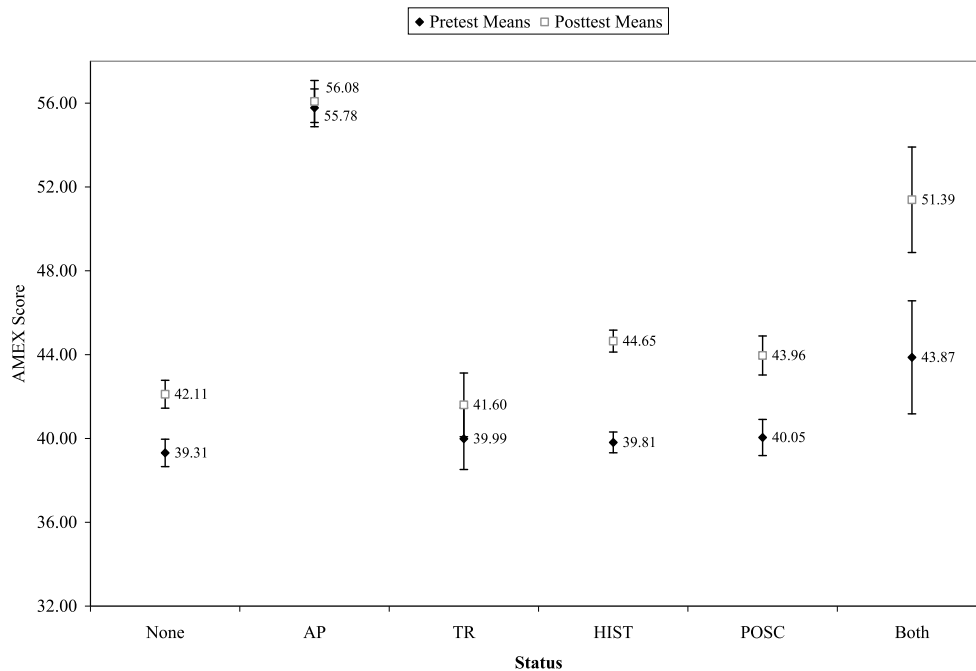
Figure 2. Weighted pretest and posttest AMEX means by status.

Note. Means are surrounded by 95% confidence intervals.

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.

1a. What is the average effect size?

The average effect size estimated from the unconditional model 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?

We did find that the effect sizes significantly differed from one another. The variation of the popu 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 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

Table 3
Unconditional 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 4
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. χ^2 values computed by comparing models to the unconditional model.

among the population effect sizes remained once accounting for status. Negligible gains were found for the AP (δ=.04) and transfer students (δ=.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. Similar reasoning applies to the TR group since 96% of the TR group completed their transfer course prior to pretest. We would not expect to see much of a gain over time for the TR group since a large majority of these students were not exposed to any form of "treatment" between pretest and posttest.

The effect size for the group of students yet to fulfill the requirement was somewhat higher (δ=.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. Post hoc analyses revealed that 45% of students in the None group were enrolled in GHIST 225 or GPOSC 225 during the semester in which the posttest was administered. The sample effect size estimated for these students was quite larger (d=.46) than those students in the None group who were not enrolled (d=.15). The presence of students in the None group enrolled in the course at the time of posttest explains why the effect size of this group is somewhat higher than the effect size of transfer students. To capture a true control group, it is suggested that future assessment reports exclude students from the None group who are enrolled in American Experiences courses at the time of posttest.

Table 5
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		

Students who completed either GHIST 225, GPOSC 225 or both courses prior to the post-test administration date had significantly larger effect sizes (δ=.42, δ=.54, and δ=.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 adopting this conclusion consider: (a) the uncertainty regarding the exact value of this gain and (b) the relatively high pretest average score for this group (see Figure 2). Because students in the Both group may have a higher interest in the content area, they may come into college knowing more about the American Experience and as a result, may also be more sensitive to the "treatment," thus make larger gains over time. Evidence that these students do have a heightened interest in the content area is the fact that 80% of the students are majors in either political science, history, interdisciplinary liberal studies or a related discipline. Because of the unique characteristics of the students in this group, we are hesitant to conclude that the same gains would be made by other students if they were to complete both courses.

Keeping in mind that the large gain made by the students taking both courses does not differ significantly from the HIST group and that the students in the Both group may not be reflective of the typical student, further investigation is still warranted as to why this relatively larger gain exists. For instance, further research is needed to pinpoint the cause of the relatively larger gains. Is the larger improvement for this group attributable to the different kinds of knowledge acquired in the two courses or to the reinforcement of the same knowledge in both courses? Further studies should also examine whether the same gains could be expected from typical students taking both courses. Results from such research together with the findings in the current study may prompt the general education program to consider increasing the requirement for the American History program. This exemplifies what is meant by "closing the loop" in assessment, which is the use of assessment results to make informed decisions when reforming educational programs.

Suggesting that further research examine whether the gain found in the Both group holds for the typical student illustrates one of the complications in assessment research. How can we obtain a large enough sample of typical students completing both courses when students cannot be assigned to complete their general education requirement in a particular way? The current study was not immune to such a complication. Although we used terms such as "treatment" and "control," it should be kept in mind that students were not randomly assigned to such groups; instead, students self-selected themselves into the various groups. Thus, the differences in gains found among groups are attributable both to the various treatments and to the unique characteristics of students choosing to complete their general education requirement in a particular way.

There are several factors that need to be taken into consideration when interpreting the effect sizes found in our study. First, the fact that 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 coursework their first semester and others the semester previous to the posttest. To explore the effect of semester in which the course was taken on gains, students in the HIST and POSC groups were distinguished by whether they had completed the course the semester prior to the posttest or in the previous academic year. Students who had completed GHIST 225 in the semester prior to posttest had larger effect sizes (d =.67) than students who had completed the course in the previous year (d =.49). Similar results were found for GPOSC 225, with larger gains (d =.49) for students who had recently completed the course and smaller gains (d =.40) for students who had completed the course in the previous year. Although not thoroughly investigated here, the time elapsing between course completion and posttest seems to have an impact on general education gains. If one's sample consists largely of students who have not recently completed the course, effect sizes may appear low. It is therefore important to report results with consideration of the time that has elapsed between

course completion and posttest. This information is also useful in that it allows one to assess the retention of knowledge and skills over 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 be stronger students in general.

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 reach or exceed an acceptable level.

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 students' 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 our 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 and 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.

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

Fong, B. (2004). Looking forward: Liberal education in the 21st century. *Liberal Education, 90*, 8-13.

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.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6,* 107-128.

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

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, A.S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics, 10,* 75-98.

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

Sheu, C-F., & Suzuki, S. (2001). Meta-analysis using linear mixed models. 102-107.

Singer, J.D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 24,* 323-355.

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.