

# Report

Tom Gause and Liam O'Brien

2022-05-22

## Introduction

For the past four years at Middlebury College, a small team of students who write for the Middlebury Campus, the school's student newspaper, have designed a survey called the Zeitgeist. The intent of this survey is to get a snapshot of student behavior and attitude on campus. The survey is totally anonymous, and this lets the Zeitgeist team ask a range of sensitive questions such as "Have you ever broken the Honor Code?", "Have you been a victim of sexual assault?", and "Are you happy"? Each year, this survey, which takes 10-20 minutes to complete, is sent to the entire student body. When the responses are collected, the Middlebury Campus analyzes the data and releases an article summarizing some of the significant results of the survey (note that this article shares raw statistics only. The Zeitgeist team avoids any opinion writing and performs minimal statistical analysis). The Zeitgeist project is the largest and most comprehensive student census at Middlebury.

In Spring 2022, the Zeitgeist team released the Zeitgeist 4.0. They collected 1134 responses of 2837 total undergraduates—approximately 40% of the student population—and published a summary article. For our final project, we chose to answer the following question: are the Zeitgeist 4.0 respondents representative of the Middlebury student population?

Our null hypothesis is: The Zeitgeist is a random sample from the Middlebury student population. To evaluate this hypothesis, we used 5 metrics—Class (class year), Geographic Distribution (hometown), Gender (gender identification), Race (ethnicity), and Major (academic)—and collected data for each of these metrics from the Zeitgeist summary post referenced above. For our Middlebury College population data, we collected responses from various administrators and public datasets.

## Methods

As we are testing our hypothesis with a large number of metrics, we pre-planned the following steps to avoid a type I error, i.e. avoid p-hacking. First, we decided to only run permutation tests. We used the formula  $\alpha^* = 1 - (1 - \alpha)^{\frac{1}{k}}$  where  $\alpha = 0.05$  and  $k = 5$  to determine the value  $\alpha^* \simeq 0.0102$  at which the p-value is considered statistically significant for each test. Next, there is some variation in the features of the Zeitgeist and Middlebury Administration college data. To do analysis, we need to combine or remove some features from both dataset. To avoid bias, these data mutations were made before running our first phase of statistical tests. If the p-values from any of the Phase 1 metric tests came up significant, we would engage in a series of Phase 2 tests where we would alter the data assumptions made about the metrics before Phase 1 and re-test against the hypothesis. This process will help ensure any detected significance is not due to falsely made assumptions.

For each of our five categories (geographic location, graduation year, gender, race, and major), we perform a permutation test. In each, we treat the Zeitgeist data like a sample from the administration data, which represents the Middlebury population. Our null hypothesis  $H_0$  is that the Zeitgeist respondents are a random

sample from the Middlebury student population. We use mean absolute error (MAE) as our test statistic for our Zeitgeist data.

$$\text{MAE} = \frac{1}{n} \sum |\text{observed} - \text{expected}|.$$

We are summing over each categorical variable, and  $n$  is the number of categorical variables. The observed count is the number of students in a category, and the expected count is the number of students that would be in the category if the proportion of students in the Zeitgeist sample was equal to the proportion in the population at Middlebury—as provided by the administrative data.

## Data Collection

The Zeitgeist data was collected from this article. It is unknown what data modification/cleaning was done before publishing the article, so we treat this data as ground truth for Zeitgeist. As the Zeitgeist survey was done during the Spring semester in 2022, we retrieved Middlebury administrative metrics collected in the same period. The data from Middlebury’s administration was collected from the *Office of Assessment and Institutional Research*. Their *Spring 2022 Enrollment* document provided us information about the total enrollment at Middlebury by gender and by ethnicity. The Director of Assessment and Institutional Research at Middlebury, Adela Langrock, generously put together geographic data for us; it totals the states of residence for the enrolled students in Spring 2022. We did not have the number of students in each graduating class enrolled in Spring 2022; instead, we multiplied the proportion of each class in Fall 2021 by the total enrollment for Spring 2022. We are assuming that the proportions of each graduating class are the same in the fall and spring semesters. The enrollment by graduation year for the Fall is contained in the *Fall 2021 Student Profile* on the Office of Assessment and Institutional Research webpage linked above.

## Data Modification Prior to Phase 1

All of the questions in the Zeitgeist included an “I prefer not to answer” option. This option was provided for numerous reasons—some respondents may have felt that the options provided did not apply to them, others may have been concerned that an adversary on the Zeitgeist Team would discover their identity from the provided information and use their data in a malicious manner. For the metrics we chose to test, less than 1% of respondents selected the “I prefer not to answer” option. For this reason, we have omitted this category from the data for all metrics for the initial testing phase. When describing the options available to respondents, we will omit the “I prefer not to answer” option.

After removing the “I prefer not to answer” data, the Class Year and Academic Major metrics matched one-to-one, so we deemed these data ready for analysis.

**Geographic Location** We did not receive the number of international students enrolled during the spring of 2022; therefore, we are excluding international students from our statistical analysis.

**Gender** The Zeitgeist included six multiple-choice options for gender-identification: “Cisgender Female”, “Cisgender Male”, “Non-Binary”, “Transgender Male”, “Transgender Female”, and “These Options Don’t Define Me”. Administration only had two categories: Female, Male. We chose to consider only the following statistics: “Cisgender Female” (Zeitgeist) :: “Female” (Admin) and “Cisgender Male” (Zeitgeist) :: “Male” (Admin). This decision was made to uphold statistical rigor and avoid mis-gendering Zeitgeist respondents. We regret that folks whose gender identities are not “Cisgender Female” or “Cisgender Male” were forced to select Female or Male in their College Applications.

**Race** For the Race section of the Zeitgeist, students were given eight options from which they were free to choose as many as they wished. These options were: “White”, “Asian”, “Hispanic or Latino Origin”, “Black or African American”, “Middle Eastern or North African”, “American Indian or Alaskan Native”, and “Native Hawaiian or other Pacific Islander”. There were 1267 responses, of 1134 respondents, meaning at maximum, 153 respondents selected more than one race. From our data, we have no way to determine what selections these 153 respondents made, meaning the data has unexplainable variance. The Middlebury Admin data only allows a single selection. It does not include a category for “Middle Eastern or North African” and includes categories for “International”, “Race and/or ethnicity unknown”, and “Two or more races”. Rather than make assumptions about respondents’ race, we only selected features from datasets that matched one-to-one. As we have no way to remove the variance from respondents who selected multiple races, we must inspect results from permutation testing on this data with extreme scrutiny.

## Phase 2 Testing

If we see any significant differences, we want to change our assumptions and run the test again to ensure that we are not making any false claims based on false assumptions. In this section, we outline our changes in assumptions for Phase 2. Note that this section was written prior to doing any statistical tests on our data.

As our Class Year and Geographic Location metrics matched one-to-one after removing the “I prefer not to answer” and “International” categories so we were not making any significant assumptions and will not run Phase 2 testing for these categories.

**Gender** We know that Zeitgeist surveyed respondents with six multiple-choice options for gender identification. Middlebury, however, only had two options to fill in: “Female” and “Male”. For our Phase 1 testing, we just ignored the Zeitgeist responses that were neither “Female” nor “Male”. In Phase 2 testing, we will make the assumption that people in other gender categories were forced to respond either “Female” or “Male”. We will assume that such a respondent’s choice of “Female” or “Male” was random with probability 0.5 of choosing either category. Our reasoning is that when non-binary people apply to Middlebury and have to select one of two categories which don’t apply to them, they will pick randomly. After randomly re-assigning gender non-binary people, we will run our statistical test again.

**Race** For Phase 1 testing, we made some big assumptions about how we should test our data (see data modification above). For Phase 2, we want to try a different approach. We will assume that those in Admin Data who selected “Two or more races” would have selected exactly two races if they were given the Zeitgeist survey. We will randomly assign each “Two or more races” selection in the administration data to two of the race categories according to the distribution at Middlebury given by the Admin Data (excluding the “Two or more races” category). Here we are adjusting our assumptions about the population rather than the Zeitgeist sample. After our changes we will run the test again with our new data.

**Major** At Middlebury, students often don’t declare their major until the end of their second year—even after they know what they want to study. In Phase 2 we will assume that some students marked down their intended major in the Zeitgeist survey even though they were still “undeclared” according to Middlebury’s records. To account for this, we will remove the “undeclared” category from both data sets—making the assumption that people who gave their intended major on the Zeitgeist survey, while being undeclared, are not distributed in any special way. That is, they follow the distributions of majors already present in both the Middlebury and Zeitgeist data.

# Results

## Phase 1

See the Appendix for the R code.

**Class P-value: 0.17437** When looking at our class category, we observe a p-value of about 0.17437, meaning that we have about a 17% chance of observing something more extreme than the test statistic of our sample given that the proportion of students in each class is randomly sampled from the Middlebury population. This is not rare enough to reject the null hypothesis because the p-value is larger than  $\alpha^*$ .

**Geography P-value: 0.04467** For our geography category, we get a p-value of about 0.04467 meaning there is about a 4% chance of observing something more rare than our sample's test statistic given that Zeitgeist is randomly sampled from the administration's population data. This is still larger than  $\alpha^*$ , so we do not have enough evidence to reject the null hypothesis.

**Gender P-value: 0** Here, we got a p-value of 0. In our 100,000 simulations of random samples from Middlebury's data, we did not find a sample with an MAE as extreme as the Zeitgeist data. Therefore, it is extremely unlikely that the Zeitgeist data was a random sample from Middlebury's population. Zero is less than  $\alpha^*$ , so we will reject our null hypothesis that the Zeitgeist data was randomly sampled from Middlebury's data.

**Race P-value: 1e-05** Again, we got a p-value of nearly zero when testing the race data, so it is extremely unlikely that the Zeitgeist data was a random sample of the population. Since the p-value is less than  $\alpha^*$ , we have enough evidence to reject the null hypothesis.

**Major P-value: 0** When testing the major data, we also got a p-value of zero. By the same logic as before, we reject our null hypothesis.

## Phase 2

After adjusting our assumptions, we got the following results. Since neither the class nor geography tests yielded statistically significant results, we will not test them again. We will, however, retest the other three categories after changing our assumptions.

**Gender P-value: 1e-05** After changing our gender assumptions, we still got a p-value near zero. Now we can be confident that our significant result from Phase 1 was not due to the assumptions we made. This result gives us more evidence that the Zeitgeist is not a random sample from the Middlebury population.

**Race P-value: 4e-05** Again, changing our assumptions still yielded a near zero p-value, showing that our significant result in Phase 1 was not due to our assumptions.

**Major P-value: 0** Lastly, we changed our assumptions about the major data and still received a near zero p-value. Therefore, we still reject our null hypothesis that the Zeitgeist was a random sample from the Middlebury population.

## Conclusion

We wanted to know if the Zeitgeist survey was a random sample from the population of Middlebury students. After looking at the demographic of students who took the survey by class year, gender identity, race/ethnicity, geographic distribution (hometown), and academic major, we found evidence that the Zeitgeist was not a random sample. For each we simulated 100,000 random samples from Middlebury's population and found that the Zeitgeist sample had exceptionally rare distributions by gender, race, and major. This is not surprising because we could imagine that people from certain groups on campus are more likely to take the Zeitgeist survey. It does, however, mean that we should be careful when reading the statistics from Zeitgeist—they may not reveal anything about the Middlebury population as a whole.

## Acknowledgement

We would like to thank The Director of Assessment and Institutional Research at Middlebury, Adela Langrock, who generously compiled Middlebury's geography data for us. Lastly, we would like to thank Professor Alex Lyford who taught us statistics and advised us on this report.