Documentation of Assessment Results: A Guide for Practitioners

AARON J. MYERS, SARA J. FINNEY, and ANDREA M. POPE

STUDENT AFFAIRS ASSESSMENT SUPPORT SERVICES
JAMES MADISON UNIVERSITY

Contents

Purpose and How to Use this Guide

The purpose of this guide is to provide the common vocabulary, text, tables, and figures associated with quantitative results commonly interpreted when assessing program effectiveness. Guidance is facilitated by presenting example interpretations. This document is not intended to be a teaching guide, but instead intended to be a reference guide to assist student affairs practitioners when writing and communicating assessment results that are quantitative in nature. This guide does not address presentation of qualitative data.

Ideally, the reader could use the examples, tables, and figures as templates for their own assessment results. Note, the values reported in this guide were fabricated and are intended to be placeholders only. Further, several examples reference statistics (e.g., $\bar{X}$ *t*, *F*) in both the text and tables or figures. The reader can choose the method of communication that best serves their audience to avoid this redundancy. That is, the text may be clearer by omitting some or all statistics and conveying this information via tables or figures instead.

To assist the reader, the first section contains terms, abbreviations, and definitions one might find in other's reporting of statistical results (e.g., publications, presentations, assessment reports, program self-study, regional accreditation reports) and in statistical computer output. If the reader needs assistance with understanding statistical output from common statistical software packages (e.g., SPSS, SAS), an excellent resource (including annotated output) can be found at the following UCLA website. Further guidance on writing, interpreting, and understanding statistical results can be found in Green and Salkind's (2016) book, *Using SPSS for Windows and Macintosh: Analyzing and Understanding the Data*, available from the JMU library. General guidance on writing and displaying statistical results via tables and figures can be found in the *Publication Manual of the American Psychological Association* (6th ed.), Nicol and Pexman's books, *Presenting Your Findings: A Practical Guide for Creating Tables* (6th ed.) and *Displaying Your Findings: A Practical Guide for Creating Figures, Posters, and Presentations* (6th ed.), and the Purdue Online Writing Lab website.

To facilitate understanding, we use the following example throughout this guide. The Community Service Learning Office is implementing a program intended to influence the number of hours of community engagement activities in which students voluntarily engage. Students are randomly assigned to (a) a group that participates in the program (i.e., experimental group) and (b) a group that will not participate in the program (i.e., control group). For both groups, the number of hours of voluntary community engagement and attitudes toward community engagement were measured at three time points: before the program, midway through the program, and after the program.

Terms & Abbreviations

- **Variable**: Any attribute of a person or the environment that can take on different values.
  - Level of a variable: A value that a variable can take on.
    - For example, the Community Service Learning Office measured students' attitudes toward community engagement (e.g., the *variable* of interest). This variable is measured using items where 1 = *Strongly Disagree*, 2 = *Disagree*, 3 = *Neutral*, 4 = *Agree*, and 5 = *Strongly Agree* (i.e., the *levels* of the variable). Levels of a variable can be categorical. For example, participation in the community engagement program is a variable: participation (coded as 1) versus non-participation (coded as 0) would be the levels.
- **Population**: All cases with specified characteristic(s) to which one wants to generalize. In the community engagement example, all JMU students may be considered the population to which the Community Service Learning Office wants to generalize.
  - Parameter: An index of a population generally symbolized with Greek letters (e.g., μ, σ).
    - For example, the population mean (μ) is a parameter.
- **Sample**: A subset of a population. In the community engagement program example, we have a sample of students who participated in the program and a sample of students who did not participate in the program.
  - Statistic: An index of a sample.
    - For example, $\bar{X}$, *SD*, *t* are sample statistics that take on different values (e.g., $\bar{X} = 5.00$).
- **Case**: A unit of analysis. Entity on which a variable is measured. For the example, a student is a case.
- **Hypothesis**: A statement about the relation between variables.
  - Null (H_O): The null hypothesis is generally a statement specifying *no relation* between variables or *no difference* between groups in the population with respect to the outcome variable. For our example, we have two groups: program participants and program non-participants (i.e., control group). The null hypothesis would be participation in the program has no influence on students' average number of hours of community engagement activities in the population (H_O: $\mu_{participated} = \mu_{control}$ or, equivalently, H_O: $\mu_{participated} - \mu_{control} = 0$).
  - Alternative (H_A): The alternative hypothesis is generally a statement specifying a *relation* between variables or a *difference* between groups in the population. For our example, the alternative hypothesis would be participation in the program has an influence on average number of hours of engagement (H_A: $\mu_{participated} \neq \mu_{control}$ or, equivalently, H_A: $\mu_{participated} - \mu_{control} \neq 0$).

- **Descriptive statistics**: Indices that describe characteristics of a data set. May apply to either a population or a sample.
  - Examples: frequency of occurrence; indices of central tendency such as mean, median, and mode; indices of variability such as standard deviation.
  - Examples from our scenario: the number of students participating in the program (i.e., frequency), the average number of hours of community engagement activities the students engaged in, and the variability (e.g., standard deviation) in the number of hours.
- **Inferential statistics**: Indices from statistical significance tests used to make generalizations from a sample to infer the characteristics of a population.
  - For example, 200 JMU students were randomly assigned to either participate in the program or not participate in the program. Our goal is not to simply describe characteristics of the 200 sampled students (e.g., descriptive statistics); we would like to make an inference about how the program may have influenced the population of JMU students.
- **Alpha** (α): An a priori user-specified value (commonly, $\alpha = .05$) which represents the probability of incorrectly rejecting the null hypothesis. In other words, based on our sample, we are willing to accept a 5% chance of concluding there *is* a difference between groups in the population with respect to the outcome variable, when in reality there is not a difference between the groups in the population.
- **Statistical significance**: Very basically, statistical significance refers to how unlikely the results from a sample are, assuming the null hypothesis is true. For example, for the community engagement program scenario, imagine those participating in the program averaged three more hours of community engagement activities after the program than before the program. Results from a statistical significance test determine how likely a difference of three hours is for this sample if there was *truly* no difference in hours in the population. Generally, *p*-values from statistical tests are compared to an a priori specified alpha (α). If the resultant *p*-value is greater than the alpha value, we presume the results are not unlikely given the null hypothesis is true in the population (i.e., the difference of three hours is not statistically significant; no statistical difference in hours from before to after the program). If the resultant *p*-value is smaller than the alpha value, we presume the results are unlikely given the null hypothesis is true in the population (i.e., the difference of three hours is statistically significant; there was a statistical difference in hours from before to after the program).
- **Practical significance**: Refers to whether the relations between variables or the differences between groups is practically meaningful. Practical significance is in the eye of the beholder. Only content experts (i.e., those creating programming who know the theory associated with intended outcomes)

can judge if an effect is practically meaningful *given the variables under study and characteristics of the programming*. Practical significance is typically assessed via effect sizes.

- Effect sizes are quantitative indices of the strength of the effect (e.g., relations, differences).
    - *Unstandardized effect sizes* are on the metric of the variable of interest. For example: hours of community engagement activities increased, on average, from two to five. Is an increase of three hours of community engagement activities practically meaningful or practically important? This unstandardized effect size must be interpreted considering the context of the study. Practitioners may ask themselves: Would I expect or hope for a larger increase considering the length and strength of the program? After reviewing the literature, what effect sizes have others found? What effect size justifies the money, time, and energy to implement this programming? Effect sizes *should be* stated when articulating program outcomes (i.e., the "degree" in the ABCD method of writing objectives is asking for effect size articulation).
    - *Standardized effect sizes* are on a standard, or common, metric independent of the metric of the variable of interest. For example, imagine we wanted to compare participation in community engagement activities from two studies. Study A measured the number of hours a student participated in community engagement activities. Study B measured the number of minutes a student participated in community engagement activities. Study A's unstandardized effect size of three hours is not directly comparable to Study B's unstandardized effect size of two-hundred-ten minutes. Fortunately, we can compare the average differences found in the two studies by evaluating their standardized effect sizes.
        - *d* (Cohen's *d*) is a standardized effect size used to index average differences. It indicates the difference in means in standard deviation units.
            - For example, the participant group's average number of hours was 0.50 standard deviation units higher ($d = 0.50$) after completion of the program than before the program.
        - R^2 or η^2 (i.e., eta-squared) is standardized effect size used to index variance explained in the outcome variable. Often, they are multiplied by 100 and reported as a percentage of variance accounted for.
            - For example, attitudes toward community engagement scores account for 10% of the variance in hours of community engagement ($R^2 = .10$).
            - For example, 20% of the variance in number of hours can be explained by if a student participated or not in the community service program ($\eta^2 = .20$).

Descriptive Statistics

- *N*: The number of cases (e.g., students) in the total sample.
- *n*: The number of cases in a subset of the total sample.
    - Of the total sample of students (*N* = 200), a subset of students were female (*n* = 100).
- Minimum (Min): The minimum observed score.
- Maximum (Max): The maximum observed score.
- Central Tendency Statistics:
    - Mean (i.e., *M*, $\bar{X}$, μ): The arithmetic average of a variable.
    - Mode: The most frequently occurring value.
    - Median (i.e., *Mdn*): The middle, or centermost number, of a sorted list of numbers.
        - Consider the values: 1, 2, 2, 4, and 6. Statistics are *M* = 3, *Mode* = 2, *Mdn* = 2.
        - If there is an even number of observations in the sample, the median will be the arithmetic average of the two centermost values.
- Variability Statistics:
    - Variance (i.e., s^2, σ^2): A measure of variability of scores on the squared-metric of the variable. As a result, variance is not easy to interpret and, therefore, may not be reported.
    - Standard deviation (i.e., *SD*, *s*, σ): A measure of variability of scores on the metric of the variable. Conceptually, it is the average deviation of observed scores from the mean.
        - For example, on average, students participated in 9 hours of community engagement activities (*M* = 9.00). The standard deviation (*SD* = 3.00) can be interpreted as, students' typical hours of engagement deviated about 3 hours below (i.e., 6 hours) to about 3 hours above (i.e., 12 hours) the mean of 9 hours (see Table 1). Relatively higher standard deviation values indicate more variability in scores. Standard deviations from different scales should not be compared, as they are reflective of the metric of the variable of interest.[1]

Table 1

Hours of Community Engagement Activities by Participant Group

		Hours			
Group	*n*	Minimum	Maximum	Mean	*SD*
Participant (Experimental)	100	1.45	23.47	9.00	3.00
Non-Participant (Control)	100	0.00	11.11	5.06	2.20

[1] For example, the metric of annual family income (e.g., *SD* = $10,000) is much larger than the metric of years of higher education (e.g., *SD* = 4).

Reliability of Scores

- **Internal Consistency Reliability.** Internal consistency reliability refers to the consistency or dependability of scores (e.g., Cronbach's coefficient alpha [i.e., α]). In general, consistency of scores refers to the repeatability of scores. That is, if we were to use a sample of students to measure an outcome (i.e., hours of engagement) repeatedly, we could index the reliability of those scores on the outcome by examining if the scores were relatively equivalent across repeated measurements. Internal consistency reliability is often used to index the interrelatedness of items on a test. Coefficient alpha ranges from values of 0 to 1, with higher values indicating higher reliability.
    - For example, imagine two students complete the attitudes toward community engagement scale. Examinees respond to 4 items using a 5-point Likert response scale (e.g., 1 = *Strongly Disagree*, 5 = *Strongly Agree*), with higher values indicating more favorable attitudes toward community engagement. Callie receives a relatively high total score and Avery receives a relatively low total score. Given the scale has high internal consistency reliability, we would expect Callie to score relatively high on most items on the scale and Avery to score relatively low on most items.
- **Test-Retest Reliability.** Test-retest reliability refers to the consistency, or stability, of scores across different time points. That is, test-retest reliability quantifies the extent to which the rank order of students' scores remains stable over time. One might assess test-retest reliability in situations where scores will be collected at different time points.
    - For example, the CSL office is interested in assessing students' attitudes toward community engagement over the course of a month. They measured attitudes at the beginning of each week. They then correlated the scores from week 1 and week 2. A relatively high correlation (i.e., test-retest reliability) would indicate students did not tend to change rank order from week 1 to week 2. Relatively low test-retest reliability would indicate students changed differentially over time.
- **Interrater Reliability.** Interrater reliability refers to the consistency of ratings obtained from different raters when rating performance assessments (e.g., essay, speech).
    - For example, Luna and Riley each rate 3 students' 30-second elevator pitches regarding the importance of community engagement. Luna assigns the following ratings using a 5-point scale: Esme 4, Luis 3, and Yasin 2. Riley assigns the following ratings using the 5-point scale: Yasin 4, Luis 3, and Esme 2. Notice Luna and Riley are not *consistent* with respect to the rank order of ratings; thus, the interrater reliability would be relatively low. Had Luna and Riley rated the students equally, interrater reliability would be relatively high.

Reliability is a critical property of scores that must be evaluated and reported. Given low reliability of scores, statistical results cannot be trusted and thus interpretations may not be justified. Moreover, the *Standards for Educational and Psychological Testing* assert that evidence of reliability should be collected and reported to justify interpretation of each intended score use (AERA, APA, & NCME, 2014).

Tests of Mean Differences across Groups

Comparison of Averages across Two Groups

Perhaps we want to compare average community engagement hours between two groups (those who participated in programming and those who did not) to provide evidence of program effectiveness. One could use central tendency statistics (e.g., mean, median, mode) to describe the sample, or one could use inferential statistics with the goal of making inferences to a larger population. The most common inferential test used to compare average differences between two groups is an independent samples *t*-tests.

Independent Samples *t*-test. Statistical test used to compare average levels of a variable across 2 independent groups to determine if the groups' averages are statistically significantly different.

- Example Write-Up: An independent samples *t*-test was used to determine whether students who participated in a community engagement program (i.e., experimental group) voluntarily engaged in more hours of community activities than students who did not participate in the program (i.e., control group). Average hours of community engagement for those who participated in the program ($M = 9.00$) was statistically significantly higher than hours for those who did not participate ($M = 5.06$), $t(198) = 2.22$, $p = .01$ (see Table 2). The 95% confidence interval bounding the 3.94-hour difference in means is narrow, ranging from a 2.50 to 5.00-hour difference between the groups. The interval does not include the value of 0 as a plausible difference in means. The effect size ($d = 0.31$) indicates a moderate effect of the program on hours of community engagement. Practitioners deem the unstandardized effect size (3.94-hour increase) to be a practically meaningful effect of the program on hours of community engagement. Thus, results indicate a statistically and practically significant effect of the community engagement program on hours of community engagement (See Figure 1).

Table 2

Hours of Community Engagement Activities by Participant Group

		Hours				
Group	*n*	Minimum	Maximum	Mean	*SD*	95% CI
Participant (Experimental)	100	1.45	23.47	9.00	3.00	[8.40, 9.60]
Non-Participant (Control)	100	0.00	11.11	5.06	2.20	[4.62, 5.50]

Note. 95% confidence intervals (CIs) around sample mean hours of engagement. The CI represents a range of plausible values of the true mean in a given population (e.g., participant, non-participant).

Confidence Intervals. Reporting confidence intervals is standard in statistical reporting. Confidence intervals represent a range of plausible values of the true population parameter (e.g., M, $M_{difference}$, b). The specific confidence interval (e.g., 95%) reflects the a priori user-specified alpha value (e.g., if $\alpha = .05$, then $1 - \alpha = .95$, and .95*100 = 95%). A confidence interval can be used to test specific values of a null hypothesis.

- For an example of a confidence interval around a mean difference, consider the null hypothesis specifying that participation in the community engagement program has no influence on students' average number of hours of community engagement activities (H_O: $\mu_{participant} - \mu_{control} = 0$). If we find an average difference of 3.94 hours of community engagement activities between the participant and non-participant (control) groups in our sample ($M_{difference} = 3.94$), given the estimated 95% CI around our mean difference (95% CI [2.50, 5.38]), we would conclude the null hypothesis (H_O: $\mu_{participant} - \mu_{control} = 0$) is not plausible in the population. That is, the interval of 2.50 to 5.38 does not include 0; thus, we reject the null hypothesis that the population mean difference is zero. However, if we find an average difference of 1 hour of community engagement between the participant and non-participant groups ($M_{difference} = 1.00$), and a 95% CI [-0.25, 2.25], we would conclude the null hypothesis (H_O: $\mu_{participant} - \mu_{control} = 0$) *is* plausible in the population because the interval of -0.25 to 2.25 *does* include 0; thus, we would *fail* to reject the null hypothesis.

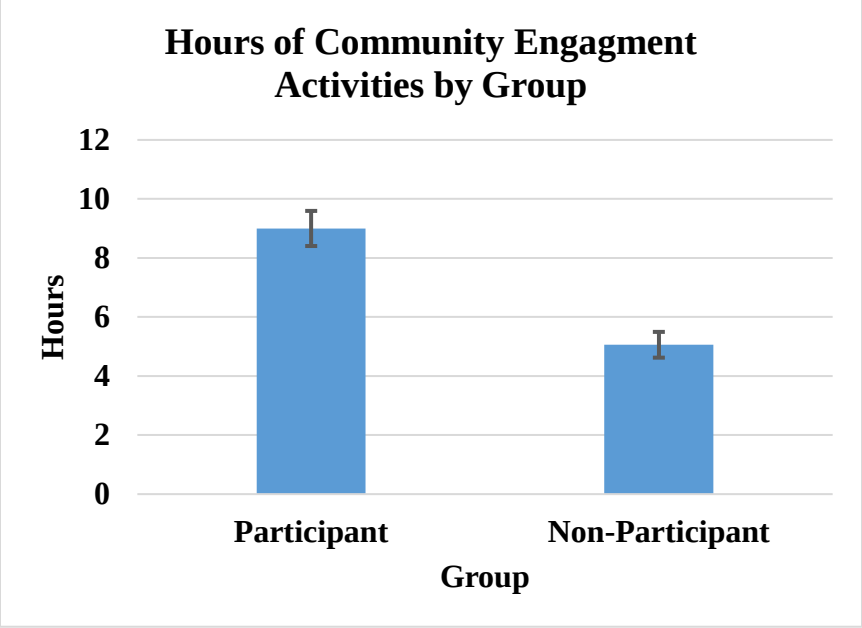

Figure 1. Hours of community engagement by participation group. Error bars represent 95% confidence intervals around mean hours of engagement. Average hours of engagement for students in the participant group, $M = 9.00$, 95% CI [8.40, 9.60], and for the non-participant (control) group, $M = 5.06$, 95% CI [4.62, 5.50]. No overlap in the 95% confidence interval bars suggests statistically significantly different average hours of engagement across groups.

Comparison of Averages across Three or More Groups

Perhaps we want to compare average community engagement hours between *three or more* groups.

One-Way Between-Subjects ANOVA. A statistical test used to compare average levels of a variable across three or more different, independent, groups.

- Example Write-Up: A one-way between-subjects analysis of variance (ANOVA) was conducted to determine whether community engagement program (new program, established program, no program) was related to students' hours of voluntary community engagement. The results indicate that hours of voluntary community engagement were statistically significantly related to type of community engagement program, $F(2, 56) = 5.33$, $p = .04$. The effect size ($\eta^2 = .08$) indicates 8% of the variance in hours of community engagement can be explained by type of community engagement program. To evaluate which programs were significantly different, Tukey HSD post hoc comparisons were conducted. The results indicated that hours of community engagement for those students in the control group ($M = 5.25$) were statistically significantly lower than those students participating in the established ($M = 8.52$) and new engagement programs ($M = 9.00$), $p = .01$ and $p = .02$ respectively (see Table 3 and Figure 2). Students' participating in the established and new programs were not significantly different with respect to number of hours, $p = .06$.

Table 3

Hours of Community Engagement Activities by Group

		Hours				
Group	*n*	Minimum	Maximum	Mean	*SD*	95% CI
New Program	100	1.45	23.47	9.00_a	3.00	[8.40, 9.60]
Established Program	100	3.21	22.13	8.52_a	2.11	[8.10, 8.94]
No Program (Control)	100	0.00	11.11	5.25_b	2.20	[4.81, 5.69]

Note. Means with no subscripts in common are statistically significantly different, $p < .05$.

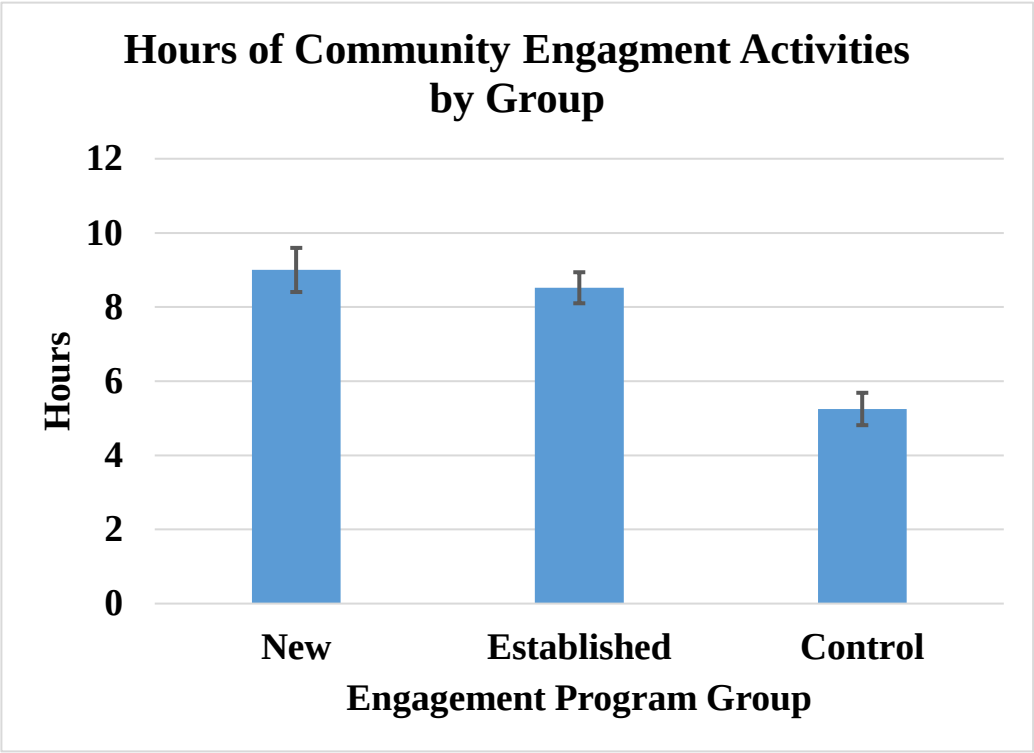

Figure 2. Hours of community engagement activities by program group. Error bars represent 95% confidence intervals around average hours of engagement for students participating in the new program, *M* = 9.00, 95% CI [8.40, 9.60], established program, *M* = 8.52, 95% CI [8.10, 8.94], and no program (control), *M* = 5.25, 95% CI [4.81, 5.69]. No overlap in 95% confidence interval bars between the control group and new and established program groups suggests statistically significantly different average hours of engagement. Overlap in 95% confidence interval bars between new and established program groups suggests their average hours of engagement are not statistically significantly different.

Comparisons of Averages across Two Variables with Two or More Levels

If we want to compare average hours of community engagement across two (or more) categorical between-subjects variables with *two or more* levels, factorial between-subjects analysis of variance (ANOVA) is an option. For example, one could compare average hours across year in college (first-year and sophomore) and community engagement group (new, established, and control).

Factorial Between-Subjects ANOVA. A statistical test used to compare average levels of an outcome variable across two or more grouping variables with two or more levels.

- Example Write-Up: Students' hours of community engagement were examined using a 2 (year in college: first-year and sophomore) by 3 (community engagement program: established, new, and control) factorial analysis of variance (ANOVA). The interaction between year in college and engagement program on hours of engagement was nonsignificant, $F(2, 98) = 1.71$, $p = .19$ (see Table 4). The effect size ($\eta^2 = .03$) indicates 3% of the variance in hours of engagement can be explained by the interaction between year in college and engagement group. The main effect of year in college on hours of engagement was nonsignificant, $F(1, 98) = 1.92$, $p = .06$; $\eta^2 = .04$, indicating year in college has little influence on hours of engagement. There was, however, a statistically significant main effect of engagement program on hours of engagement, $F(2, 98) = 9.15$, $p = .01$; $\eta^2 = .16$, indicating engagement program has a moderate effect on hours of engagement. Post hoc Tukey HSD test was conducted to assess which engagement programs differed from one another with respect to hours. Students in the control group participated in statistically significantly fewer hours of community engagement ($M = 5.25$) than students in the established ($M = 8.52$, $p = .001$) and new engagement program groups ($M = 9.00$, $p < .001$; see Table 5 and Figure 3). Community engagement hours did not differ between students participating in the established program and students in the new program, $p = .74$.

Table 4

Community Engagement Hours as a Function of Year in College and Program Group

Effect	*df*	*F*	*p*	η^2
Year in College Main Effect	1, 98	1.92	.06	.04
Program Group Main Effect	2, 98	9.15	.01	.16
Year by Group Interaction	2, 98	1.71	.19	.03

Table 5

Average Community Engagement Hours as a Function of Year in College and Program Group

Year	Engagement Program Group		
	New	Old	Control
First-Year	8.70_a	8.62_a	5.15_b
Sophomore	9.30_a	8.22_a	5.75_b

Note. Means within rows or within columns with no subscripts in common are statistically significantly different, $p < .05$.

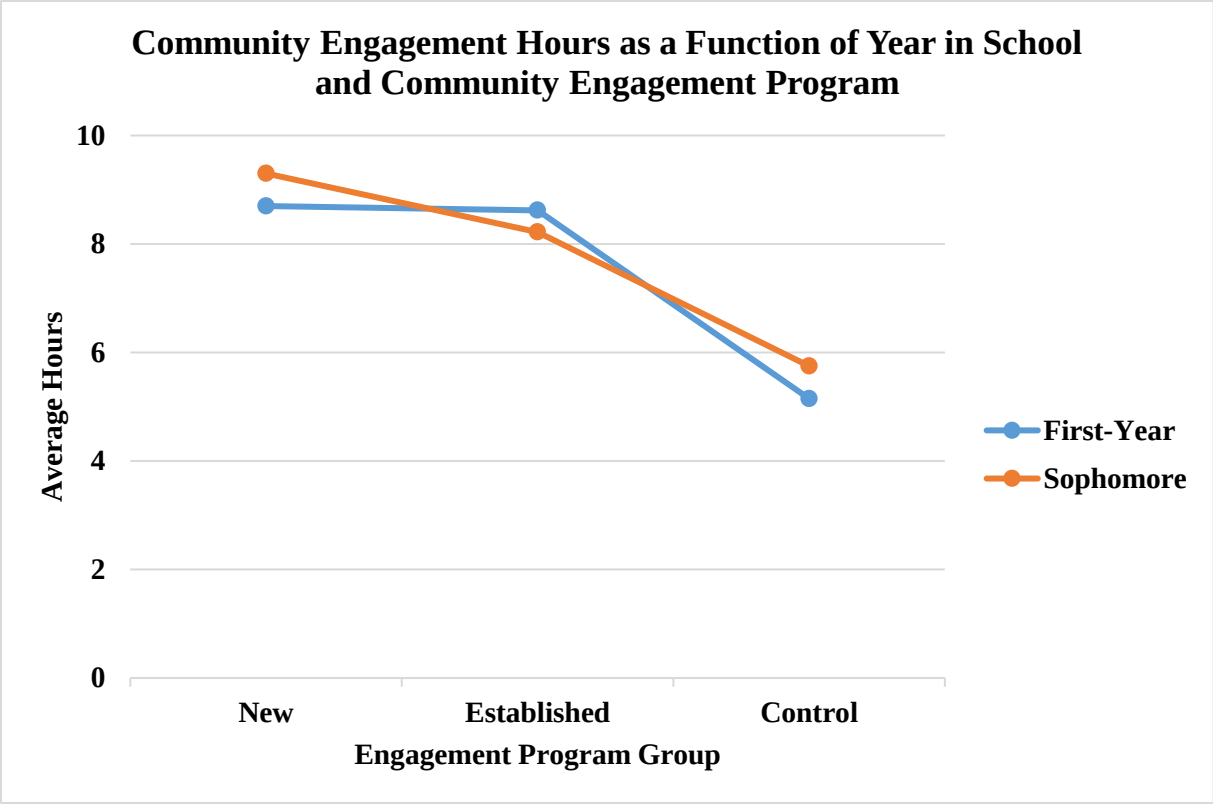

Figure 3. Average hours of community engagement by year and engagement program type. Notice the average difference in hours of engagement between first-year and sophomore students follows essentially the same pattern across program group, which indicates year in college and engagement group do not interact with respect to hours of engagement.

Comparison of Averages over Time

Perhaps, for those students participating in the engagement program, we want to compare average community engagement hours across *two* time points (e.g., before and after completing the program). The most common inferential test used to compare average differences over two time points is a dependent samples *t*-tests (also termed repeated-measures *t*-test or within-subjects *t*-test). Or, perhaps we want to compare average community engagement hours across *three or more* time points to determine if hours increased or decreased at various stages throughout the program. The most common inferential test used to compare average differences over three or more different time points is a repeated measures analysis of variance (ANOVA). Or, we may want to compare average hours of engagement across *two or more* groups (i.e., program vs. no program; between-subjects) and over *two or more* occasions (i.e., time; within-subjects). A factorial mixed-subjects analysis of variance (ANOVA) allows us to compare the outcome across levels of the between-subjects variable (i.e. program type) and across levels of a within-subjects variable (i.e., time). The latter is a very common approach to evaluate program effectiveness, as it assesses if there is differential change in the outcome depending on if students experienced or did not experience the program.

Repeated Measures *t*-test. (i.e., paired samples *t*-test) A statistical test used to compare average levels of an outcome variable from the same group of participants at two occasions (e.g., pretest, posttest), to determine if the groups' averages are statistically significantly different.

- Example Write-Up: A dependent samples *t*-test was conducted to evaluate change in students' hours of community engagement. The results indicated that students' hours of engagement before the program ($M = 5.25$) were statistically significantly lower than students' hours of engagement after the program $M = 9.00$, $t(99) = 4.97$, $p < .001$ (see Table 6 and Figure 4). Moreover, the 95% confidence interval of the difference in means [1.75, 5.25] indicates 0 is not a plausible difference in average hours. The effect size ($d = 0.45$) indicates students hours of engagement increased by 0.45 standard deviation units from before to after the program.

One-Way Repeated Measures ANOVA. A statistical test used to compare average levels of a variable across three or more different time points.

- Example Write-Up: A one-way repeated measures analysis of variance (ANOVA) was conducted to determine if average hours of community engagement activities changed throughout the program (before, midway, and after program completion). The results indicated average hours of engagement changed over time $F(2, 354) = 205.48$, $p < .001$ (see Figure 4). The effect size ($\eta^2 = .54$) indicates 54% of the variance in hours of engagement can be accounted for by time in the program. Tukey's HSD test assessed which time of measurements differed with respect to hours of engagement. Students

participated in more hours of community engagement activities after the program (M = 9.00) than midway through the program (M = 7.65, $p < .001$), and participated in more hours of activities midway through the program than before the program, M = 5.25, $p < .001$ (see Table 6). Thus, the results indicate that students' hours of community engagement tended to increase throughout the program.

Table 6

Average Community Engagement Hours as a Function of Participation in Engagement Program

Statistic	Time of Measurement		
	Before	Midway	After
Mean	5.25_a	7.65_b	9.00_c
SD	2.34	2.81	3.00

Note. Means with no subscripts in common are statistically significantly different, $p < .05$.

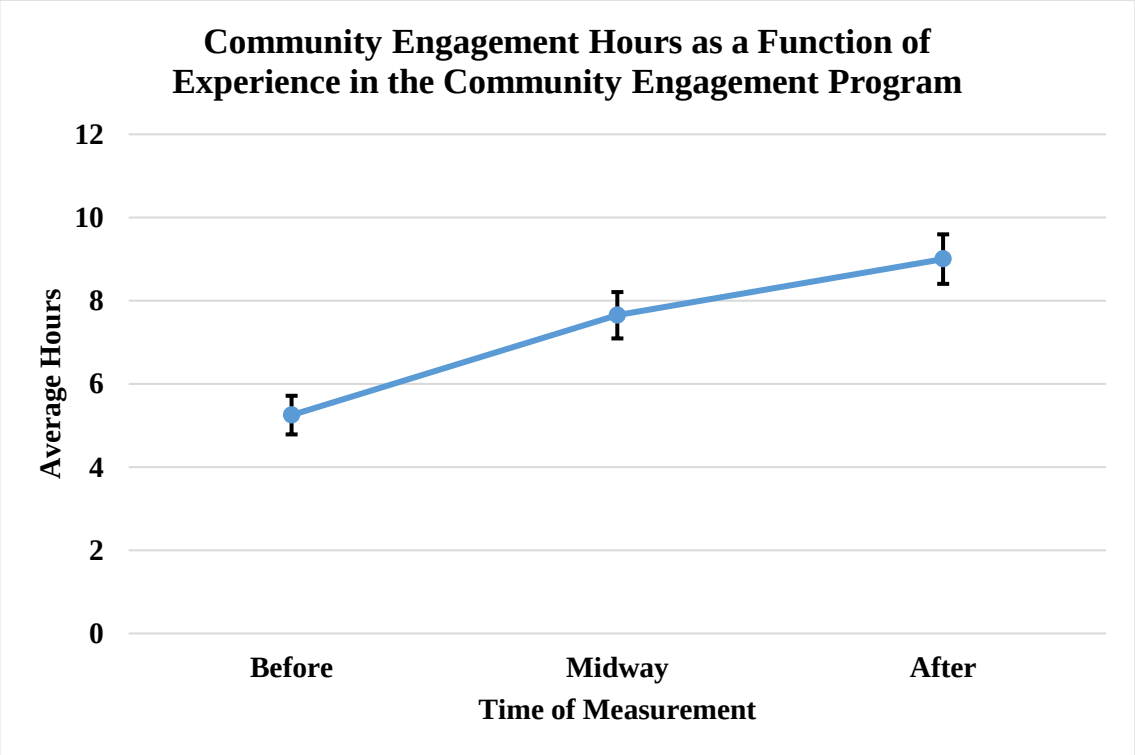

Figure 4. Hours of community engagement by time of measurement. Error bars represent 95% confidence intervals around average engagement hours. No overlap in the 95% confidence intervals suggests statistically significantly different average hours of engagement among times of measurement.

Factorial Mixed-Design ANOVA. A statistical test used to compare average levels of a variable across one or more between-subjects variables and one or more within-subjects variables.

- Example Write-Up: Students' hours of voluntary community engagement activities were examined using a 2 (participant and non-participant group; between-subjects) by 3 (time of measurement; within-subjects) mixed design factorial analysis of variance (ANOVA). The interaction between group and time of measurement on hours of engagement was statistically significant, $F(2, 101) = 6.75$, $p = .02$, which indicates the effect of time of measurement on hours of engagement depends on group membership. The effect size (partial $\eta^2 = .33$) indicates this interaction effect accounts for 33% of the variance in hours of engagement, after controlling for group and time of measurement. Given the statistically significant interaction effect, the main effects were not directly interpretable (see Table 7). Tests of simple main effects were conducted to probe the interaction. The simple main effect of time for the non-participant group was not statistically significant. That is, for students in the non-participant group, average community engagement hours were not statistically significantly different after completion of the program ($M = 5.06$) than their hours midway through the program ($M = 5.21$) or before the program ($M = 4.71$). The simple main effect of time was statistically significant for the program participant group. Thus, for the program participant group, Tukey's HSD test was conducted to assess which times of measurement differed from one another with respect to hours of engagement. For program participants, average community engagement hours were statistically significantly higher after completion of the program ($M = 9.00$) than their hours midway through the program ($M = 7.65$), which were statistically significantly higher than hours before the program ($M = 5.25$; see Table 8 and Figure 5). Thus, average hours of engagement increased throughout the program for program participants but remain essentially unchanged for non-participants.

Table 7

Community Engagement Hours as a Function of Group and Time of Measurement

Model	*df*	*F*	*p*	η^2
Group	1, 101	8.42	.02	.44
Time	2, 101	4.86	.03	.24
Group by Time Interaction	2, 101	6.75	.02	.33

Table 8

Average Community Engagement Hours as a Function of Group and Time of Measurement

Group	Time of Measurement		
	Before	Midway	After
Program Participants	5.25_{a}	7.65_{b}	9.00_{c}
Non-Participants (Control)	4.71_{a}	5.21_{a}	5.06_{a}

Note. Means within rows or within columns with no subscripts in common are statistically significantly different, $p < .05$. Means with subscripts in common across time of measurement within the non-participant group indicate the simple main effect for control group was not statistically significant. Different subscripts across time of measurement within the program participant group indicate a statistically significant simple main effect for the participant group; thus, post hoc comparisons were needed to test for statistically significant differences in average hours across the three times of measurement.

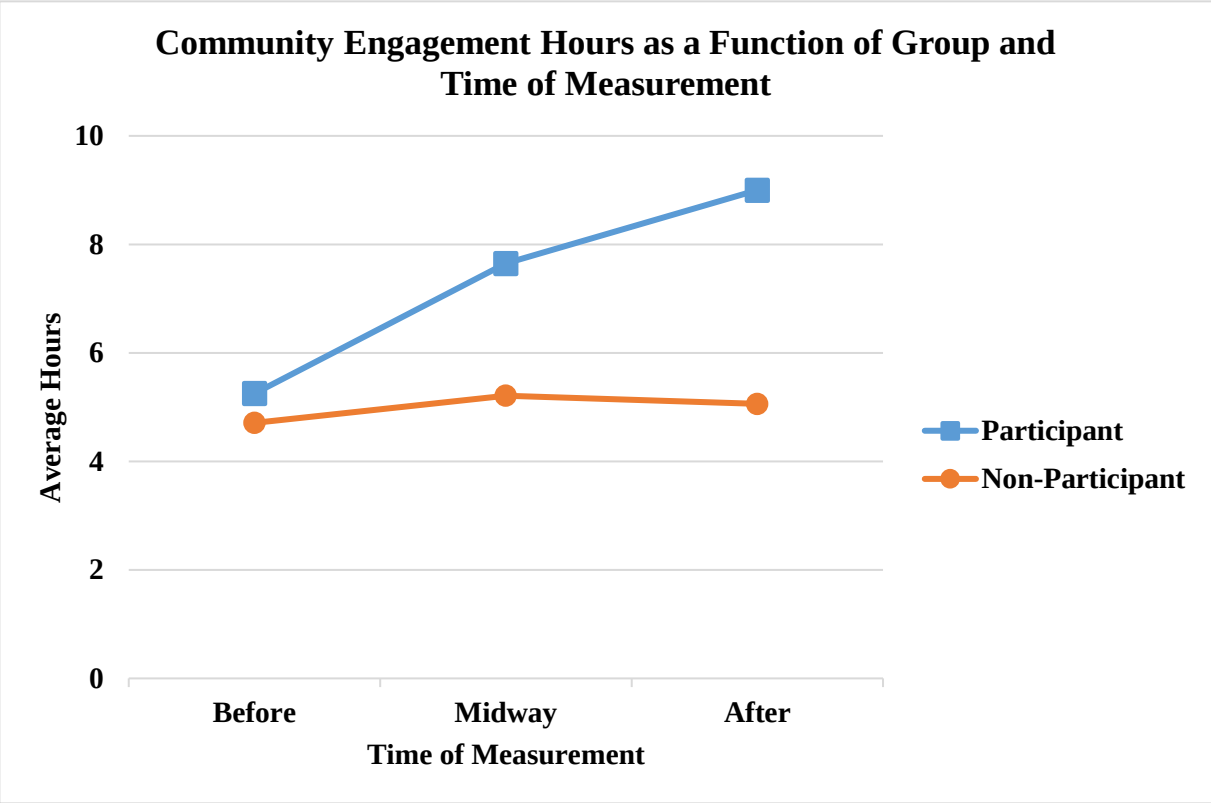

Figure 5. Hours of community engagement by group and time of measurement. Notice average hours of engagement increase throughout the program for the program participant group, but remain essentially unchanged for the non-participant group, which suggests an interaction between program group and time of measurement with respect to hours of engagement.

Evaluating the Relation between Two or More Continuous Variables

If we are interested in evaluating the relation between two continuous variables (such as hours of community engagement activities and attitudes toward community engagement scores), then we could estimate a correlation and/or conduct a regression analysis.

Correlation

Pearson Correlation. A statistical index used to evaluate the *strength* and *direction* of a linear relation between two continuous variables. Coefficients range from -1.00 to 1.00, with higher absolute values indicating a stronger relation. Values closer to zero indicate weaker relations.

- Example Write-Up: A Pearson product-moment correlation was estimated to determine the extent to which students' hours of community engagement activities are linearly related to attitudes toward community engagement scores (see Figure 6 for scatterplot illustrating a positive linear relation). Descriptive statistics can be found in Table 9. The correlation coefficient was statistically significantly different from 0, $r(98) = .61$, $p < .001$. The effect size ($R^2 = .36$) indicates attitudes toward community engagement scores account for 36% of the variance in hours of community engagement. Thus, the strong positive relation indicates that as attitudes toward community engagement scores increase, hours of community engagement tend to increase (and vice versa).

Table 9

Descriptive Statistics for Attitudes Toward Community Engagement Scores and Hours of Community Engagement Activities

	Hours			
Variable	Minimum	Maximum	Mean	*SD*
Attitude Scores	8.04	17.97	12.72	1.91
Engagement Hours	3.40	14.88	9.93	2.09

Note. Students responded to 4 Attitudes Toward Community Engagement items using a Likert response scale ranging from 1 (*Strongly Disagree*) to 5 (*Strongly Agree*). Thus, attitude scores can range from 4 to 20 with higher scores indicating more favorable attitudes toward community engagement.

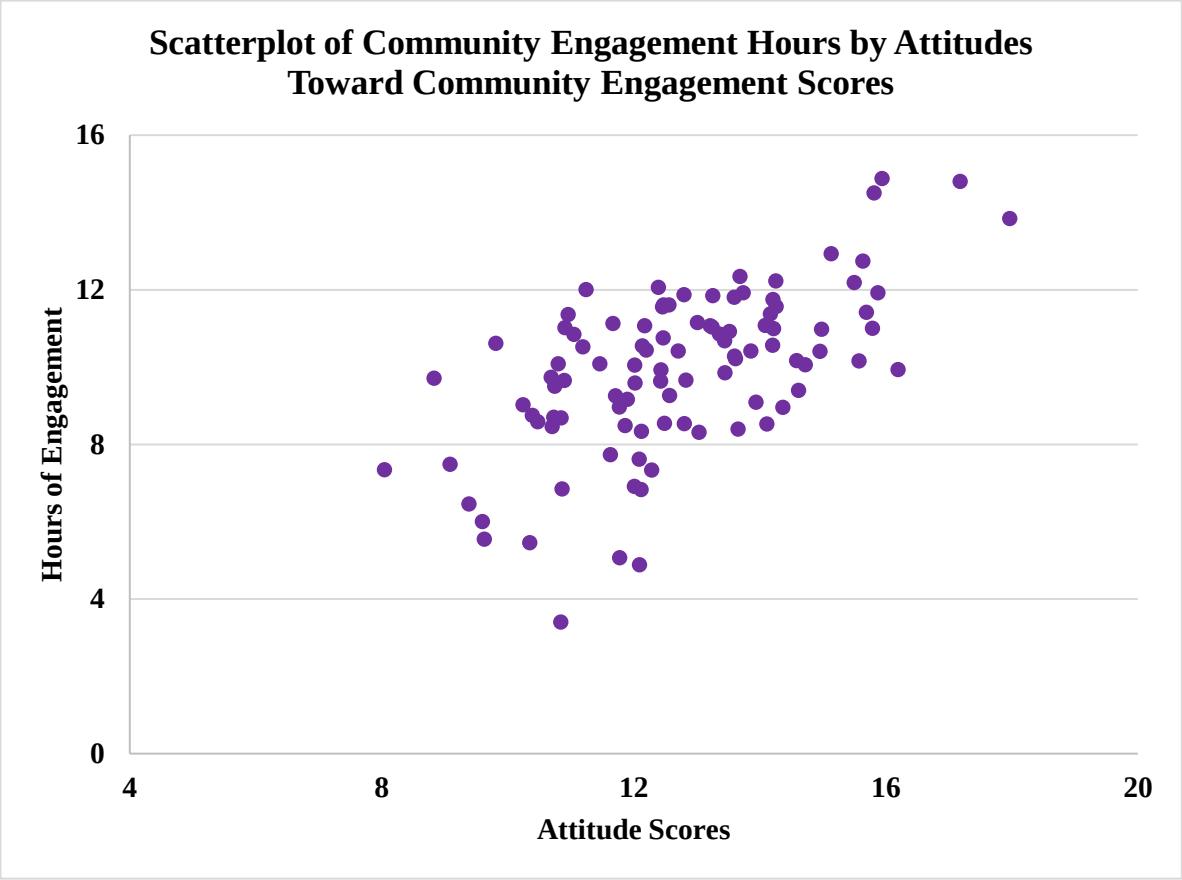

Figure 6. Plot of hours of community engagement by attitudes toward community engagement scores. Notice the positive linear relation between hours of engagement and attitude scores (r = .61). The positive relation indicates that as attitudes toward community engagement scores increase, hours of community engagement tend to increase (and vice versa).

Regression

Multiple Regression. A statistical test generally used to evaluate the relation between one or more independent variables (categorical or continuous) and a continuous dependent variable.

- Example Write-Up: A multiple regression analysis was conducted to examine if hours of community engagement activities could be predicted by attitudes toward community engagement scores and if that relation is moderated by students' grade point average (GPA). Examination of the bivariate scatterplots (see Figures 6, 7, and 8) and Pearson product-moment correlations (see Table 10) allows us to foreshadow the results of the multiple regression analysis. As expected, the relation between hours of engagement and attitudes toward community engagement was positive, linear, and statistically significant. The relations between GPA and hours of engagement and between GPA and attitudes toward community engagement were nonsignificant.

  Attitudes toward community engagement, GPA, and their interaction accounted for a statistically and practically significant portion of variance in hours of community engagement, R^2 = .42, $F(3, 96) = 22.90$, $p < .001$ (see Figure 9). The interaction between

attitudes and GPA was not statistically significant, b = -0.06, 95% CI [-0.37, 0.24], p = .69, sr^2 < .01 (see Table 11). Thus, the relation between attitudes toward community engagement and hours of engagement does not depend on student GPA. As hypothesized, attitudes toward community engagement was the strongest and only statistically significant predictor, b = 0.86, 95% CI [0.03, 1.68], p = .04, sr^2 = .03, explaining about 3% of the variance in hours of community engagement after controlling for GPA. The b coefficient can be interpreted as for every unit increase in attitudes toward community engagement, community engagement increases by 0.86 hours, after controlling for GPA. GPA was not a statistically significant predictor of hours of engagement, b = 1.47, 95% CI [-2.42, 5.36], sr^2 < .01.

Table 10

Correlations, Means, and Standard Deviations for Hours of Engagement, Grade Point Average, and Attitudes Toward Community Engagement Scores

Variable	Variable			Statistic	
	Engagement Hours	GPA	Attitude Score	Mean	*SD*
Engagement Hours	--			9.93	2.09
GPA	.14	--		2.60	0.62
Attitude Score	.61*	-.10	--	12.72	1.97

Note. *p < .05.

Table 11

Regression Analysis Predicting Grade Point Average from Hours of Community Engagement Hours and Attitudes Toward Community Engagement Scores

Predictor	*b*	*t*	*p*	β	95% CI of *b*		sr^2
					LL	*UL*	
Intercept	-2.73	0.52	.61	--	-13.25	7.78	--
GPA	1.47	0.75	.45	.43	-2.42	5.36	< .01
Attitudes	0.86	2.07	.04	.79	0.03	1.68	.03
GPA by Attitudes Interaction	-0.06	0.41	.69	-.27	-0.37	0.24	< .01

Note. LL and *UL* represent lower and upper confidence interval limits, respectively, b = unstandardized coefficient, β = standardized coefficient, sr^2 = squared semi-partial correlation.

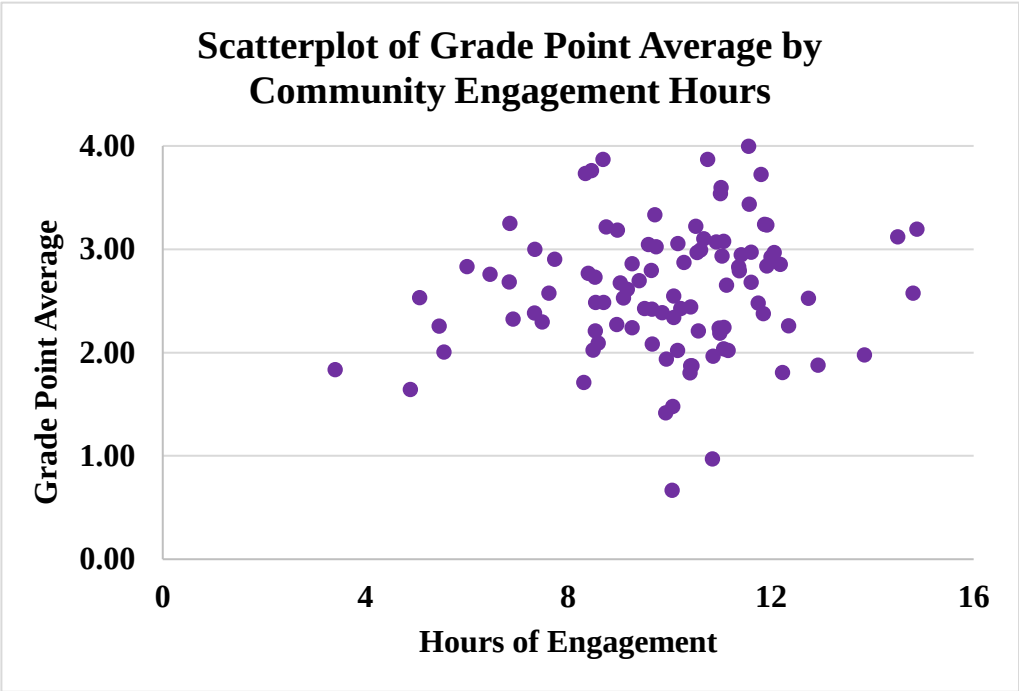

Figure 7. Plot of grade point average by hours of community engagement. Notice a slight, but nonsignificant, positive relation between GPA and hours of engagement (r = .14).

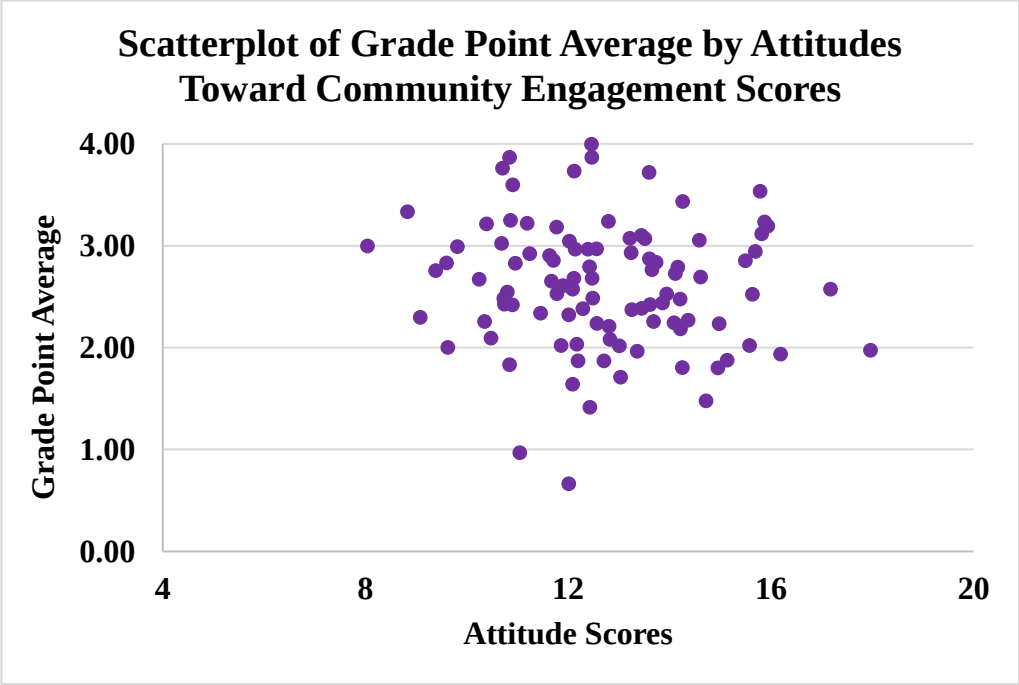

Figure 8. Plot of grade point average by attitudes toward community engagement scores illustrating no relation between GPA and attitudes toward community engagement (r = -.10).

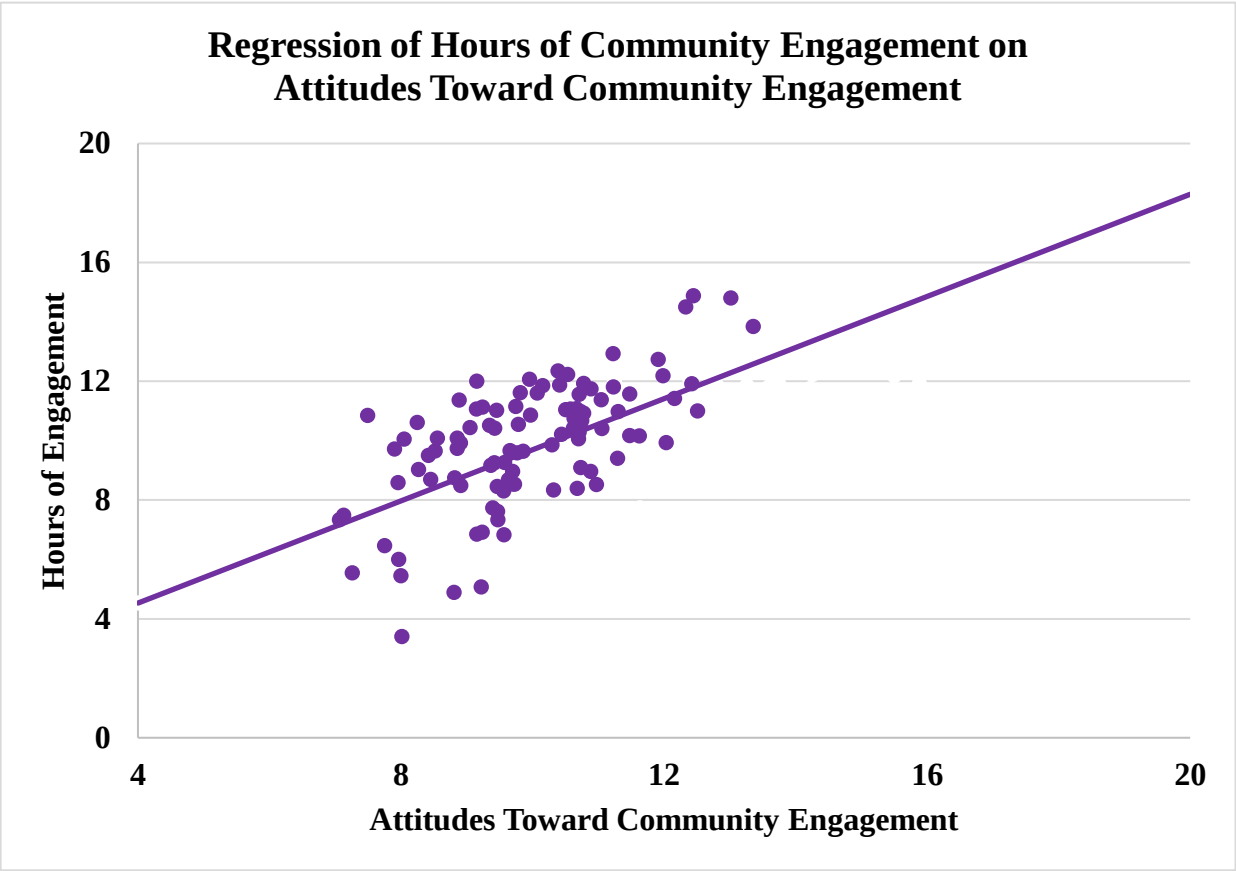

Figure 9. Predicted hours of engagement as a function of attitudes toward community engagement scores, controlling for GPA (i.e., when GPA is held constant at its mean).

Evaluating the Relation between Two Categorical Variables

Let's say students were not randomly assigned to one of three community engagement program groups. Instead, students opted into the program of their choosing. In turn, student characteristics could be related to group membership and thus complicate the interpretation of the impact of programming on hours of community engagement. Accordingly, we want to assess if students' year in college (e.g., first-year, sophomore) is related to their self-selected group. If it is, then any relation between program group and hours of activity could simply be due to year in school. A common statistical test used to evaluate the relation between categorical variables (e.g., self-selected group and year in school) is a chi-square test of independence.

Chi-Square

Chi-square test of independence. A statistical test generally used to evaluate the relation between two categorical variables (e.g., year in school, academic major, gender).

- Example Write-Up: A 2 x 3 chi-square test of independence was conducted to determine whether students' year in college (i.e., first-year or sophomore) was related to their selection of community engagement program groups (i.e., new, established, control). Students' self-assignment to community engagement group was not statistically significantly related to year in college, $\chi^2(2) = 3.50$, $p = .17$. Of the students who self-selected into the new engagement program, 38% were first-year students. Of the students who self-selected into the established community engagement program, 43% were first-year students. Of the students who self-selected into the control group, 54% were first-year students. See Table 11 and Figure 10 for observed and expected frequencies.

Table 12

Frequency of Assignment to Community Engagement Group by Year in School

	Program Group					
	New		Established		Control	
Year in School	Observed	Expected	Observed	Expected	Observed	Expected
First-Year	25	30	29	30	36	30
Sophomore	41	36	38	37	31	37

Note. Expected frequencies represent the expected frequency for each cell if there were no relation between year in school and program group (i.e., null hypothesis is true).

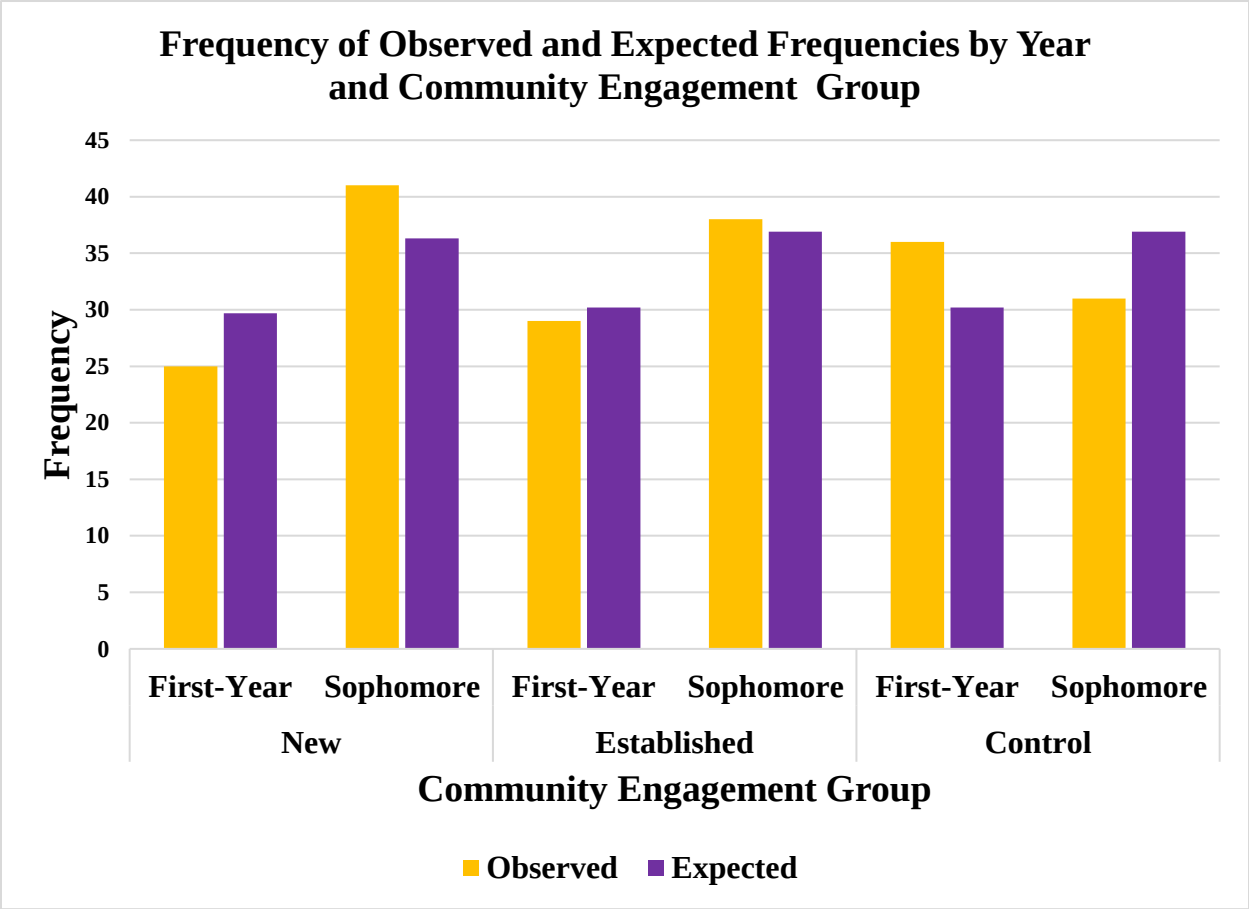

Figure 10. Observed and expected frequencies by year and community engagement group. Given self-assignment to community engagement group was not related to year in school, we see similar observed and expected frequencies associated with year in school and within engagement group. Only compare adjacent columns (year in school) within engagement groups. For example, notice that first-year observed (gold) and first-year expected (purple) frequencies are similar within the new group, established group, and control group, respectively. If self-assignment to community engagement group had been related to year in school, we would have seen relatively large differences between observed and expected frequencies associated with year in school within engagement group.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association. Available at http://www.aera.net/Publications/Books/Standards-for-Educational-Psychological-Testing-2014-Edition

APA Tables and Figures 1 and 2 from the Purdue Online Writing Lab (OWL). Retrieved from https://owl.english.purdue.edu/owl/resource/560/19/

Green, S. B., & Salkind, N. J. (2016). *Using SPSS for Windows and Macintosh: Analyzing and understanding the data* (8th ed.). Prentice Hall Press.

Introduction to SAS. UCLA: Statistical Consulting Group. Retrieved from https://stats.idre.ucla.edu/other/annotatedoutput/

Nicol, A. A. M., & Pexman, P. M. (2010). *Displaying your findings: A practical guide for creating figures, posters, and presentations* (6th ed.). Washington, DC: American Psychological Association. Available at http://www.apa.org/pubs/books/browse.aspx?query=series:APA%20Style%20Series

Nicol, A. A. M., & Pexman, P. M. (2010). *Presenting your findings: A practical guide for creating tables* (6th ed.). Washington, DC: American Psychological Association. Available at http://www.apa.org/pubs/books/browse.aspx?query=series:APA%20Style%20Series

VandenBos, G. R. (Ed). (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association. http://www.apa.org/pubs/books/browse.aspx?query=series:APA%20Style%20Series