ICPSR | INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH

ICPSR 38588

College and Beyond II (CBII) Course Content Data, [United States], 2000-2021

College and Beyond II Series User Guide

Inter-university Consortium for
Political and Social Research
P.O. Box 1248
Ann Arbor, Michigan 48106
www.icpsr.umich.edu

College and Beyond II (CBII) Course Content Data, [United States], 2000-2021

Annaliese Paulson
University of Michigan

Paul N. Courant
University of Michigan

Allyson Flaster
University of Michigan

Susan Jekielek
University of Michigan

Margaret Levenstein
University of Michigan

Timothy A. McKay
University of Michigan

Kevin M. Stange
University of Michigan

Terms of Use

The terms of use for this study can be found at:
http://www.icpsr.umich.edu/web/ICPSR/studies/38588/terms

Information about Copyrighted Content

Some instruments administered for studies archived with ICPSR may contain in whole or substantially in part contents from copyrighted instruments. Reproductions of the instruments are provided as documentation for the analysis of the data associated with this collection. Restrictions on "fair use" apply to all copyrighted content. More information about the reproduction of copyrighted works by educators and librarians is available from the United States Copyright Office.

College and Beyond II Series User Guide

December 2022

Prepared By

D'Wayne Bell
W. Carson Byrd
Allyson Flaster
Benjamin Koester
Susan Leonard
Kayla Manley
Raphael Nishimura
Anna Paulson
Kevin Stange

Acknowledgements

College and Beyond II is a joint project of researchers from the University of Michigan's College of Literature, Science, and the Arts; Gerald R. Ford School of Public Policy; Inter-university Consortium for Political and Social Research; and School of Education. College and Beyond II was supported by the Mellon Foundation [Grants G-1802-05485 and G-1910-07255] under the auspices of the Mellon Research Forum on the Value of Liberal Arts Education. The project was led by Paul Courant, Kevin Stange, Allyson Flaster, Susan Jekielek, Tim McKay, and Maggie Levenstein.

Many people contributed their time and expertise to make College and Beyond II happen. We wish to thank Mariët Westermann, Armando Bengochea, Heather Kim, Camilla Somers, Mike McPherson, James Shulman, Maya Kobersy, Johanna Bleckman, Patrick Woods, and the College and Beyond II Advisory Committee for their vision, advice, and guidance. Anya Ovchinnikova, Alison Sweet, Alex Toma, Shane Redman, and Elizabeth Meier-Austic kept the project organized and on track. We appreciate Jenni Brady for her ability to elucidate the aims of the project and value of the liberal arts; Mark Umbricht, Krystina Engleman, and Marissa Thompson for their expert data skills; and Maureen O'Brien, Pete Batra, and the data collection staff at the Survey Research Operations for shepherding the survey to completion. Anne Gere and her research team provided extensive feedback, particularly on the Alumni Survey and Open Response studies. A big thank you to the leadership and staff of the College and Beyond II partner institutions for sharing their data and expertise in order to advance research on undergraduate education.

Citation

Please cite the data in publications and presentations. See why here. Below is an example for the Administrative Data.

Courant, Paul N., Flaster, Allyson, Jekielek, Susan, Levenstein, Margaret, McKay, Timothy A., and Stange, Kevin M. College and Beyond II (CBII) Administrative Data, [United States], 2000-2020. Inter-university Consortium for Political and Social Research [distributor], 2022-08-18. https://doi.org/10.3886/ICPSR38488

Table of Contents

1.0 Introduction to College and Beyond II

1.1 Background and Potential Uses

It is well-documented that a college education provides benefits to individuals and society. Exactly *how* college transforms students' lives is less well understood. In particular, little is known about the mechanisms that link the undergraduate educational experience with long-term outcomes. In recognition of this, in 2019 the Mellon Foundation commissioned a team of researchers to collect data from postsecondary institutions and other sources that could be used to answer a wealth of questions about the nature and long-term value of undergraduate education. The resulting study, *College and Beyond II*: *Outcomes of a Liberal Arts Education*, serves as the foundation for the data described in this user guide.

This user guide provides general and study-specific guidance on how to understand and use data from the College and Beyond II series (CBII). CBII contains data on bachelor's-seeking undergraduates at 19 public institutions from approximately 2000-2021. These 19 institutions represent seven different postsecondary systems and are diverse in terms of student population, size, mission, and geography. For a subset of students, CBII contains later-life outcomes data on a variety of domains – including health and well-being, civic and democratic engagement, labor market participation, and openness to diversity – more than a decade after they graduated. CBII data is made available to approved researchers by the Inter-university Consortium on Political and Social Research (ICPSR) at the University of Michigan.

The Mellon Foundation was motivated to support CBII by a specific interest in learning about the impact of a liberal arts education. However, thanks to the comprehensive nature of the data, CBII can be used to examine student experiences outside of the liberal arts as well. The study was designed so that researchers could answer a wide variety of research questions about the undergraduate experience and its consequences. The data is particularly well-suited to answering questions that require access to student-level data on the full population of students at several universities across many cohorts, rather than a sample of students across a large number of institutions. Analysis seeking to compare or benchmark institutions as a whole would not be appropriate.

Preliminary uses of the data have included an examination of the post-graduation effects of high-impact practices; how curriculum relates to later-life career adaptability, civic engagement, and labor market outcomes; measuring curricular breadth and interdisciplinarity using course descriptions and networks; creating new measures of racial diversity; thematic analysis of students' most significant experiences; and many others.

1.2 Series Overview

The CBII series consists of eight studies that can be linked together to answer a wealth of research questions. Each study contains de-identified data on a common topic or from a common data source.

The **Administrative Data** serves as the core of the CBII series. It contains student record and transcript data on bachelor's-seeking undergraduates enrolled at 19 institutions nested within seven different postsecondary systems. The data spans from 2000-2021 and contains records for over 1.3 million students, including both degree completers and non-completers. Graduate students and students that were not seeking a bachelor's degree (i.e., students exclusively enrolled in community colleges) are not included in the data. The records contained in the Administrative Data are used to define the samples used in the other studies in the CBII series.

The **Alumni Survey** measures later-life outcomes and impressions of the college experience approximately a decade after bachelor's receipt for a subsample of students in the Administrative Data. Survey respondents graduated during the 2009-2010 academic year and took the survey in 2021. Data collection coincided with the COVID-19 pandemic. Surveys were sent to a sample of 15,000 CBII institution alumni, and valid responses were received from 2,801, representing a 19% response rate.

The **Alumni Survey Open Response** study includes respondents' written responses to two open-ended questions that were asked in the CBII Alumni Survey. These questions allowed respondents to write freely about their perceptions and experiences. One variable contains respondents' reflections on the college experiences that had an impact on their personal and professional development, and the other contains respondents' final thoughts at the conclusion of the survey. The study only contains data for respondents who answered one or both of the open-ended questions. In total, there are 2,462 observations in the data.

Enrollment and Awards data contains information from three systems about students' periods of enrollment and awards earned (e.g., degrees, certificates) at institutions that report data to the National Student Clearinghouse. It also contains a file of derived variables created by the CBII team that summarize students' enrollment patterns and highest degrees earned. The Enrollment and Awards data is ideal for studying topics related to educational transitions and attainment, such as undergraduate transfer and graduate school enrollment.

The **Course Content** study contains course section-level information on the academic content of student's courses for nearly all sections enrolled in by bachelor's-seeking undergraduate students at CBII institutions between 2000 and 2021. This includes the department the course was offered in, the course College Course Map (CCM) code, and - at four partner systems - the course catalog description and full title associated with that course.

Student Advanced Placement (AP) data contains information on student-level Advanced Placement test results and course credit provided by four CBII systems. The data includes test scores, test names, and course credit received. AP test results are often used as measures of students' academic preparation and the availability of pre-college learning resources. They also factor into the accounting of credits earned and credits taken by students during college, as they are often accepted in college as credit toward an academic degree. Student AP data can be linked to other student-level data in the series.

Student Experience Analytics are summary measures of the student experience created by researchers and derived solely from information extracted from the Administrative Data. The study consists of two sets of derived measures: Course Diversity and Transcript of the Future. Both make use of information for the full universe of undergraduate students at CBII institutions. They are provided as examples of the types of aggregate measures that can be distilled from transcript data. Users are encouraged to explore their potential for measuring aspects of students' undergraduate experience and to develop their own measures in a similar vein.

The **Contextual Data** provides two datasets that can be used to contextualize other data in the CBII series: the Integrated Postsecondary Education Data System (IPEDS) and the National Neighborhood Data Archive (NaNDA). The IPEDS data cover 2004-2021 and the NaNDA data cover 2008-2017. These are well-documented, publicly available data provided within the virtual data enclave (VDE) for researchers' convenience.

Variable metadata for all the studies can be seen here. Most studies' files are organized by postsecondary system. **Table 1.1** shows which systems are represented in each of the studies that contain institutional data.

Table 1.1: Availability of Study Data by CBII System

Study	CBII System						
	A	B	C	D	E	F	G
Administrative Data	■	■	■	■	■	■	■
Alumni Survey	■	■	■	■	■	■	■
Alumni Survey Open Response	■	■	■	■	■	■	■
Enrollment & Awards		■				■	■
Course Content	■	■	■	■	■	■	■
Student Advanced Placement	■			■		■	■
Student Experience Analytics	■	■	■	■	■	■	■

■ = All files available ■ = Some files available ■ = No files available

Note:
Administrative: System G does not have a Term file available.
Alumni Survey: Graduates from one institution at System E were sampled for the survey.
Course Content: Only Systems A, B, C, & D have course descriptions available (description & description_raw).
Contextual Data: Is not organized by system and is excluded from this table.

1.3 Linking Data Across the Series

CBII data files contain anonymized identifiers for important entities represented in the study. These entities are students (**id_person**), postsecondary systems (**id_system**), institutions (**id_college**), and courses (**id_course**). Data files can be linked together within and across studies in the series using these identifiers, either alone or in combination with other variables to create a unique key. **Table 1.2** lists all the identifiers and codes in the CBII series that can facilitate linkage.

Some data files – such as the Term and Course files in the core Administrative Data – are panel data, with multiple observations for each student. For such files, unique keys impose the constraint that there are no duplicate rows for a given student. **Table A1** in the Appendix lists the unique keys for each data file in the studies.

Note that sometimes identifiers have slightly different variable names across data files. For instance, in the Administrative Data's Student file, **id_college** has several iterations, such as **id_college_start** (id_college for the first CBII institution the student attended), **id_college_start_bach** (id_college for the first CBII institution attended in which the student attempted a bachelor's degree), and **id_college_degree** (id_college for the first CBII institution from which the student received a bachelor's degree)**.** Despite the different variable names, all forms of **id_college** refer to the same set of anonymized, four-digit institutional IDs and thus can be used to link records associated with attendance at specific colleges within and across studies in the series.

Some studies also contain codes that can be used to link records to additional information. Examples include standardized term codes created by the CBII study team and ZIP codes from the US Postal Service. Much like with **id_college**, these codes may have slightly different variable names from file to file, but they refer to the same entities.

1.4 Variable Naming and Labeling Conventions

Variable names, variable labels, and value labels were designed to be as concise as possible while also providing sufficient description and detail for data users. Generally, names and labels follow these guidelines:

- All variable names are lowercase and start with an alpha character.[1] Variable names do not exceed 28 characters.
- Underscores are used to separate parts of a variable name. The first part of the variable name is the primary construct (e.g., act_). The following parts of the variable name contain sub-constructs and keywords (e.g., act_math_) and, sometimes, a description of the variable type (e.g., act_math_derived).

[1] Exceptions include the variable names in IPEDS and NaNDA, which were not created by CBII.

Table 1.2: Variables That Facilitate Linkage Across CBII Studies

Identifier / Code	Description	Available In	Variable Names
CBII Person ID	Anonymized student identifier.	Administrative Data Alumni Survey Alumni Survey Open Response Enrollment and Awards Student Advanced Placement Student Experience Analytics	id_person
CBII System ID	Anonymized postsecondary system identifier.	Administrative Data Student Advanced Placement Student Experience Analytics Course Content	id_system
CBII College & University ID	Anonymized institutional identifier.	Administrative Data Alumni Survey Alumni Survey Open Response Enrollment and Awards	id_college id_college_start id_college_start_bach id_college_degree
CBII Course ID	Course identifier.	Administrative Data Course Content Student Experience Analytics	id_course
CBII Term Code	Standardized, chronological term identifier.	Administrative Data Alumni Survey	term_code entry_term_code entry_term_bach_code degree_term_code_first degree_term_code_second
CIP code	Classification of Instructional Programs code developed by the US Department of Education.	Administrative Data Alumni Survey Course Content Contextual Data	major_cipcode_01 major_cipcode_02 major_cipcode_03 major_cipcode_term_01 major_cipcode_term_02 major_cipcode_term_03 department_cipcode CIPCODE CIPCODE1
Unit ID	Unique identifier assigned to entities that report to IPEDS by the US Department of Education.	Alumni Survey Enrollment and Awards Contextual Data	col_first_lookup_00_ID col_serious_lookup_01_ID col_serious_lookup_02_ID col_serious_lookup_03_ID unitid UNITID
OPE ID	Unique identifier assigned to institutions by the Office of Postsecondary Education.	Enrollment and Awards Contextual Data	nsc_college_code OPEID
ZIP code	US Postal Service code for geographic area.	Administrative Data Alumni Survey Contextual Data	address_zip_raw address_zip_derived zip_current zipcode ZIP

- Words were spelled out in variable names whenever possible, especially for the primary construct. However, abbreviations are also used to shorten variable names where needed.
- Acronyms are used only when most users are likely to understand what the acronym stands for (e.g., 'ACT' – a common standardized test used in college admissions).
- Anonymized study identification variables begin with id_.
- Variable labels, available in selected files, do not exceed 80 characters.
- Value labels, available in selected files, include both the data value as well as the description of the value (e.g., "0 Freshman" and "1 Transfer").

Naming and labeling conventions specific to a study are explained in their respective sections, when applicable.

1.5 Accessing the Data

Colleges and universities provided data to CBII under the condition that individuals' and institutions' privacy and confidentiality be protected. Thus, all CBII data are restricted-use and available only for approved researchers in the ICPSR Virtual Data Enclave (VDE).

To request data access, researchers begin by completing an online application that can be accessed from the CBII ICPSR webpage. The application requests information about the lead researcher (investigator) and the topic of the proposed study. It also requires a Restricted Data Use Agreement signed by both the investigator and a legal representative of the investigator's institution. Once approval to access the data has been granted, investigators must purchase annual VDE user licenses for each member of the study team who will access the data. As of December 2022, the cost of an annual user license is $484.

The VDE provides access to restricted-use data through a virtual machine launched from the researcher's own desktop but operating on a remote server, similar to remotely logging into another physical computer. Users analyze restricted-use data using several software options available within the VDE. All research output must undergo disclosure risk review by ICPSR staff to ensure that the output cannot be used to identify an individual or institution. Depending on the volume of requests, a review can take up to 10 business days. Guidance on how to submit a disclosure review request is available in the VDE.

1.6 File Types and Documentation in the VDE

Users of the CBII data will find that the types of files, metadata, and documentation available in the VDE vary from study to study (see **Table 1.3**). The Administrative Data and Alumni Survey, which are core aspects of the CBII study, have been curated and documented at a higher level of detail than the other studies in the series. The Alumni Survey and Administrative Data files are provided in commonly used statistical package formats, ready for use. Additional studies in

the series are provided in comma-separated-values (CSV) format for users to read into the statistical software of their choice. Some studies also have Excel or Stata file formats available.

This User Guide serves as the main source of information about the CBII study series and will be available for data users to consult in the VDE. Data users will also have access to the CBII **Technical Appendix**, which is only available in the VDE and contains detailed information about the construction and composition of variables. Specifically, for each variable the Technical Appendix contains:

1. The variable name and variable description.
2. Notes on the construction of the variable across systems.
3. Notes on the construction of the variable at a specific CBII system.

The Technical Appendix also documents minor errors that have been discovered after the data was produced.

Several of the additional studies have **Variable Coverage Tables** available. Variable Coverage Tables are spreadsheets that indicate the percentage of non-missingness for each variable in the data file, by system and students' entry term. The spreadsheets are color coded so that users can see, at a glance, how sparse the coverage may be for a given variable and cohort. **Table 1.3**, below, shows all the types of files and documentation available in the VDE.

Table 1.3: Files and Documentation in the VDE, by Study

Study	Type of Files Available in the VDE	Documentation Available in the VDE
Core Data		
Administrative Data	SAS, SPSS, Stata, R, ASCII	Codebook, Technical Appendix
Alumni Survey	SAS, SPSS, Stata, R, ASCII	Codebook
Additional Studies		
Alumni Survey Open Response	CSV	Read Me
Enrollment & Awards	CSV, Stata, XLSX	Read Me, Technical Appendix, Variable Coverage Table, CIP Code table, Key to Values in Raw NSC Data
Course Content	CSV, Stata, XLSX	Read Me, Technical Appendix, Variable Coverage Table, CIP Code and CCM Code tables, CCM Technical Report
Student Advanced Placement	CSV, Stata	Read Me, Technical Appendix, Variable Coverage Table
Student Experience Analytics	CSV	Read Me, Technical Appendix
Contextual Data	CSV, Stata	Read Me, IPEDS and NaNDA documentation

1.6.1 Working with Large Files

Given the large size of the Administrative Data, users who are linking or appending files are advised to carefully consider which variables and observations to retain. Large files can slow data processing and even result in the incomplete execution of analytical tasks. When using large data files, particularly in the Administrative Data, we recommend:

- Only read in and keep variables you need.
- Convert string variables to numeric.
- Avoid merges that retain all variables from multiple files.
- Regularly delete old or unused versions of large intermediate data files and instead rely on archived code that creates such data files.

When working with data at the scale of the Administrative Data's Term and Course files, typical workflows one might use with smaller datasets can be inefficient or impossible due to time or memory constraints. We advise that users consider how to optimize memory usage and efficient application of operations. Recognizing that optimizing code may be unfamiliar to many users, here we provide some guiding principles.

To optimize code:

1. Experiment with code on subsets of the data and estimate total run times before attempting to run code on full datasets.
2. Instead of beginning with full CBII files for analysis, consider writing-out minimal working versions of the data to disk that are optimized to work within the scope of the research questions.

To optimize memory:

1. Keep only variables that you plan on using in your analysis. This is especially important when merging files across levels of granularity. Attempting to merge all variables from the Student file with all variables from the Course file will consume an unreasonable amount of memory and may cause the VDE machines to crash.
2. Convert strings to numeric or categorical values if possible
3. Store data in appropriate data formats. For instance, when working with text data derived from course descriptions or networks derived from student transcripts, consider storing data in sparse matrices rather than memory-intensive dense matrices.
4. If necessary, break tasks into intermediate parts and save the results. For instance, if performing a memory-intensive computation on the Student file, one might work on each

term of the file, saving files for each term. Note that this will sacrifice computational efficiency for the sake of memory usage by requiring many IO (input/output) operations.

5. Make use of built-in functionality. For instance, the “gc()” command in R clears unused memory, and the memory usage widget in R Studio can help to understand memory limits.

To optimize computational efficiency:

1. When possible, use vectorized operations and avoid using loops. If loops are necessary, include only the minimum number of functions within the loop and examine how much time each function call within the loop requires. Functions that achieve similar outputs may vary in terms of computational efficiency and you need to find efficient versions of functions. For instance, base R’s merge function may be less computationally efficient than dplyr’s join function to achieve the same end.

2. If merging files across multiple levels of aggregation, perform as many operations as possible on individual files before merging. For instance, if working with variables from the Student and Course files, derive all student-level variables before merging the files. Similarly, if intending to work with aggregated measures from the Term or Course files merged with the Student file, first perform these aggregations before merging the files.

2.0 Administrative Data

2.1 Study Description

2.1.1 Overview

The CBII Administrative Data contains administrative student records for nearly all bachelor’s-seeking undergraduate students at 19 public colleges and universities nested within seven postsecondary systems from 2000 to 2021. We refer to these 19 institutions as “CBII institutions.” The names of CBII systems and institutions are masked in the data and in this user guide to maintain the confidentiality of the institutions and to preclude benchmarking of colleges and universities against each other.

The CBII institutions provided a wealth of information about students from their administrative information systems, including student demographics; family background; entry and completion terms; college major; admissions test scores; term-by-term information on majors, credits, and grades; and information on all courses taken. Students’ administrative data can be linked to other student-level CBII data using the anonymized student identifier, **id_person**.

2.1.2 Study Aims and Content

The study aimed to collect detailed quantitative longitudinal information about undergraduate students and their experiences in several domains:

- Demographics and family background: sex, race/ethnicity, parent education, family income, citizenship, permanent address zip code, high school state, high school code, residency
- Academic background and preparation: ACT/SAT score, high school grade point average (GPA), transfer student status, transfer credits
- Extracurricular and co-curricular activities: Greek participation, athletics participation, residential college, study abroad
- Final academic outcomes: degree date, major field(s) of study, cumulative GPA, cumulative credits, honors
- Term-level outcomes: credits attempted, credits earned, term GPA, cumulative GPA, declared major, college of enrollment
- Courses taken and outcomes: subject, number, title, credits, letter grade, course type, meeting day/time

Such data can be used to investigate questions about students' postsecondary paths and success. Demographic and background information can be used to examine inequalities in access to different types of collegiate experiences and differences in student success by background, and also serve as controls for pre-college factors that may influence outcomes. Term-level outcomes can be used to examine students' paths through college, including the effect of time-varying factors, such as academic performance, on persistence and degree completion. Course-level data can be used to richly characterize the curricular experiences of students longitudinally, including investigating the role of specific courses (e.g., "gatekeeping" courses such as college algebra) in student progress, or to examine the diversity of courses and subjects taken, even by students graduating with the same majors. When combined with other data in the CBII series, the core Administrative Data can be used to examine the long-term consequences of undergraduate experiences.

2.2 Study Design

2.2.1 Target Population

The target population is bachelor's-seeking undergraduates who entered CBII institutions since 2000, including students who did not complete a degree and students that previously enrolled in another institution prior to a CBII institution. Graduate students and students that were not

seeking a bachelor's degree (i.e., students exclusively enrolled in community colleges) are not included in our target population.

2.3 Data Collection

We asked institutions to provide as many of the fields in the domains described above as possible arranged in tables: student-level data, term-level data, and course-level data. Data collection took place between 2019 and 2021, with raw flat text files transferred to the CBII team at various times during this period.

Not all variables are available for all institutions, particularly extracurricular activities. We included variables in the study that are only available from a few or even a single institution if we viewed these variables as particularly useful for the aims of CBII.

2.3.1 Sample Restrictions and Sizes

We restricted our sample to students that appear in all three core data files provided by institutions: Student, Term, and Course. Since a Term file is not available for System G, students are only required to appear in the Student and Course files to be included from that system. The sample is restricted to students with a Fall 2000 start date or later (Fall 2003 start date for System D and System G).[2] Since System E includes community colleges, the System E sample is restricted to students who ever pursued a bachelor's degree or attended one of the bachelor's-granting institutions.

Because sample inclusion is conditional on start date, the data does not include *all* enrolled undergraduates at these institutions in each academic year until several years into the study period. That is, Term and Course data for Fall 2000 will only contain records for new entrants in that year; Fall 2010 will contain records for nearly all students, including Fall 2010 entrants and students who entered in the prior decade. Thus, analysts wishing to use information about all students enrolled concurrently with students who are the focus of their analysis should examine terms that occur four or five years into the CBII study period.

Table 2.1 shows the number of unique students and observations for each data file by system. The final sample includes records for 1,311,818 unique students, 9,171,187 unique student-terms, and 46,863,737 unique student-course enrollments.

[2] Approximately 40,000 students (3% of the sample) were dropped due to a missing or invalid start term. Most of these students were missing other information or interpreted to be possibly out of the study scope (e.g., graduate students).

Table 2.1: Core Administrative Data File Counts

System	Student File	Term File	Course File
	Unique observations (student)	Unique observations (student-term)	Unique observations (student-course)
A	162,414	1,174,484	7,746,957
B	37,774	240,067	1,046,518
C	196,737	1,169,781	4,466,554
D	25,670	180,051	876,829
E	594,891	5,370,977	20,241,408
F	163,729	1,035,827	5,846,516
G	130,603	–	6,638,955
Total	1,311,818	9,171,187	46,863,737

2.3.2 Institutional Coverage

Our student-level samples capture most enrollment reported by CBII institutions to the Integrated Postsecondary Education Data System (IPEDS). **Table A2** in the Appendix compares the number of new unique students enrolling in each Fall entering cohort in CBII to those reported in IPEDS. Note that IPEDS counts from 2000 to 2005 do not include transfer students, so IPEDS will undercount the number of new students in those years. The table also reports the "coverage rate," which is the ratio between these two numbers. Generally, the CBII coverage rate is high, exceeding 90% for most entering cohorts and systems. Typically, CBII includes fewer students than reported to IPEDS, which is unsurprising given our restriction that we must have received valid student, term, and course records in order for a student to appear in our final dataset. CBII entrants in Fall 2006 exceed those in IPEDS at three institutions; this appears to be a transitory disconnect due to when a period of enrollment growth is reported to IPEDS.

System E is an exception, with about 70% coverage throughout our sample period. This incomplete coverage could be due to a number of factors that we continue to investigate. For example, it is possible that IPEDS counts include a broader set of individuals than we include in the CBII sample or that IPEDS responses at an institution level double-count individuals who enroll at multiple institutions in a system. System E also has many students that enrolled in non-Fall terms (which are excluded from the comparison table), which may be treated differently in the CBII administrative records and in IPEDS. Nonetheless, the CBII sample includes the majority of bachelor's-seeking new enrollees at all CBII institutions, including those in System E.

2.3.3 Final Quality Checks, Variable Standardization, and Redaction

The administrative data we received directly from institutions needed extensive checking, documenting, cleaning, and standardization to make it into a cohesive dataset that is usable for researchers. This work occurred over several years and is beyond the scope of this document

to fully describe. Once we had a set of preliminary files, we performed a number of final quality and consistency checks, described below.

File-level quality checks included checking that individuals were included in all three files (two files for System G) and that records that preceded the study period were excluded. We performed some basic consistency checks across files, including confirming that the entering cohort term (Student file) generally aligns with the first term a student appears in the Term and Course files, confirming that the degree term (Student file) aligns with the final term a student appears in the Term and Course files, and confirming that bachelor's degree recipients generally graduated with about 120 credits. While the data generally passes these checks for most students, we did not require that the data be internally consistent across the different data tables obtained directly from institutions. Users may wish to impose their own internal consistency requirements on the data depending on their use.

Many categorical variables, for example, race/ethnicity and parent education have inconsistent categories across institutions. Thus, we created derived variables with categories that are as consistent as possible across institutions. In many cases, we also include "raw" versions of the variables so users can impose their own standards and coding systems on the variables.

All string variables were run through a redaction code to remove any text that specifically identifies the CBII institution or system, including institution names, abbreviations, nicknames, mascots, and city/location. This text was replaced with generic language and included in brackets; for example, "University of Michigan" was replaced with "[INSTITUTION]." The final redacted files were inspected manually before being shared with ICPSR for distribution to researchers.

2.4 Administrative Data Files

2.4.1 Overview

The Administrative Data is arranged into three separate files for each of the seven postsecondary systems for a total of 20 data files (a term file is not available for System G). The same file type can be stacked ("appended") across systems, as all variable names, storage types, and labels are consistent across systems. However, given the large file sizes, users should refrain from stacking files until necessary. Consult the guidelines in Section 1.6.1 about working with large files in the VDE environment.

2.4.2 Student Data File

The Student data file contains one record for every student included in the administrative data collection. Each student is uniquely identified by **id_person**. There are 1,311,818 observations. The Student data file contains information that generally does not change over the course of students' records. These include information in five domains:

- Study information: student identifier, CBII system, first CBII institution, entry term, participation in CBII Alumni Survey
- Demographics and family background: sex, race/ethnicity, parent education, family income, citizenship, permanent address zip code, high school state, high school code, citizenship
- Academic background and preparation: ACT/SAT score, high school GPA, transfer student status, transfer credits
- Extracurricular and co-curricular activities: Greek participation, athletics participation, honors program
- Final academic outcomes: degree receipt, degree award term, major(s) field of study, cumulative GPA, cumulative credits

Students' **entry_term** was obtained directly from institutions. In cases where it was missing, we filled it in as the first-term students appeared in the Term file. Given the variety of academic calendars followed by CBII institutions, variables denoting terms, including entry term and degree term, were recoded into a standard scheme designed by the CBII team, as described in Section 2.4.7.1.

For ease of use, the Student file contains derived variables. Final cumulative GPA (**gpa_cum_final**) is the cumulative GPA taken from the last term the student appeared in the Term data. For System G, for which a Term file is not available, final cumulative GPA is computed from data in the Course file. There are also indicators for whether the student earned a bachelor's degree within four (**grad_4years**), six (**grad_6years**), and eight years (**grad_8years**) of first entering the CBII institution. These variables are constructed from **degree_term_code_first** and **entry_term_code** and are missing if a full four (six, eight) years had not passed since **entry_term** when institutions provided the data. Therefore, these variables will correctly measure graduation rates consistently across cohorts and institutions.[3]

The Student file provides up to three major fields of study for each student. Major field of study for students' bachelor's degrees was provided by institutions as an unstructured string and, for some institutions, as a six-digit Classification of Instructional Programs (CIP) code. The Student file contains the raw major degree title (**major_descr_01**) as provided by institutions, with limited edits for redaction. Most analysts will want to use instead the standardized six-digit CIP major codes and descriptions (**major_cipcode_01** and **major_cipdescr_01**) which are comparable across institutions. The process used to standardize majors into six-digit CIP codes is described in Section 2.4.7.2.

Additional details on how each variable in the Student file was constructed are contained in the Technical Appendix. **Table 2.2** lists the variables that are missing in the Student file for entire systems.

[3] See Technical Appendix's 'Data Errors' tab for exceptions

Table 2.2: Variables Missing for Entire Systems, Student File

Variable	System						
	A	B	C	D	E	F	G
degree_term_second degree_term_code_second		X	X				
major_descr_02 major_cipdescr_02 major_cipcode_02			X				
major_descr_03 major_cipcode_03 major_cipdescr_03	X	X	X		X		
transfer_credits_course					X		
transfer_credits_notap		X			X		X
transfer_credits_other		X	X	X	X		X
act_math_derived act_engl_derived sat_verb_derived sat_math_derived sat_comp_derived					X	X	
gpa_hs_raw gpa_hs_derived					X		
greek_life	X	X	X		X	X	X
honors_program		X			X	X	
athlete		X	X		X	X	X
parent_educ_derived				X	X		

2.4.3 Term Data File

The Term file contains one record for each term students in the Student file enrolled at CBII institutions. For System E, we only include terms in which the student was seeking a bachelor's degree, though these terms will include enrollment at community colleges in the system. Each observation is uniquely identified by the combination of **id_person**, **id_college**, and **term_code**. Terms may include Fall, Winter, Spring, and Summer. A Term file is not available for System G. See Section 2.4.7.1 for details on how terms were standardized.

The Term file includes measures in five broad domains. Most measures are time-varying, and thus their values can change across terms:

- Study information: student identifier, term, CBII system, CBII institution
- Student and enrollment characteristics: residency, full/part time status, school/college
- Extracurricular and co-curricular activities: honors program participation, residential college participation, Greek life participation, study abroad

- Current term academic outcomes: credits attempted and earned, GPA, current major(s)
- Cumulative academic outcomes: cumulative credits attempted and earned, cumulative GPA

The current credit and cumulative credit information contained in the Term file has been modified very little from the raw files obtained from institutions. As such, we have not imposed internal consistency between these variables or between credits measured in the Term and Course files. We thus include **credits_attempted_derived** and **credits_earned_derived** in the data, which compute the number of attempted and derived credits in a given term from the course records contained in the Course files.

Current major field of study on a term-by-term basis was provided by institutions as an unstructured string variable and, for some institutions, as a six-digit CIP code. The Term file includes up to three major fields of study per student. The Term file contains this raw major descriptor as provided by institutions (**major_descr_01)**, with limited edits for redaction. Most analysts will want to use instead the standardized six-digit CIP major codes and descriptions (**major_cipcode_term_01** and **major_cipdescr_term_01**) which are comparable across institutions. For System E, some Term file CIP codes (e.g., 99.9999) have differing raw descriptors; thus, we did not label these with a CIP code descriptor. The process we used to standardize major descriptors into six-digit CIP codes is described in Section 2.4.7.2. Additional details on how each variable in the Term file is constructed are contained in the Technical Appendix available in the VDE. **Table 2.3** below lists the variables that are missing from the Term data for entire systems.

Table 2.3: Variables Missing for Entire Systems, Term Data

Variable	System					
	A	B	C	D	E	F
school_code_first school_descr_first		X		X	X	X
school_code_second school_descr_second		X	X	X	X	X
residency_code		X				
major_descr_term_02 major_cipcode_term_02 major_cipdescr_term_02			X		X	
major_descr_term_03 major_cipcode_term_03 major_cipdescr_term_03	X	X	X		X	
honors_program_term		X	X	X	X	X
residential_college		X	X	X	X	X
study_abroad	X	X	X		X	X
greek_life_term	X	X	X		X	X

Note: A Term file is not available for System G.

2.4.4 Course Data File

The Course data file contains one record for every course section each student in the Student file took at the CBII institutions. Each observation is uniquely identified by the combination of **id_person**, **id_college, id_course, course_code_comp,** and **course_grade_basis**. For System E, only courses taken during terms when the student was seeking a bachelor's degree are included.

The Course file includes measures in three broad domains:

- Study information: student identifier, term, CBII system, CBII institution, course identifier
- Course identity: course title, subject number, grading basis, day, time, and location
- Course performance: course grade (letter and numeric), credits attempted and earned

The variable **id_course** is a concatenation of course subject (**course_catalog_subject**), course number (**course_catalog_number**), term (**term_code**), and course section number (**course_section_code**). This variable can be used to identify students that took the same section of the same course in the same term. A course with multiple components (e.g., a lecture and a lab) will have multiple entries. These will receive a distinct **id_course** if the institutions provide a different **course_section_code** for each course component. A course that has multiple grading bases will also have multiple records, though this is not common. Typically, only one course component will have credits and a course grade attached. Users should keep only one course component when constructing measures of total courses taken.

Note that, because the raw values for **course_catalog_subject** can contain text that could identify institutions within our data, the CBII team redacted potentially identifying text within the course subjects, including institution names, abbreviations, nicknames, mascots, and cities/locations. Information that was redacted was replaced with a descriptor enclosed in brackets and capitalized, for example [INSTITUTION]. Thus, **id_course** can contain redactions as well.

Academic performance can be measured by course grades and credits attempted and earned. We obtained a variable denoting the units associated with a course from institutions, though there was variability across institutions in whether this variable represented credits attempted or earned. We used **course_units_raw** along with **course_grade_letter** and **course_grade_number** to construct new derived measures of course credits attempted and earned that are standardized across institutions. **Table 2.4** describes the construction of these two derived course unit variables.

Table 2.4: Construction of Derived Course Credit Variables

System	course_units_attempted_derived	course_units_earned_derived
A	= course_units_raw	= course_units_raw = 0 if (course_grade_number = 0 or missing)
B	= course_units_raw	= course_units_raw = 0 if (course_grade_number = 0 or missing) = course_units_raw if (course_grade_letter = “S”)
C	= course_units_raw	= course_units_raw = 0 if (course_grade_number = 0 or missing) = course_units_raw if (course_grade_letter = “S”)
D	= course_units_raw = modal credits if (course_grade_letter = “W”)	= course_units_raw = 0 if (course_grade_number = 0 or missing) = course_units_raw if (course_grade_letter = “P”)
E	= course_units_raw	= course_units_raw
F	= course_units_raw = modal credits if (course_grade_letter = “W,” “F,” or “S”)	= course_units_raw = 0 if (course_grade_number = 0 or missing) = course_units_raw if (course_grade_letter = “P”)
G	= 3	= 3 = 0 if (course_grade_number = 0 or missing) = 3 if (course_grade_letter = “CR”)

The derived measures have *attempted* credits greater than zero—even if a student fails or withdraws from a course—but are set to zero *earned* credits in these cases. Generally, the course units provided by Systems A, B, C, D, and F correspond to attempted credits and those for System E correspond to earned credits, with a few exceptions. In cases where attempted credits are not provided, we impute it with the modal number of credits earned in each course-term. That is, if most students earned three credits in the class, then we assign three credits attempted for anyone that has zero earned credits. Course credits were not provided by System G, so we assigned three credits attempted for all course enrollments in that system. Earned credits are set to zero for any students that fail or withdraw from a course at System G.

The Course file includes limited information about the content of the courses taken. Most useful are the subject code (e.g., “PSYCH”) and course title (e.g., “INTRO TO PSYCHOLOGY.”). The Course Content study contains several additional variables that enrich the information in the Course file. See Section 6.4 for more details.

Additional details on how each variable in the Course file is constructed are contained in the Technical Appendix in the VDE. **Table 2.5** lists the variables that are missing from the Course file for entire systems.

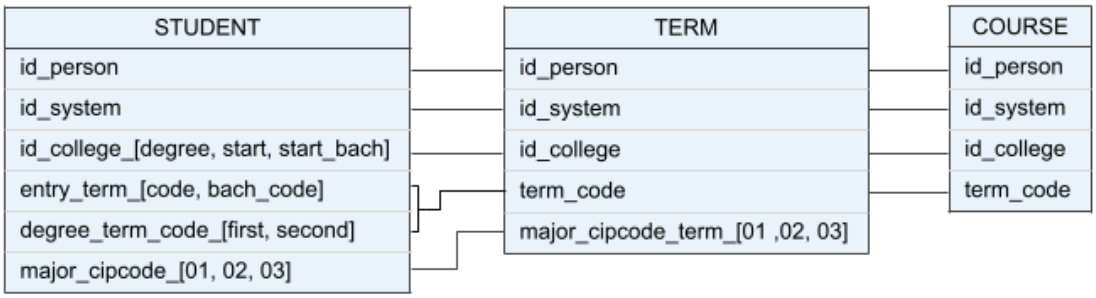
Table 2.5: Variables Missing for Entire Systems, Course Data

Variable	System						
	A	B	C	D	E	F	G
course_code_comp_raw					X		
course_facility_code					X	X	
course_meeting_day					X	X	X
course_meeting_time					X	X	
course_units_raw							X
course_grade_basis							X

2.4.5 Linking Records

Each important entity in the CBII series has a unique identifier. The anonymized student identifier **id_person** is found in the Student, Term, and Course files. The files were constructed so that every **id_person** included in a given system's Student file is also included in its Term file (if available) and Course file. In addition to identifiers, records can be linked across files by term codes and CIP codes. **Figure 2.1** below shows the linkages that can be made across the Student, Term, and Course files with identifiers and codes.

Figure 2.1: Linkage Between Administrative Data Files



The Student file also contains a variable that links to information about individuals an colleges in the CBII Contactual Data study. Respondent's permanent ZIP Code, **address_zip_derived**, will link to ZIP Code Tabulation Areas in the National Neighborhood Data Archive (NaNDA) data (see Section 9.3.4.5). See Section 1.3 for more information about linking data across files in the series.

2.4.6 Missing Values

Missing values are set to -999 in all administrative data files. Values can be missing for a number of reasons:

- “Logical” missing. For example, degree major code in the Student file will be missing for students that did complete a degree.
- Fields were not provided by the institution for any students. These situations are described in the tables above.
- A field was provided by the institution for some students, but not for this particular student or record.
- A record was assigned a known missing code in the source data obtained from institutions. This occurred with some demographic information, such as race/ethnicity.

Before attempting to impute missing values, users should consider whether the value is missing for logical or other reasons.

2.4.7 Code Details

2.4.7.1 Term Codes

We created standard term codes to facilitate the ordering and sequencing of terms within an institution across the different administrative files. Just as importantly, naming and numbering conventions were necessary to standardize term names across institutions. Creating the codes required accommodation of variation across institutions and over time in how terms are used.

The standard term codes we constructed are a two- or three-digit integer that identifies both the academic year and the within-year term. Beginning in Fall of 1996, terms were assigned an integer code, starting with 11, and we defined four terms each year thereafter: Fall, Winter, Spring, and Summer. Codes increment one digit each term and then jump to the next tens digit in a new academic year.

The administrative data files available at ICPSR begin in Fall 2000, so the earliest entry term code appearing in the data is 51. **Table 2.6** provides three years of standardized terms, as well as their corresponding term codes, to illustrate the term coding conventions used in the data. Additional details about the academic calendars at CBII systems can be found in Section 11.1.1 of the Appendix.

2.4.7.2 Major Codes

We standardized all text descriptions of majors (**major_descr**) into the appropriate six-digit CIP codes (**major_cipcode**) and corresponding CIP label descriptions (**major_cipdescr**). Major

codes are available in the Student and Term files. See Section 11.1.2 in the Appendix for more detail about assigning and standardizing major codes.

Table 2.6: Term Coding Convention Examples

Standardized Term Name	Standardized Term Code
Fall 2000	51
Winter 2001	52
Spring 2001	53
Summer 2001	54
Fall 2001	61
Winter 2002	62
Spring 2002	63
Summer 2002	64
…	…
Fall 2004	91
Winter 2005	92
Spring 2005	93
Summer 2005	94
…	…

2.4.8 Variable Naming

The general naming conventions discussed in Section 1.4 of this user guide are applicable to the Administrative Data. Additional specific conventions are as follows:

- Some variable names have an additional part that indicates the timing at which the variable was measured to distinguish it from other measures of the same construct. For example, **greek_life_term** in the Term file and **greek_life** in the Student file measure the same construct, but one at the term level to indicate if the student participated in Greek life during that term and one at the student level to indicate if the student ever participated in Greek life across all terms.

- Variables that end in **_raw** or **_derived** are part of a constructed variable series. The **_raw** variable represents the variable exactly as we received it from the institution, and the **_derived** variable is our standardization of that variable.

3.0 Alumni Survey

3.1 Study Description

3.1.1 Overview

The CBII Alumni Survey was administered to individuals living in the United States who earned a bachelor's degree from one of seven CBII institutions during the 2009-2010 academic year. Most respondents were in their early- to mid-thirties at the time of data collection. Data collection occurred via a self-administered online survey during a period of six months in 2021, which coincided with the COVID-19 pandemic. Surveys were sent to a sample of 15,000 alumni, and valid responses were received from 2,801, representing a 19% response rate. Respondents were asked about their demographic backgrounds, life course events, college experiences, and post-college life outcomes. Respondents' survey data can be linked to other data in the CBII series, including the core Administrative Data.

3.1.2 Questionnaire Content and Development

Questionnaire development began in Summer 2019 with the convening of a working group of experts to advise on questionnaire content and item selection. This was followed in late 2019 with a pilot survey of randomly selected bachelor's recipients who graduated in 2003-2004 from a single institution. The results of this pilot and cognitive interviews were used to hone the final questionnaire content further.

The final Alumni Survey questionnaire contained **seven sections** and **335** items. Previously validated scales were used to measure constructs whenever possible. Two items were open-response questions that allowed respondents to write freely about their perceptions and experiences. The answers to these questions are available in the CBII Alumni Survey Open Response study. The other six sections are included in the survey data file. **Table 3.1** shows the types of information collected in each section that is included in the main survey data file.

3.2 Study Design

3.2.1 Target Population

The Alumni Survey target population are domestic bachelor's degree recipients who earned their degrees in academic year 2009-2010 from CBII institutions. Institutions were asked to provide student records and their most recent contact information for all individuals who met these criteria.

Table 3.1: Section Descriptions

Section 1: College Experiences Participation in extracurricular activities Impact of extracurricular activities Classroom experiences and connections Contact with faculty and staff Sense of belonging in college Challenges encountered in college Living situation in college	**Section 4: Employment & Wealth** Labor market outcomes, such as employment Job duties, benefits, and overall satisfaction Home ownership Student loans Career strengths and adaptability
Section 2: Arts & Culture Openness to diversity Pluralistic orientation Engagement in and appreciation of the arts Frequency of generative and prosocial behavior	**Section 5: Civic & Democratic Participation** Voting information Frequency of civic engagement activities Opinions on democratic issues
Section 3: Health & Well-Being Sense of continued development Perception of relationships with others Attitudes about self Perception of independence and social pressures Rating of physical and mental health	**Section 6: Background** Demographic information Perception of discrimination Household composition Social class growing up Education level of parent(s) and self High school grades Other colleges applied to Religious frequency Political orientation

3.2.2 Sample Design

With the assistance of the Survey Research Operations (SRO) unit at the University of Michigan, we used the target population data to construct a sampling frame using several steps.

First, we selected a single institution from each system to sample from, resulting in seven eligible institutions: 4002, 4006, 4008, 4013, 4016, 4029, and 4030. The sampling frame contained **24,529** members of the target population across these seven institutions. Next, we compiled the most recent directory information each institution had on file for the 24,529 individuals. This information included mailing address, email address, and phone number.

Table 3.2 shows the percentage of those from the sampling frame who were missing specific components of the directory contact information before making an attempt to update.

Next, SRO submitted the directory contact information to a database of public contact information maintained by Accurint for checking and updating against their records. This provided SRO with updated contact information for all but 422 individuals. Following contact information updating, only three members of the sampling frame lacked contact information of any kind. These observations were ineligible for sampling due to the impossibility of sending them an invitation to take the survey.

Table 3.2: Sampling Frame Missing Contact Information Before Updating, by Institution

Contact Information	Institution						
	4002	4006	4008	4013	4016	4029	4030
Total frame members	5,905	5,497	5,619	1,076	4,215	1,204	1,013
% Missing							
–address	0.51	0.02	2.17	0.19	0.00	0.58	0.49
–email	0.05	1.09	9.84	0.00	0.05	0.42	0.99
–phone	17.09	2.80	97.21	4.46	18.27	0.91	100.00
–contact info of any kind	0.00	0.00	0.25	0.00	0.00	0.00	0.00

In order to ensure accurate representation of sub-groups in the study data, we then stratified the sampling frame by the three-way interaction of degree-granting institution, major field, and underrepresented minority (URM) status,[4] as shown in **Table 3.3**. Seven strata with no members were discarded, leading to **91** strata from which to sample.

Table 3.3: Sub-groups Used for Stratification

Degree-granting Institution (seven groups)	First Major Degree Field (seven groups)	URM Status (two groups)
4002	Liberal arts: Humanities	Underrepresented minority = Yes
4006	Liberal arts: Physical and biological sciences	Underrepresented minority = No
4008	Liberal arts: Social sciences	
4013	Liberal arts: Other (includes multidisciplinary fields)	
4016	Professional: Business	
4029	Professional: Engineering	
4030	Professional: Other (includes education, public health, social work, and other fields)	

3.2.3 Sample Selection

Out of the **24,529** members of the sampling frame, **15,000** cases were selected for participation in the survey. The selection procedure was as follows:

[4] URM status was defined at the institutional level. Generally, institutions considered students to be URM if they identified with one or more of the following racial/ethnic groups: African American/Black; Latino/a/x, Chicano/a/x, Hispanic; and Native American, Native Alaskan, Native Hawaiian or Pacific Islander. However, there was variation across institutions in how they categorized individuals, particularly regarding multi-racial/ethnic individuals. Note that the URM status variable was used only for sampling purposes and is not available to CBII data users. Users can create their own URM indicator using the **ident_ethnic** variables in the main survey file.

- 100% of URM individuals were selected in an effort to provide more precise estimates of the experiences of graduates from underrepresented groups.
- 100% of individuals who graduated from institution 4016 were selected because reliable information about URM status was not provided by this institution, and we did not want to prevent any URM individuals from taking the survey.
- 100% of individuals who graduated from institutions 4013, 4029, and 4030 were selected due to the limited number of cases available at those institutions.

Altogether, **8,947** individuals from the sampling frame were selected with certainty.

For non-URM individuals at institutions 4002, 4006, and 4008, a simple random sample without replacement was conducted. These **6,053** cases were allocated proportionately to their stratum population sizes. **Table A3** in the Appendix provides a more detailed breakdown of how the survey sample was allocated across strata.

3.3 Data Collection

Data collection services were provided by SRO at the University of Michigan using Illume, a web-based survey platform. Respondents self-administered the survey over the web after receiving a personalized link via email or letter. The survey could be completed on any web-enabled device, including computer, tablet, or smartphone. Illume allowed respondents to work on the survey over multiple sessions if needed.

3.3.1 Incentives

Initially, each respondent received a **$30** check as a token of appreciation for completing the survey. After approximately 20 weeks of data collection, the incentive was increased to **$50** in order to encourage participation as response rates decreased. SRO processed and mailed all respondent checks.

3.3.2 Overview of Communications

SRO interviewers from the Survey Services Lab (SSL) in Ann Arbor, Michigan were responsible for all communications with potential respondents. Respondents were invited to participate by a mailed letter as well as an email. Both the letter and the email included the survey links. Up to five email reminders were sent to those respondents who had not yet completed the survey at the time the communication was sent. All respondents who had not yet completed an interview by June 22, 2021 were also sent a notification that the incentive had been increased. **Table 3.4** below shows the timeline of communications along with the response rate for that week. Section 11.3 in the Appendix contains copies of all communications sent to potential respondents.

Table 3.4: Communications and Milestones

Key Milestones	Date	Cumulative Response Rate
Project Launch, Invitation Letter Mailed	2/05/2021	0.00%
Invitation Email Sent	2/09/2021	5.73%
Reminder Email #1	2/18/2021	9.45%
Reminder Email #2	2/24/2021	11.52%
Reminder Calling Start	2/25/2021	11.52%
Reminder Email #3	3/06/2021	12.73%
Reminder Email #4	4/22/2021	15.67%
Text Messaging Start	5/24/2021	16.79%
Incentive Increased to $50, Email Sent	6/22/2021	17.71%
Reminder Email #5	7/09/2021	18.13%
Last Day of Data Collection	7/26/2021	18.67%

3.3.3 Tracking Procedures and Responsive Design

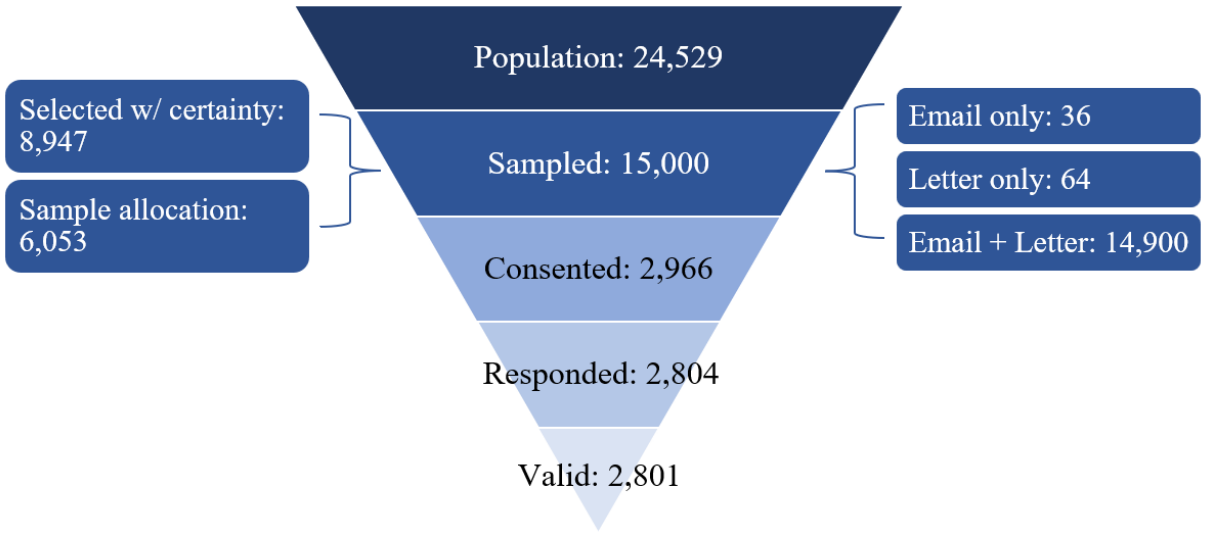
SRO employed both reminder calling and manual locating strategies for non-responders. Manual locating, also referred to as tracking, was carried out by two experienced field locators at SSL. The goal of tracking was to identify more accurate respondent phone numbers and email addresses than what had been provided by the contact information vendor. Locators made use of various locating tools and websites to attempt to find the most up-to-date contact information. Despite our best efforts to reach potential respondents, it is possible that some members of the survey sample never received their invitation to participate in the Alumni Survey.

In a process called “Responsive Design,” sample groups with lower response rates were strategically prioritized for reminder calling, manual locating, and text messaging by SSL staff. **Table A4** in the Appendix provides more detail about CBII responsive design efforts.

Reminder calls began after the second reminder email, the week of February 22, 2021. Reminder callers attempted to reach respondents by telephone. When contact with a respondent was made, the interviewer confirmed the respondent’s email and date of birth, then re-emailed the survey link. If a reminder call was unsuccessful, cases were eligible for manual locating. Manual locating began on May 8, 2021 for targeted groups with low response rates.

Text message follow-ups were introduced on May 24, 2021 for sample members whose phone numbers had been confirmed during reminder calling but had not yet completed the survey. After sending a text message, SSL staff followed up by resending the last email reminder that included the survey link. By the time the field period ended on July 26, 2021, **2,801** individuals had consented to take the survey, responded, and provided a valid survey response. **Figure 3.1**, below, provides a visual overview of the data collection process and resulting counts.

Figure 3.1: Overview of Selection, Participation, and Invitation Type Counts



3.3.4 Response Rates

The final response rate for the survey was approximately 19%. **Figure 3.2** shows the overall response rate over time along with when key milestones took place.

Table 3.5 shows the final percentage of nonresponse and response rates by institution, field of study, and URM status. “No Contact” indicates that the respondent did not respond to our efforts to make contact via telephone or text message. “Refusal” indicates a verbal refusal on the respondent’s part. The “Other” category includes all other instances of a non-valid response, such as an international, incarcerated, or deceased respondent, or not consenting to take the survey.

Figure 3.2. Response Rate Over Time

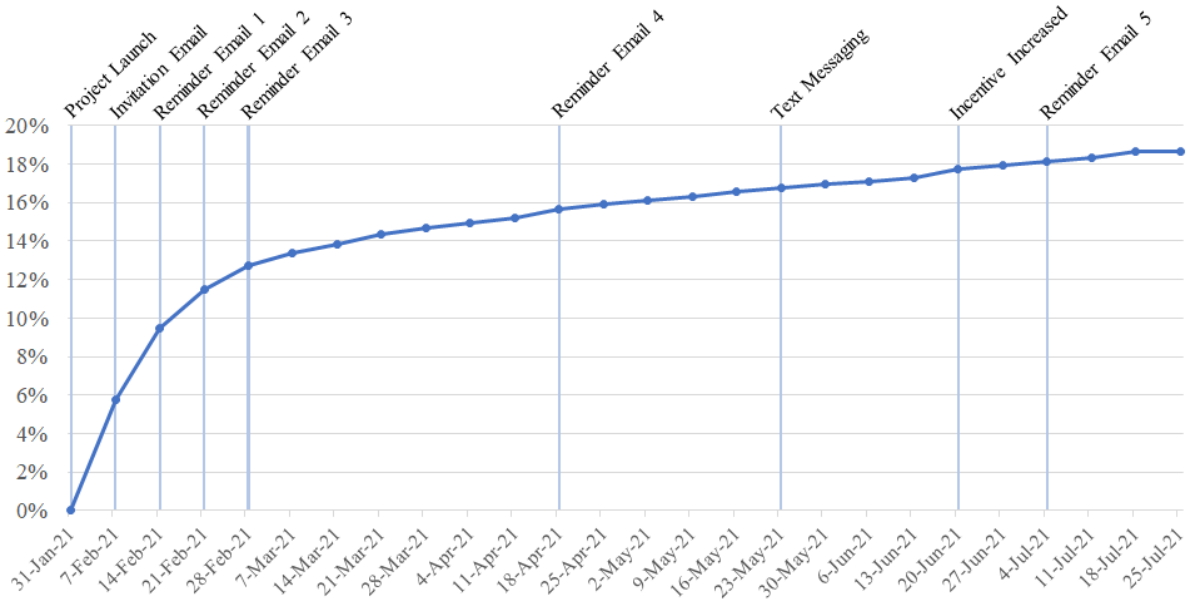


Table 3.5: Nonresponse and Response Rates by Institution, Field of Study, and URM Status

Demographic	Response = No			Response = Yes
	No Contact	Refusal	Other	
Institution				
4002	62.72%	2.14%	10.53%	24.61%
4006	71.76%	3.28%	8.32%	16.64%
4008	63.06%	2.72%	14.33%	19.89%
4013	56.23%	2.14%	12.08%	29.55%
4016	72.38%	3.91%	10.51%	13.19%
4029	63.70%	6.64%	13.79%	15.86%
4030	60.12%	4.74%	16.19%	18.95%
Field of Study				
Liberal Arts	66.19%	3.36%	11.54%	18.90%
Non-Liberal Arts	66.51%	3.54%	11.54%	18.41%
URM Status				
Non-URM	66.13%	3.44%	11.72%	18.72%
URM	67.41%	3.49%	10.66%	18.44%

3.3.5 Respondent Characteristics

Table 3.6 reports unweighted descriptive characteristics about the sampling frame and the survey respondents and compares it to a nationally representative sample of public college graduates from 2007-2008 from the Baccalaureate and Beyond (B&B) study.

The results indicate that, relative to the B&B sample, our sample and survey respondents had slightly fewer underrepresented minorities and more students who graduated in liberal arts and sciences fields (as opposed to other professional fields).

Table 3.6: Summary Statistics of CBII Samples and National Sample

Demographic	CBII Sampling Frame (*n* = 15,000)	CBII Survey Respondents (*n* = 2,801)	B&B 2007-2008 (PUBLIC) (*n* = 8,742)
Gender			
Female	0.54	0.55	0.55
Male	0.46	0.45	0.44
URM Status			
Non-URM	0.83	0.84	0.79
URM	0.17	0.16	0.21
Field of Study			
Arts & Humanities	0.15	0.18	0.12
Social Science	0.22	0.22	0.18
Biology/Physical Science	0.11	0.12	0.08
Engineering	0.08	0.08	0.06
Business	0.16	0.13	0.20
Multi/Lib Arts	0.05	0.03	0.04
Professional	0.22	0.24	0.32

Note: For the CBII sample, URM status was defined at the institutional level. Generally, institutions considered students to be URM if they identified with one or more of the following racial/ethnic groups: African American/Black; Latino/a/x, Chicano/a/x, Hispanic; and Native American, Native Alaskan, Native Hawaiian or Pacific Islander. However, there was variation across institutions in how they categorized individuals, particularly regarding multi-racial/ethnic individuals. From the B&B sample, “White” and “Asian” were considered Non-URM. URM included all other races/ethnicities.

3.4 Data File Descriptions

3.4.1 Overview

The CBII Alumni Survey data consists of two files: 1) individuals' responses to the web questionnaire (main survey data file) and 2) study participation indicators (participation data file). As with the other studies in the CBII series, underscores are used to separate variable name parts (primary constructs, sub-constructs, keywords, etc.). Variables that are created from other survey items, such as indexes, have [DERIVED] at the end of the variable label. Variables that also appear in the Core Administrative Data have [ADMIN] at the end of the variable label.

3.4.2 Survey Data File

The main survey data file contains respondent-level data for 2,801 valid survey responses. Among the information included in the survey data file are individual and institution IDs, responses from the survey, occupation and labor codes, weights, derived indexes, and selected administrative variables. The following subsections describe the different types of data included.

3.4.2.1 Linking Records

The main survey file contains one record per individual. Individuals are identified with the anonymized study ID for individuals, **id_person**. This variable can be used to link the survey data with other student-level data. However, users should be aware that there are 50 survey respondents who are not included in the Administrative Data and its related studies. This occurred because the survey sample was selected before the administrative data collection was finalized, and the CBII team was unaware that these 50 people lacked the data needed to be included in the Administrative Data population.

In the main survey file, the institutions from which the students graduated are identified using the anonymized study ID for CBII colleges and universities, **id_college**. The main survey file also contains variables that link to information about individuals and colleges in the CBII Contextual Data study. Respondent's current ZIP Code, **zip_current**, will link to ZIP Code Tabulation Areas in the National Neighborhood Data Archive (NaNDA) data (see Section 9.3.4.5). Several survey variables contain the Unit IDs of colleges and universities that the respondent applied to or attended prior to attending the CBII college they graduated from; these link to data from the Integrated Postsecondary Education Data System (IPEDS) data (see Section 3.4.2.3).

Section 1.3 provides more information about linking data across the CBII series.

3.4.2.2 Occupation and Industry Coding

CBII uses 2010 Census Occupation and 2007 Census Industry codes to describe respondents' employment. These are four-digit codes maintained by the U.S. Census Bureau and the Bureau of Labor Statistics. Trained coders from SRO assigned these codes based on open-text

responses to questions about respondents' work, work activities and duties, and job titles. Questionnaire items in **Table 3.7** were replaced with a corresponding occupation or industry code. These coded versions of the variables have **_code** at the end of their variable name and **[DERIVED]** in the variable label.

Table 3.7: Occupation and Industry Variables

Items in Questionnaire	Coding Applied	Variable in Data
labor_empl_work labor_empl_duties labor_empl_title	2010 Census Occupation	labor_empl_work_code
labor_notempl_work labor_notempl_duties labor_notempl_title	2010 Census Occupation	labor_notempl_work_code
labor_first_work labor_first_duties labor_first_title	2010 Census Occupation	labor_first_work_code
labor_empl_industr	2007 Census Industry	labor_empl_industr_code
labor_notempl_industr	2007 Census Industry	labor_notempl_industr_code
labor_first_industr	2007 Census Industry	labor_first_industr_code

3.4.2.3 College Lookup IDs

Aside from the college they graduated from, participants provided the names of up to three colleges they applied to (**col_serious_lookup_01 - 03)** as well as the first college they attended (**col_first_name**), if applicable. Along with the name of the college, the data file contains the corresponding IPEDS identification code (Unit ID). The Unit ID can be used to link to information about these colleges and universities in the CBII Contextual Data (see Section 9.2.4.5).

3.4.2.4 Weights

Sample weights were computed to allow for the computation of design-unbiased estimates and decrease non-sampling error bias due to coverage and nonresponse in the survey estimates. The weighting consists of three components:

- Design weights
- Nonresponse adjustment
- Calibration

Table 3.8 presents descriptive statistics of the three computed weights for CBII. All the weights sum up to the frame population size (24,529), an expected property of the weights. The unequal weighting effect of the final weight is 1.278. This means that, assuming there is no correlation between the weights and the study variables, the sampling variance of the estimates will be 1.278 larger than if no weights are used.

Only the final calibrated weight is available in the main survey file. This weighting variable, **weight**, and the stratification variable, **stratum**, are available for each respondent. See Section 11.1.3 in the Appendix for a detailed description of the weight calculation procedure.

Table 3.8: Descriptive Statistics for Weights

weight	*n*	min	Q1	Q2	Q3	max	SD	sum	Coef. of variation	Unequal weighting effect
Base	15,000	1.00	1.00	1.00	2.57	2.59	0.77	24,529	0.47	1.22
Nonresponse adjusted	2,801	2.50	5.25	8.02	11.78	25.00	4.23	24,529	0.48	1.23
Final calibrated	2,801	2.32	5.34	7.74	11.23	52.49	4.62	24,529	0.53	1.28

Note: SD = Standard deviation.

3.4.2.5 Scales and Indexes

Numerous validated scales are included in the survey. Examples include the Ryff Scales of Psychological Well-being (42-item version), Career Adapt-Abilities Scale--Short Form, Pluralistic Orientation Scale, Openness to Diversity and Challenge Scale, and the Everyday Discrimination Scale (short version). See the CBII Alumni Survey questionnaire for a comprehensive listing of scales and their sources.

Indexes were created for scales where possible and include **[DERIVED]** in the variable label. **Table A5** in the Appendix includes a list of indexes and how they were generated. Indexes for respondents with partial missing data were not created. Thus, indexes were created only for respondents with non-missing values for each of the variables that comprised the scale. This was not true for indexes that contained variables with coded skip logic. Cronbach's alpha was calculated for each mean index. All alpha coefficients are above **.60,** and all but two – **index_challen** and **index_artdone** – are above **.70.**

3.4.2.6 Administrative Data Variables

Select variables from the core Administrative Data were included in the main survey data file. These variables include **[ADMIN]** in the variable label. See Section 2.4 for further explanation of these variables.

3.4.3 Participation Data File

The participation data file contains one record for each of the 24,529 individuals in the sampling frame. It contains a total of eight variables, five of which are binary indicators of each population member's engagement with the survey, which is also summarized in the categorical variable **particip**. A small number of sample members lacked the contact information needed for SRO to send a letter or email invitation to the survey – this is indicated in the variable called

invitat_type. **Table 3.9** contains a summary of each type of participation observed in the survey population and its corresponding binary indicator in the data file.

Table 3.9: Participation Data Summary

Frequency	Participation	Indicator
24,529	Member of the sampling frame	populat
15,000	Sampled / invited to participate	sampled
2,966	Consented to take the survey	consent
2,804	Answered at least one survey question	respond
2,801	Provided a valid survey response	valid

3.4.4 Item Nonresponse

Three nonresponse labels were generated as a way to categorize missing data. Nonresponse labels are defined and applied as follows:

1. **-999 Missing (Not Answered)**
   Missing because the respondent did not answer the question. This might be due to the respondent breaking off from the survey early or being posed a question and not providing a response.

2. **-888 Not Applicable**
   Missing because respondent wasn't posed the question due to skip logic. For example, respondents who reported they did not participate in academic clubs were not posed the question about how impactful academic club participation was to them [**impact_academ** = *-888 (Not Applicable)* because **extrac_academ** = *0 (No)*].

3. **-777 Invalid Skip**
   Missing because respondent was not shown the question when they should have been. This applies to two variables: **impact_pub** and **impact_perform**. Respondents who answered '*1. Yes, I participated as a group member'* to **extrac_pub** or **extrac_perform** were not shown **impact_pub** or **impact_perform** due to a programming error.

No missing data were imputed. Users should consider how to handle missing data before conducting analyses.

3.4.5 Redactions

To ensure respondent and institutional confidentiality, we redacted information that could have been disclosive. Short answer variables underwent two stages of checking and redacting. First, automated syntax ran through each observation and replaced easily assumed identifying information, such as institution name or city. Second, study team members read through each

response and redacted any additional identifying information not removed by the automated syntax. The following variables in the survey data file include redactions:

- extrac_other_text
- highimp_other_text
- politic_orient_text
- politic_party_text

Information that was redacted was replaced with a descriptor enclosed in brackets and capitalized, for example [INSTITUTION].

3.4.6 Open Response Data

The following open-response variables appear in the questionnaire but were excluded from the main survey data file:

essay	labor_empl_industr	labor_notempl_title	labor_first_duties
labor_empl_work	labor_notempl_duties	labor_notempl_industr	labor_first_title
labor_empl_duties	labor_notempl_work	labor_first_work	labor_first_industr
labor_empl_title	labor_notempl_duties	labor_first_duties	final

Two variables, **essay** and **final**, are available in the Alumni Survey Open Response study. For more information on these variables, see Section 4.0.

4.0 Alumni Survey Open Response

4.1 Study Description

The Alumni Survey Open Response study includes respondents' written responses to two open-ended questions that were asked in the CBII Alumni Survey. These questions allowed respondents to write freely about their perceptions and experiences. One variable contains respondents' reflections on their college experiences (**essay**), and the other contains respondents' final thoughts at the conclusion of the questionnaire (**final**). This study only contains data for respondents who answered one or both of the open-ended questions. In total, there are 2,462 observations in the data. There are 2,429 responses for **essay** and 990 responses for **final**.

Because respondents could write freely, the topics discussed in the data range substantially. Users should be aware that some responses cover sensitive topics such as experiences of discrimination, sexual assault, suicidal ideation, and substance misuse. Respondents'

experiences of discrimination include incidents of hate speech and the use of racial slurs. ICPSR preserved the data as it was collected, without masking, and does not condone the use of these words.

4.2 Study Design

4.2.1 Target Population

The Alumni Survey target population are domestic bachelor's degree recipients who earned their degrees in academic year 2009-2010 from select CBII institutions. For detailed information about the survey's target population, see Section 3.2.1.

4.2.2 Sample Design

With the assistance of the Survey Research Operations (SRO) unit at the University of Michigan, we used the target population data to construct a sampling frame. For detailed information about the sample design, including the sampling frame and sample selection, see Section 3.2.2 and Section 3.2.3.

4.3 Data Collection

Data collection services were provided by SRO at the University of Michigan using Illume, a web-based survey platform. Respondents self-administered the survey over the web after receiving a personalized link via email or letter. The survey could be completed on any web-enabled device, including computer, tablet, or smartphone. Illume allowed respondents to work on the survey over multiple sessions, if needed. For detailed information about data collection procedures including incentives, communication, response rates, and respondent characteristics, see Section 3.3.

Two questions in the survey allowed respondents to write open-ended responses:

- **Essay**: Think back on the experiences and events that made up your undergraduate years. Choose one that seems especially important and explain it. Then discuss the ways it has contributed to your life today, in personal and/or professional terms.
- **Final**: Is there anything else you would like to tell us?

Illume timed out a survey session after 30 minutes of inactivity. Because of this, some respondents were logged out while they were writing their response to the open-response questions and lost their first response. Some participants make note of this in their second attempt at responding, but it is unknown how many participants this affected.

4.4 Data File Descriptions

4.4.1 Overview

The Alumni Survey Open Response data consists of one data file containing two qualitative variables extracted from the Alumni Survey. The variable **essay** contains respondents' written descriptions of important college experiences and the ways these experiences contributed to their lives. Variable **final** contains any additional information respondents wished to share at the conclusion of the survey..

We removed invalid survey participants and responses with missing data on both open-ended questions, resulting in 2,462 valid observations: 2,429 responses for **essay** and 990 responses for **final**. The **final** variable contains many non-meaningful responses (e.g., "No," "N/A," etc.). We did not consider non-meaningful responses as missing. Because of this, researchers should be cautious about answering research questions solely based on analysis of the variable **final**.

4.4.1.1 Linking Records

The data file contains one record per individual. Individuals are identified with the anonymized study ID for individuals, **id_person**. This variable can be used to link data with other student-level data. The file also includes **id_college**. See Section 1.3 for more information about linking data across the CBII series.

4.4.1.2 Item Nonresponse

One nonresponse label was generated as a way to categorize missing data for the Alumni Survey Open Response Study. The nonresponse label was defined and applied as follows:

1. **-999 Missing (Not Answered)**
   Missing because the respondent did not answer the question. This might be due to the respondent breaking off from the survey early or being posed a question and not providing a response.

The Alumni Survey data has additional nonresponse labels that are not applicable to the open-response questions. See Section 3.4.4 for an explanation specific to the survey data.

4.4.2 Redactions

The open-response format provided respondents the opportunity to share information about their lives in rich detail and therefore include potentially disclosive information. To ensure respondent confidentiality, we redacted information that could have contained personally identifiable information (PII). The variables underwent two stages of checking and redacting PII. First, an automated syntax ran through each observation and replaced easily assumed PII. Second, study team members read through each response and redacted any additional PII not removed by the automated syntax.

Information that was redacted was replaced with a descriptor enclosed in brackets and capitalized, for example [CBII INSTITUTION]. **Table 4.1** provides a general overview of which type of information was redacted and what it was replaced with. This table does not cover every redaction instance, but rather helped guide our thinking in what type of information to redact.

Table 4.1: Redaction Decision-Making and Replacements

Instance	Replacement
Proper names	[NAME]
CBII institutions	[CBII INSTITUTION]
Non-CBII institutions	[INSTITUTION]
Community colleges	[COMMUNITY COLLEGE]
CBII-institution colleges/schools	[COLLEGE/SCHOOL OF XXX]
Campus building names	[BUILDING] / [CENTER] / etc.
Campus names	[CAMPUS]
CBII institution mascots	[MASCOT]
Student clubs or programs	[CLUB] / [PROGRAM] / etc.
Greek sororities/fraternities	[SORORITY] / [FRATERNITY]
Living learning communities	[LIVING LEARNING COMMUNITY]
Student boards	[BOARD]
Scholarships	[SCHOLARSHIP]
Student newspapers	[NEWSPAPER]
Locations in the U.S. (counties, cities, states, etc.)	[COUNTY] / [CITY] / [STATE] / etc.
Study abroad cities	[CITY]
Employers or internships	[RESTAURANT] / [STORE] / [COMPANY] / etc.

To help ensure individual and institutional confidentiality, we searched for information about student clubs/activities or Greek sororities and fraternities via an online search engine (e.g., Google) or the CBII institution website to determine how prevalent the activity or program was. Activities or programs present at less than three CBII institutions were redacted, and activities or programs present at three or more CBII institutions were not redacted.

Institution colleges/schools include a standard academic description. For example, “School of Business” was redacted to [COLLEGE/SCHOOL OF BUSINESS]. The list of standardized colleges/schools include:

- Architecture
- Arts
- Business

- Communications
- Computer Sciences
- Education
- Engineering
- Health Sciences
- Liberal Arts and Sciences
- Medicine
- Natural Sciences
- Nursing
- Pharmacy
- Public Affairs and Policy
- Public Health
- Social Sciences
- Social Work

Student group and program redactions include additional descriptions to provide information about the type of programming. **Table 4.2** includes the type of descriptions and a brief overview of what they include.

Table 4.2: Standardized Student Programming Descriptions

Descriptor	Overview
Academic - [field of study]	Clubs or societies related to a major or other academic interest.
Cultural	Multicultural, international, or other identity-based student group.
Honors	Honors programs. This includes academic-based programs that students must apply to or qualify for. This does not include honors classes.
Interest	Clubs for special interests or hobbies.
Internship	Programs that provide internship opportunities. May include travel to other U.S. cities.
Leadership	Programs focused on developing leadership skills.
Mentorship	Programs that provide students with mentors. This does not include programs where the student was a mentor. Rather, the program provides mentorship support for the student.
Political	Student groups with a political or civic engagement focus. Does not include student-led governing bodies, such as student government.
Religious	Spiritual or religious groups.
Research	Research-based programs.
Service	Service organizations, on or off campus.
Social Justice	Programs focused on social justice or diversity efforts.
Student Success	Programs that support student success. May include financial support, advising support, etc.
Study Abroad	Study abroad programs. Must include travel outside of the U.S.

5.0 Enrollment and Awards

5.1 Study Description

The Enrollment and Awards study provides information on where students enrolled and the credentials they earned before and after attending CBII institutions. Thus, it is ideal for studying phenomena such as undergraduate transfer and graduate school enrollment. The study consists of National Student Clearinghouse (NSC) data on students' terms of enrollment and receipt of degrees, certificates, and other awards. It also contains a file of derived variables that summarize aspects of students' enrollment patterns and degree attainment.

The NSC is a nonprofit organization that facilitates postsecondary data reporting and academic verification. The NSC (2022a) estimates that their data covers 99% of postsecondary institutions in the United States. Thus, it serves as the most comprehensive source of educational attainment data on college students in the U.S.

5.2 Study Design

The target population for the Enrollment and Awards study is identical to that of the Administrative Data: bachelor's-seeking undergraduates who enrolled at CBII institutions since 2000. See Section 2.2.1 for more information about the Administrative Data population.

5.3 Data Collection

Systems B, F, and G provided CBII with NSC data for their members of the target population. To obtain this data, systems queried the NSC enrollment reporting and degree verification databases using a batch process. The lists used to define the target population were created in August, 2021. The queries were executed between August 2021 and March 2022. The NSC then returned student-level records for all terms of enrollment and awards earned by members of the target population across all institutions that report data to the NSC. These data reflect information about students at the time of the query.

For Systems B and F, less than 1% of students in their target populations could not be found in the NSC's records. All of the students from System G's target population were found in NSC's records. The NSC (2022b) states that there are multiple reasons why student records are not returned. These include FERPA blocks placed by the student and inconsistent information between the school and NSC records. Additionally, CBII has determined that a small number of students were omitted from the target population list that returned the NSC records, for unknown reasons.

5.4 Data File Descriptions

5.4.1 Overview

NSC data covers time periods before, during, and after students' enrollment in the CBII institutions. The data consists of three files for each system that provided data: an Enrollment file, an Awards file, and a Derived file. Every person who has a record in the Student file of the Administrative Data is represented in each of the three files. For a complete guide to understanding the files returned by the NSC, please refer to the StudentTracker Detailed Report.

The enrollment and awards data have been minimally processed by CBII. We did not deduplicate records and we retained the major descriptions and CIP codes exactly as they were provided by the NSC, even if the values are invalid. Before using this data, users should note that data cleaning will likely be necessary.

5.4.2 Enrollment File

Each row in the Enrollment file represents a student's period of enrollment (e.g., a single semester) on record with the NSC. Students who did not have enrollment records returned by the NSC have only one row, with missing values for all variables. Users should note that System G is missing data on several variables, as indicated in **Table 5.1**

Table 5.1: Variables Missing for Entire Systems, Enrollment File

Variable	System		
	B	F	G
class_level			X
major_descr_enroll_01			X
major_cipcode_enroll_01			X
major_descr_enroll_02			X
major_cipcode_enroll_02			X

5.4.3 Awards File

Each row in the Awards file represents an award earned by a student on record with the NSC. Students who did not have award records returned by the NSC have only one row, with missing values for all variables. There are no variables missing for entire systems.

5.4.4 Derived File

We created the derived variables from the raw enrollment and award data returned by the NSC as a convenience for users. Each row is unique at the student level.

The raw NSC data contains seemingly endless variations of award names. For instance, there are over 2,700 unique award titles in the System F raw data alone. In an effort to facilitate research on educational attainment, the CBII team created a set of derived variables that indicate the levels of awards students earned. These degree indicator variables are **degree_doctoral**, **degree_masters**, **degree_bachelors**, **degree_associates**, **degree_certificate**, and **degree_other**.

The degree indicator variables were created using the **degree_title** variables associated with each row of data on the raw NSC data. For example, if a row showed a student receiving a “Bachelor of Arts” in the **degree_title** variable, that student was identified as earning a bachelor’s degree and has a value of “1/Yes” for the derived **degree_bachelors** variable. These steps were taken for all awards levels.

Table 5.2 contains examples of the type of awards that were included in each award level. Note that the examples in the table are not an exhaustive list of all awards included. They are provided merely to illustrate the types of awards included in the indicator.

Table 5.2: Degree Indicator Variables in Derived File

Variable Name	Label	Example of Types of Awards Included
degree_doctoral	Student earned a doctorate, including professional	Doctor of Philosophy, Doctor of Medicine, Doctor of Pharmacy, Juris Doctorate
degree_masters	Student earned a master’s degree	Master of Business Administration, Master of Science, Master of Education
degree_bachelors	Student earned a bachelor’s degree	Bachelor of Arts, Bachelor of Science, Artium Baccalaureus
degree_associates	Student earned an associate degree	Associate of Arts, Associate of Science, Associate in Nursing
degree_certificate	Student earned a certificate	Teacher Certification, Certificate in Applied Science, Certificate Graduate
degree_other	Student earned other degree or award	Diploma, Non-Degree, High School Diploma

The **degree_highest** variable was created by identifying the highest award earned by a student. From highest to lowest, the awards are ranked doctoral, master’s, bachelor’s, associate’s, certificate, and other.

The **degree_highest_date** was created by identifying the date associated with the highest degree awarded to the students. Users should note that **degree_highest_date** does not necessarily indicate the most recent date, but the date of when the highest degree was awarded.

5.4.5 Linking Records

Individuals are identified with the anonymized study ID for individuals, **id_person**. This variable can be used to link the Enrollment and Awards data with other student-level data. NSC records also include an institutional identifier called **nsc_college_code**. This variable identifies the institution where a student enrolled or earned an award. Once the dash in **nsc_college_code** is removed, the code is identical to the U.S. Department of Education's Office of Postsecondary Education identifiers (OPE IDs) and thus can be linked to the IPEDS data included in the CBII Contextual Data. Similarly, the NSC data includes an institutional identifier called **unitid**. Unit IDs are also contained in the IPEDS data in the CBII Contextual Data study.

Because institutions that participate in CBII may not be identified in the data, these institutions have had their **nsc_college_code** and **unitid** replaced with the anonymized CBII **id_college** variable. This means that users cannot link to the Contextual Data for CBII institutions. See Section 1.3 for more information about linking data across the CBII series.

5.4.6 Missing Values

Missing values are set to -999 in all data files. Values can be missing in the enrollment and award files because that information was not provided by the NSC or a student was not found in the NSC search. Values can be missing in the derived files because a student was not found in the NSC search or a student did not have the necessary data to create a value for a variable (e.g., if a student did not receive any award, the **degree_highest** and **degree_highest_date** will be coded as missing).

6.0 Course Content

6.1 Study Description

The CBII Course Content study contains course section-level information on the academic content of students' courses for nearly all courses enrolled in by bachelor's-seeking undergraduate students at CBII institutions between 2000 and 2021. This includes standardized variables about the department the course is offered in, its College Course Map (CCM) code, and – at four partner systems – the course catalog description and full title associated with that course.

6.2 Study Design

The target population for the Course Content data is derived from the target population of the Administrative Data, in particular the Course data. The Course Content data contains records for sections of courses taken by bachelor's-seeking undergraduates enrolled at CBII institutions since 2000. See Section 2.2.1 for more information about the Administrative Data population.

6.3 Data Collection

6.3.1 Overview

Data for the Course Content study were collected from three sources: variables derived from the Administrative Data's Course files, variables collected from publicly available course catalogs, and variables created using machine learning models trained on four waves of the National Center for Education Statistics' Postsecondary Education Transcript Studies.

6.3.2 Course Content Variables Derived from Course Data

Many of the variables in the Course Content study are drawn from the Administrative Data's Course file or are derived from information in the Course files. Variables drawn directly from the Course files include **course_catalog_subject**, **course_catalog_number**, and **course_section_code**.

The variables **department_cipcode**, **ccm_code_two**, **ccm_code_prob_two**, **ccm_code_four**, **ccm_code_prob_four**, **ccm_code_six**, and **ccm_code_prob_six** were derived in part by using the values of **course_catalog_subject** and **course_catalog_title**. See Section 6.4 for further information concerning these variables' derivation.

6.3.3 Course Catalog Data Collection

The CBII study team collected course descriptions from course catalogs at four partner systems: A, B, C, and D. At System A, the CBII study team was provided access to an administrative API to access course catalog information. At System B, the CBII study team was provided course catalog data from the Registrar's Office. At System C, the CBII study team scraped publicly available online course catalogs. At System D, the CBII team used a combination of scraping publicly available online course catalogs and annotating information from online course catalogs.

The first term of course description data for each system is summarized in **Table 6.1** below. At Systems C and D, when course descriptions were collected through scraping publicly available online course catalogs, we include both the text of the description in **description** and the full HTML text associated with the description in **description_raw**.

Table 6.1: Availability of Course Description Data in Course Content Study

Variable	System						
	A	B	C	D	E	F	G
description	Available beginning Fall 2004	Available beginning Fall 2000	Available beginning Fall 2004	Available beginning Fall 2003	Not available	Not available	Not available
description_raw	Not available	Not available	Available beginning Fall 2004	Available beginning Fall 2007	Not available	Not available	Not available

After a first wave of data collection, the collected course descriptions were merged against the course sections present in the Administrative Data's Course files. A CBII study team member then manually checked the system's course catalog to examine any missing courses. If a course description was found, the course record was added to the dataset. If a course description was missing, the course description was marked as not present in the data.

6.3.4 College Course Map Data Collection

The Course Content data includes six variables relating to course sections' predicted College Course Map (CCM) codes. The CCM is a standardized hierarchical taxonomy of postsecondary course topics developed by the National Center of Education Statistics (NCES) to be applied to postsecondary transcripts in national surveys as part of the Postsecondary Education Transcript Studies (PETS). See the NCES website for additional information about the CCM. The CCM code structure is analogous to the Classification of Instructional Programs (CIP) code structure used to characterize the subject focus of postsecondary programs.

Using four waves of PETS, the CBII team trained a logistic regression model to predict a course's two-, four-, and six-digit CCM code using the text of **course_catalog_subject** and **course_catalog.** The four waves of PETS the CBII team used are associated with the High School Longitudinal Study of 2009, the Baccalaureate and Beyond Longitudinal Study of 2008-2012, the Beginning Postsecondary Students Longitudinal Study of 2004-2009, and the Beginning Postsecondary Students Longitudinal Study of 2012-2017 and we refer users to their respective documentation for more details on their data collection.

6.4 Data File Descriptions

6.4.1 Overview

For each unique course section in the Administrative Data's Course files, the Course Content data provide variables for the department the course is housed in, the specific course type, and – at four partner systems – a course catalog description associated with the course. The Course Content data is unique at the **id_college, id_course**, and **course_catalog_title** level.

In the following sections, we discuss the process for cleaning and redacting course descriptions, annotating **course_subject_codes** with CIP codes, and predicting CCM codes. Additional details about the creation of course content variables on a system-by-system basis are available to users in the Technical Appendix in the VDE.

6.4.2 Course Description Cleaning

Because course descriptions contain text that could identify institutions within our data, the CBII team redacted potentially identifying text including institution names, abbreviations, nicknames, mascots, and cities/locations. Information that was redacted was replaced with a descriptor enclosed in brackets and capitalized, for example [INSTITUTION].

In some academic years, course descriptions were collected from online course catalogs at System C and D. In these cases, the variable description contains a cleaned string containing the online course catalog description after removing any HTML tags. However, because HTML markup may be useful for extracting aspects of a course description, such as prerequisites, we also provide the full HTML text associated with the course catalog descriptions found through websites in **description_raw**.

6.4.3 Annotation of Department CIP Codes

To allow for comparison of course sections across schools, the CBII team annotated all **course_catalog_subject** values at Systems A, B, C, D, F, and G and the majority of **course_catalog_subject** values at System E with a six-digit CIP code: **department_cipcode**. The annotations were assigned such that a department's CIP code is the major that most closely matched the department's curricular content. See Section 11.1.4 in the Appendix for more detail about the process for annotating department CIP codes.

In the process of annotating, the study team identified a recurring set of courses that did not fit into a CIP code but may be of interest to academic researchers. These include **course_catalog_subjects** comprised entirely of study abroad courses, internships, first year and transition experience courses, living-learning communities, undergraduate research, honors courses, and service learning. These courses are annotated with a six-digit code beginning with 70. In some cases, a **course_catalog_subject** contained a mixture of these courses. We annotated them with the six-digit code 70.9999. Users who have access to the data in the VDE will receive a corresponding file titled 'department_cipcodes.xlsx' that provides more information about these codes.

6.4.4 Prediction of College Course Map Codes

To standardize courses across systems, the CBII team labeled the majority of course sections with the corresponding CCM Code. As discussed in Section 6.3.4, the CCM is a hierarchical taxonomy developed to standardize postsecondary transcripts across institutions. The CCM taxonomy is inspired by the CIP taxonomy and shares a common hierarchical structure. Each

course is associated with a two-, four-, and six-digit CCM code with two-digit codes indicating the general subject, four-digit codes narrowing the focus to a subcategory, and six-digit codes providing the most specific definition of a course's content. Because it is based on the CIP taxonomy, many CCM codes directly relate to CIP codes; for instance, the two-digit CCM code for business coursework, 52, is the same as the two-digit CIP code for business majors.

Users who have access to the data in the VDE will receive three documentation files that contain the corresponding labels for CCM codes at the two-, four-, and six-digit levels. The two- and six-digit level files contain the corresponding text label associated with those levels. Because the CCM Technical Report does not provide four-digit text labels, the four-digit level file contains values with the text of all six-digit labels that are associated with that four-digit level code.[5]

The scale of the CBII data makes expert human annotation of all course sections infeasible. Therefore, the CBII team developed a machine learning classification model to algorithmically predict the appropriate CCM code to apply to a given course section. Using previously annotated data drawn from four waves of PETS, the CBII team trained three logistic regression models to predict a course's CCM code at the two-, four-, and six-digit levels using features derived from **course_catalog_subject** and **course_catalog_title**. For more information about the training and evaluation of our machine learning model, see this technical report.

After training and evaluating our three models, the CBII team made predictions of 1) a course section's most likely CCM code and 2) the predicted probability that the CCM prediction is correct. Because the model heavily relies on features derived from **course_catalog_title** in its predictions, we did not make predictions for course sections for which this variable was missing.

Predicted CCM codes at the two-, four-, and six-digit level are contained in the variables **ccm_code_two**, **ccm_code_four**, and **ccm_code_six** respectively, while the model's predicted probability that this is the correct code is in **ccm_code_prob_two**, **ccm_code_prob_four**, and **ccm_code_prob_six**.

Because machine learning models typically perform worse than expert human annotation, we encourage users to consider the predicted probabilities when interpreting their results. If the model is not confident in a particular prediction and the interpretation of results relies on that prediction, users may want to perform a round of manual validation that the prediction is correct. Similarly, users may want to pre-specify a tolerance for error in the model and only use predicted CCM codes for which the model is very confident in its prediction. For instance, the user might use all machine learning predictions where the model has a predicted probability greater than .8 and manually annotate any courses with a predicted probability of .8 or lower.

[5] For instance, the four-digit level CCM code 14.01 is associated with two six-digit level CCM codes, 14.0101 - Engineering, General and 14.0102 - Pre-Engineering. The corresponding record in the four-digit level file is 14.01 - Engineering, General; Pre-Engineering.

Note that because we trained three separate models to predict CCM codes at the two-, four-, and six-digit levels, some records may have incongruous predictions across levels. For instance, a course section may have a two-digit prediction of 14 - Engineering but a six-digit prediction of 15.0101 - Architectural Engineering Technology/Technician. If consistency across two-, four-, and six-digit CCM codes is important for the user's research project, they may consider deriving two- and four-digit codes from the predicted six-digit code.

6.4.5 Linking Records

The Course Content data can be linked to the Administrative Data's Course files using the combination of **id_course**, **id_college,** and **course_catalog_title**. Once linked to the Course data, the Course Content data can be further linked to other student-level CBII data using the individual identifier, **id_person.** See Section 1.3 for more information about linking data across the CBII series.

6.4.6 Missing Values

Variables missing data are labeled "-999 Missing."

7.0 Advanced Placement (AP)

7.1 Study Description

The Advanced Placement (AP) study consists of student AP test and credit data provided by CBII institutions. Advanced Placement is a program run by the College Board that allows high school students to take college-level courses as part of their high school curriculum. AP exams, offered at the end of AP courses, are standardized tests that measure how well students learned the AP course subject matter. Some colleges and universities provide course credit or advanced placement (skipping required courses) in exchange for test scores above a certain threshold. For more information about AP tests and credit, see the College Board website.

As part of its initial data request to institutions, CBII requested student-level data on AP test results and credit received. We requested this data because AP test results are often used as measures of students' academic preparation and the availability of pre-college learning resources. They also factor into the accounting of credits earned and credits taken by students during college, as they are often accepted in college as credit toward an academic degree. For these reasons, and the fact that there was some broad availability of AP scores at several systems, CBII proceeded with collection.

7.2 Study Design

Student AP data were requested to complement the design of the CBII Administrative Data (see Section 2.2). We requested data on any AP test taken by a degree-seeking undergraduate in an

entering cohort since Fall 2000. In cases where a student took a test (e.g., Calculus AB) more than once, the maximum test was included and other instances omitted. Test scores, names, course credit, and course credit indicators were all requested. Note that students in the CBII target population who did not have any AP test data available are not included in the Student AP study.

7.3 Data Collection

Four systems were able to provide student-level AP data: A, D, F, and G. While IPEDS constitutes an external standard for the collection of the core administrative data, we were unable to locate such a standard for AP tests that would facilitate a quality control check. Therefore, as a basic quality check, we computed the fraction of students in incoming cohorts with at least one AP test, which we call the “AP coverage.” Our key takeaways are:

- Since Fall 2008, AP coverage in systems A, D, and F has increased steadily over time. We interpret this trend as increased AP test taking.
- Previous to Fall 2008, System A had AP coverage of <10%, which is likely an underreporting of the true prevalence of AP test taking by System A students.
- In Fall 2009, System G had a sudden increase in AP coverage.

Putting the key takeaways together, between Fall 2009 and Fall 2019, the coverage results might be roughly interpreted as the fraction of students that truly took one or more AP tests at the four systems. During this time, the coverage is smoothly increasing or decreasing, and shows no sudden, unexplained changes in any system. This does not exclude the possibility of systematic, long-term omission or other system-wide errors that might bias this estimate of the true fraction of students taking an AP exam. A fuller treatment of this question is beyond the scope of this document. We also did not attempt a proper accounting of the contribution of AP credit to a student’s total earned credits. AP credits could, for instance, explain apparent shortfalls in total earned credits observed in the Administrative Data’s Course and Term files.

7.4 Data File Descriptions

7.4.1 Overview

Data files are separated by system, unique at the **id_person** and **ap_test_name** (Systems A, F, and G) or **ap_course_credit** (System D) level. For example, a student that took four AP tests will therefore occupy four rows in this table, each with a different test name. **Table 7.1** shows the AP-related variables that are included in the study.

Table 7.1: Variables in the AP Student Data

Variable	Description
ap_test_name	The AP College Board Exam name. Approved users should consult the Technical Appendix in the VDE for notes on how naming conventions and the tests themselves have evolved over time.
ap_test_score	The score received for this test on a scale of 0-5.
ap_course_cred_ind	Credit received for a course (Y/N) as reported by the system.
ap_course_credit	Course subject and catalog number for which credit was received.

7.4.2 Linking Records

The individual identifier **id_person** can be used to link to other student-level data in the series. The system identifier **id_system** is included in the data for convenience, but additional student-level information (for instance college of entrance) can be merged from the Administrative Data's Student file. See Section 1.3 for more information about linking data across the CBII series.

7.4.3 Missing Values

Variables missing data are labeled "-999 Missing." **Table 7.2** shows which variables are missing for entire systems.

Table 7.2: Variables Missing for Entire Systems, AP Files

Variable	System			
	A	D	F	G
ap_test_name		X		
ap_test_score				
ap_course_cred_ind	X	X		X
ap_course_credit	X			X

8.0 Student Experience Analytics

8.1 Study Description

The CBII Student Experience Analytics Study contains measures of students' undergraduate experiences derived solely from information extracted from the Administrative Data. The study consists of two sets of derived measures: the Course Diversity measures and the Transcript of

the Future measures. They are provided as examples of the types of aggregate measures that can be distilled from CBII transcript data and demonstrate a major benefit of the CBII data vis-a-vis other sources such as the Baccalaureate and Beyond Longitudinal Study (B&B) and the Beginning Postsecondary Students Longitudinal Study (BPS). Specifically, the Student Experience Analytics are created from information on course-taking for the full universe of undergraduate students at CBII institutions. Information on the race or major of students' classmates is not available in these other national transcript studies but is present in CBII. Users are encouraged to explore the analytics' potential for measuring aspects of students' undergraduate experience and to develop their own measures in a similar vein.

8.2 Study Design

The Course Diversity and Transcript of the Future measures allow for the exploration of additional aspects of courses and broader learning contexts of students. Each of the seven CBII systems has a Course Diversity file and a Transcript of the Future file available. The Course Diversity measures identify the course-level ethnoracial composition of students in addition to the level of ethnoracial diversity in a course relative to the representation of students on campus. The Transcript of the Future measures provide further information about course context, the academic performance of students compared to other students in the course, and measures of the breadth and depth of the courses students pursue along their degree path using transcript data. The Transcript of the Future measures are reported at the student level. To generate these, course-level calculations were undertaken and then aggregated over all courses on a student's transcript.

8.3 Data Collection

Both sets of measures are derived from the CBII Administrative Data. Where applicable, columns that CBII has standardized are used as opposed to those that are raw. For example, **race_ethn_derived** is used instead of **race_ethn_raw** where required.

8.4 Data File Descriptions

8.4.1 Course Diversity Measures

Student records and transcripts allow researchers to examine the diversity among students in a variety of ways. One approach is to examine the ethnoracial diversity among students within and across CBII institutions. Previous research notes the many benefits of diversity to the educational and social development of undergraduate students, which can extend from the representation of different ethnoracial student groups within the classroom.

8.4.1.1 Measures Description

To support the exploration of student diversity's relationship to academic and social outcomes, variables were created to measure the ethnoracial diversity among students at the course level.

Courses with multiple sections offered during an academic term, such as large introductory courses, were treated as separate courses. Proportions and three indices were created for use in analyses. Using the indicator variables for whether a student was identified with a particular ethnoracial group (i.e., **black_ind**), first the total number of students identified with each group present in a course for which data is available was computed. Ethnoracial student groups include the following:

1. Black
2. Asian
3. Hawaiian or Pacific Islander
4. Hispanic/Latino
5. Native American
6. White
7. Multiracial
8. Race/ethnicity Unknown (includes non-U.S. students)

Using the total number of students in a course (**course_count_students**), a proportion for each ethnoracial group was created (i.e., **course_prop_black**). Additionally, the proportion of students of color (i.e., all students not identified as white or as race/ethnicity unknown) and underrepresented racial minority (i.e., all students not identified as white, Asian, or race/ethnicity unknown) were created for analyses.

Three diversity indices were calculated using suggested measures by Chang and Yamamura (2006): mean diversity, heterogeneity index, and diversity index. The mean diversity (*m*) was calculated by adding the proportions of each possible ethnoracial student group together and dividing by eight (**course_ethn_div_mean**). Below, the formulas for remaining indices are provided and explained, each of which range from 0 (no ethnoracial diversity) to 1 (equitable representation/diversity):

Heterogeneity index (course_ethn_div_hetero) =
1 - (proportion of largest group)

The heterogeneity index subtracts the proportion of the largest ethnoracial group from one and measures how concentrated enrollment in a given course is. The index provides a snapshot of how much one group accounts for the most students on a campus.

Diversity index (course_ethn_div_index) =
$1 - [(\text{prop Black})^2 + (\text{prop Asian})^2 + \ldots (\text{prop Race unknown})^2]$

To calculate the diversity index, the proportion of each of the eight ethnoracial student groups is squared. Then, the sum of these eight values is calculated and subtracted from one. The diversity index provides a single proportion that corresponds to the probability for having cross-racial interactions in a classroom.

In total, 13 measures of ethnoracial diversity per course were created along with a total count of students in a course. The multiple diversity measures allow researchers to be attentive to differences in group representation that may be hidden by only utilizing proportions without considering their relationship to one another; a limitation addressed by the diversity index, for example. Thus, these multiple measures allow researchers to explore the variation in the diversity of courses and its possible relationship to student experiences and outcomes in a variety of ways. These course-level measures can be used to aggregate to other levels of diversity that may be of interest to researchers such as diversity within academic majors or broader major divisions, upper and/or lower level courses, all courses taken by a student along their pathway, and identifying the ethnoracial diversity by each academic term rather than relying on a census approach to enrollment that does not capture variation in student enrollment and diversity across all terms offered at CBII institutions.

For additional information on these and other related student diversity measures that could be calculated for CBII data, see Chang and Yamamura (2006).

8.4.1.2 Linking Records

The Course Diversity file contains one record for each unique course in each term. Course Diversity measures can be joined to other CBII studies and linked to individual students by matching on the course information using the combination of **id_system, id_course**, **term_descr**, and **course_catalog_number**. See Section 1.3 for more information about linking data across the CBII series.

8.4.1.3 Missing Values

Variables missing data are labeled “-999 Missing.”

8.4.2 Transcript of the Future Measures

Two important motivations for the collection of transcript-level data are the contextualization of a student’s experiences during their studies and the alignment of those experiences with notions of a liberal arts education. With these in mind, we have begun development of analytic measures that might capture these important parts of a student’s education and release these as part of the CBII study.

The Transcript of the Future measures are inspired by the work of Koester et al. (2017), who note that the summative measures provided on typical transcripts -- cumulative GPA, degrees awarded, honors received -- reflect very crude aspects of students’ experiences and success in college. Instead, Koester et al. develop measures that answer questions not addressed by these summative measures, including:

> *Were the courses taken by the student difficult on average? Did the individual stand out from their peers? Were the courses representative of a broad intellectual experience, or did the student delve into detail in the chosen field of study? And with whom did they take courses?* (Koester et al., 2017)

These measures could potentially be combined with a portfolio of students' work to comprise a “Transcript of the Future,” improving on the outdated summative measures included in typical transcripts. The Transcript of the Future measures included with CBII are an extension of the measures in Koester et al. (2017) to multiple institutions.

The Transcript of the Future data file contains one record for each unique student in CBII, for a total of 1,311,818 observations across the entire dataset.

Measures derived from administrative data have the advantage of being measured for *all* students, not just those that completed a survey. They also are precisely defined, which is not the norm in treatments of, for instance, breadth and depth (but see Zemsky, Chapter 2, 1989). These measures have the disadvantages of not yet being rigorously validated and being drawn exclusively from the elements reported on a transcript, which omit important aspects of students’ experiences inside and outside the classroom.

Measure validation is an area of active research and will help improve and reimagine analytics that better capture the experiences we intend, and to discard those that do not. The measures described in Section 8.4.2.1 embody the first steps of this feedback process. In their development, we strive to stay closely coupled to familiar student experiences and elements of the liberal arts while creating quantitative measures that have a reasonable range. That is, those measures should show variation among individuals, and behave in a way that squares with our intuition about the experience we attempt to capture.

It is recommended that these measures be considered and compared most meaningfully at increasing levels of granularity: within systems, within colleges, and within majors. Comparisons among cohorts, and especially among systems, may be designed to investigate differences among these units, but this should be undertaken in parallel with an effort to understand how systems have evolved their majors, course classification schemes, and demographic information over time. The analytics we provide are sensitive to the eccentricities of the underlying data provided to us by the member institutions, but become less so when the comparisons are more localized. The enterprising researcher may find it appropriate, and is even encouraged, to recompute these analytics as a means to reduce sources of systematic variance.

8.4.2.1 Measures Description

transcript_course_effect and **transcript_student_effect**: These describe the student’s grades in the context of other students and the grading patterns of the courses they take. They are an effort to 1) estimate a student’s overall success at receiving high grades (student effect) relative to the other students they take courses with and relative to the grades given in a course, and 2) estimate the “difficulty” in receiving high grades in a course (course effect). The effects are estimated for all undergraduates in the system, in all courses over all time for which transcript data is available and is not relative to a cohort, major, or any other unit. The **transcript_student_effect** is reported in units of grade points and is centered on zero for

students performing at the centroid of all other students. The **transcript_course_effect** is also reported in grade points and ranges from 0-4. In Koester et al. (2017), these were called “student fixed effect” and “course fixed effect,” respectively.

In practice, every grade given to a student is modeled as a linear combination of all student and course fixed effects, estimated across the full array of student-course-term records. If there are *N* students of the time for which we have transcript data, and *M* total instances of all courses, and *C* total, the model may be written compactly as:

$$y_{ij} = \sum_{N}^{n} \beta_i \delta_{in} + \sum_{M}^{m} \beta_j \delta_{jm} + \varepsilon$$

where δ_{in} = 1 when *i=n* and 0 otherwise, and similarly for $\delta_{jm.}$, which effectively creates an *C x (N+M)* matrix that is set to 1 for every *(i,j)* that picks out a course *j*, in which this student *i* received grade y_{ij}. The rest of the students and courses are set to zero; that is these are all indicator variables. Each β_i is a student fixed effect and is identical to the **transcript_student_effect**. Each β_j is a course fixed effect. The coefficients in this model may be estimated by ordinary least-squares techniques in principle. However, because of the large size of the matrix representation, we resorted to numerical techniques to estimate these coefficients on the full course table.

From this model, each student receives a static student fixed effect, that is, a composite of their performance over all courses. Higher values indicate more success at achieving higher grades. The course fixed effect that comes out of the model is specific to a course. To actually compute the **transcript_course_effect** for a student, we compute the credit-weighted mean of course fixed effects over all courses a student took. A lower value of the course fixed effect in the model, and likewise a **transcript_course_effect**, indicates a tendency for courses to be less likely to assign high grades. This might loosely be interpreted as “difficulty,” and a low **transcript_course_effect** indicative of having taken more difficult courses.

Roughly speaking, in practice the **transcript_course_effect** is often approximately the mean grade of the course in a given term. Also, while the **transcript_student_effect** should range between -4 and 4, in practice, noise and statistical error can produce estimates that are outside of the expected range for the **transcript_student_effect** and **transcript_course_effect**.

transcript_course_format: A liberal arts education is often associated with a broad range of learning experiences. Ideally those experiences extend beyond the classroom into co-curricular settings, but to begin with they can be understood within the context of the classroom. Did a student spend most of their in-class time in one or two formats, such as a lecture or lab? Or were they spread across many formats such as independent study, studio, recitation, internships, etc.

To capture this, we compute the fraction of a student's total credits spent in each available course format at an institution (**course_code_comp**). The **transcript_course_format** is then simply the maximum fraction. For instance, a student spending 80%, 10%, and 10% of their total credits in studio, lab, and discussion respectively would receive a value of 0.8, as would a student spending 80%, 10%, and 10% in lecture, internship, and lab respectively. No course format is given a favored status in this prescription, and higher value of this fraction indicates a smaller range of experience.

transcript_breadth, transcript_depth: Depth and breadth are hallmarks of a liberal arts education. Depth in a particular field of study is encouraged, as is breadth in exposure to many different disciplines. While course depth and breadth may be conceptualized and measured in many different ways, we consider simple measures that can be directly computed from students' course-taking patterns..

Depth is simply defined as the credit-weighted mean of the catalog numbers of the courses on a student's transcript. Higher values indicate greater depth, since higher catalog numbers are generally associated with more advanced coursework in each subject or courses that require prerequisites.[6] Simplicity is the strength of this measure. It doesn't require knowledge of prerequisites, or of mapping courses to majors. These are both active areas of research. As a technical matter, catalog numbers with leading or trailing non-numerical characters are considered after removal of the offending characters.

Breadth is considered as the diversity of course subject codes among the courses each student takes:

$$Breadth \ = 1/\sum_{i}^{R} p^{-1}$$

Here, *p* is the proportion of the transcript contained in each of *R* subject codes. This formula is a specific version (with *q = 0,* it is the harmonic mean) of the Hill number (e.g., Page, 2010) that has been used in ecology to estimate the effective number of species. On the transcript, this is the effective number of subjects. A higher value indicates greater breadth.

This measure is agnostic to the specific subjects, but it has the advantage of being well-defined, which is both good and bad. In the calculation, this means that physics is implicitly considered as similar to math as it is to psychology. In practice in college curricula, breadth is often encoded in broad distribution requirements (e.g., science, social science, humanities, etc.), and thresholds are set to satisfy this breadth requirement, which often is not well defined. The Course Content study in the CBII series can be used to derive measures of course similarity directly and we encourage users interested in course content to consult that study (see Section 6.0).

[6] This is a simplification, as there are many exceptions to this pattern at CBII institutions.

transcript_div_major, transcript_div_demography: Exposure to and appreciation of diversity are increasingly cited as important liberal arts experiences. Transcript and administrative data provide a means to quantify diversity using academic (majors) or demographically based identities to characterize a student's network of classmates over their studies. Co-enrollment in a course is no guarantee of meaningful interaction; rather, it is measurement of the context in which a student is situated. Is the student's time in courses spent with other students that ultimately major in many different subjects? Do those courses serve students of many different demographic identities?

The metrics we develop are built at the course level. Each course receives a measure of the major diversity using the same formula as used for the breadth:

$$Diversity = 1/\sum_{i}^{R} p^{-1}$$

For each course, a proportion *p* is calculated for each $i \in R$ groups. In the case of **transcript_div_major**, *R* is the total number of distinct majors in the course. The student's **transcript_div_major** is then the credit-weighted mean of the diversities over all courses taken by the student. The major of record is drawn from **major_descr_01**. In the case of double majors, the field used in the calculation is drawn at random.

In the case of **transcript_div_demography**, the groups must be defined according to the available demographic measures. We consider **sex**, **race_ethn_derived**, and **citizen_usa**, and note that it is common to also include income level, but we did not do so here due to limited availability and standardizability of the income data from the CBII institutions. A student's demographic identity is defined here as the particular combination of sex, race/ethnicity, and U.S. citizenship: three categories for sex (including missing), nine for race/ethnicity (including missing), and four for citizenship (including missing). This results in 108 possible demographic identities, not all of which are realized. Each student in a course falls into one of these categories, and only one.

In the diversity calculation for a course, *R,* is the total number of these groups realized in the course, and we count the number of students in each group using the same calculation as above. As with **transcript_div_major**, a student's **transcript_div_demography** is then the credit-weighted mean of this course-level diversity over all courses.

As with the breadth, both of these diversity metrics equally weight majors or demographic groups. For instance, it implicitly assumes that Physics majors and Interdisciplinary Physics majors are just as similar to one another as Physics and English majors. And a student that is male/white/citizen is as similar to female/white/citizen as they are to female/Asian/resident. Again, users are encouraged to explore and produce alternative measures of major and demographic diversity to suit their own research questions.

transcript_stud_fac_ratio: This is akin to measures of course size. It begins by calculating a course size at the **course_id** level, as the total number of students in a term offering of a section of a course. It assumes that this instance of the course has one instructor. The **transcript_stud_fac_ratio** is then calculated as the credit-weighted mean of all course sizes on a student's transcript.

The following Student Experience Analytics variables were developed by Koester et al. (2017): **transcript_student_effect**, **transcript_course_effect**, **transcript_breadth**, and **transcript_div_major**. If you use them in your work, please cite with the following:

> *Koester, B. P., Fogel, J., Murdock III, W., Grom, G., & McKay, T. A. (2017). Building a transcript of the future. In Proceedings of the Seventh International Analytics & Knowledge Conference (pp. 209-308).*

8.4.2.2 Linking Records

Transcript of the Future measures can be joined to other CBII datasets by matching on **id_person** to other student-level data. See Section 1.3 for more information about linking data in the CBII series.

8.4.2.3 Missing Values

Variables with missing data are labeled "-999 Missing."

9.0 Contextual Data

9.1 Overview

The Contextual Data study provides two datasets that can be used to contextualize other data in the CBII series: the Integrated Postsecondary Education Data System (IPEDS), and the National Neighborhood Data Archive (NaNDA). These are well-documented, publicly available data provided within the VDE for researchers' convenience. The data packages for IPEDS are comma-separated text files (CSV) with minimal metadata. The data package for NaNDA contains Stata DTA and do-files, and an Excel file with minimal metadata. Documentation provided by the data producers is included for both studies and can also be found on their websites.

9.2 IPEDS

9.2.1 Study Description

The Integrated Postsecondary Education Data System (IPEDS) data are collected by the U.S. Department of Education and include information about institutions from 12 surveys in nine topic areas:

1. Academic Libraries
2. Admissions
3. Completions
4. Enrollment (Fall and 12-Month)
5. Finance
6. Graduation Rates (150% and 200%) and Outcome Measures
7. Human Resources
8. Institutional Characteristics
9. Student Financial Aid

The IPEDS data describe basic characteristics of institutions, enrollments, academic program completions and completers, graduation rates and other outcome measures, faculty and staff, institutional finances, institutional prices, student financial aid, admissions, and academic libraries. IPEDS data can be used to characterize the first non-CBII college a CBII Alumni Survey respondent attended after high school and the other institutions to which they applied and were accepted to. The identity and characteristics of other institutions a student applied and/or was accepted to can be used to characterize the college options a student had available to them. When linked with the CBII Enrollment and Awards data, it can also be used to characterize the non-CBII institutions students attended and earned degrees, certificates, and other awards from, including graduate school enrollment and degree completion for stop-outs and transfers.

9.2.2 Study Design

Every college, university, and technical/vocational institution that participates in the federal student financial aid programs (Title IV-eligible institutions) mandatorily reports to IPEDS. Institutions that are not eligible for participation in Title IV may request to be included. There is no information on what percentage of non-eligible institutions participate in IPEDS.

9.2.3 Data Collection

IPEDS data are reported to the federal IPEDS Data Collection System by institutions on an annual basis. Data are collected three times per year: Fall, Winter, and Spring. These data collection periods cover different survey components and have different corresponding data release dates. We highly recommend that users read the IPEDS Survey Methodology provided by the U.S. Department of Education to understand which time periods are covered by the various survey components, as they are not uniform. For example, Completions data, which is collected during the Fall reporting period, covers July 1 to June 30 of the previous academic year. The Human Resources data, which is collected in the Spring reporting period, reflects the number of employees on payroll as of November of the IPEDS collection year.

9.2.4 Data File Descriptions

9.2.4.1 Overview

The IPEDS data were downloaded from the IPEDS website on September 1, 2021, with the exception of the 2019-20 Final data and the 2020 Provisional data which were downloaded on September 8, 2022. The data cover 2004-2021. The tables are zipped by year. The ZIP files include all data tables for that year in CSV format, complete documentation in an Excel file, and a Read Me file in Word. Users are encouraged to examine this documentation before using the IPEDS datafiles.

9.2.4.2 Data Structure

Data are structured by survey, table topic, and academic year. Different survey components cover different time periods. The files are multiple record. **Table A6** in the Appendix provides an example from the 2011-2012 academic year.

9.2.4.3 Variable Names

Full variable names for each year are listed in the varTableXX tab of the Excel documentation in the VDE, in the column varName.

9.2.4.4 Missing Values

IPEDS uses blank for missing, -1 for “Not reported,” and -2 for “Not applicable.” Most variables are eligible for imputation and imputed values are noted in the FLAGS20XX table. The FLAGS20XX file has its own set of codes to indicate the status of or reason for the response. This file in addition uses -9 for “Not active.” A description of the imputation process can be found on page 18 of this IPEDS Methodology Report.

9.2.4.5 Linking Records

Every organization that has submitted data to IPEDS is assigned a unique six-digit Unit ID by the U.S. Department of Education. IPEDS data can be merged to the CBII Alumni Survey data using Unit IDs (**unitid**) as the merge variable. In the CBII Alumni Survey data, **unitid** are the values in the following variables: **col_first_lookup_00_ID, col_serious_lookup_00_ID, col_serious_lookup_01_ID, col_serious_lookup_02_ID,** and **col_serious_lookup_03_ID**.

See **Table A7** in the Appendix for linkage results associated with these variables. IPEDS data can also be merged to the CBII Enrollment and Awards data for non-CBII institutions using Unit IDs and Office of Postsecondary Education identifiers (OPE IDs). See Section 1.3 for more information about linking data across files in the series.

9.3 NaNDA

9.3.1 Study Description

The CBII Contextual Data contains the National Neighborhood Data Archive (NaNDA) data on neighborhood socioeconomic status and demographics. NaNDA is a publicly available data resource that measures neighborhood and community characteristics along numerous dimensions. NaNDA data can be used to characterize the ZIP Code Tabulation Area (ZCTA) of students' permanent address during college and where survey respondents lived at the time of the survey by merging on ZIP Code. If you use the data in your work, please cite the data using the following:

> *Melendez, R., Clarke, P., Khan, A., Gomez-Lopez, I., Li, M., & Chenoweth, M. (2020). National Neighborhood Data Archive (NaNDA): Socioeconomic status and demographic characteristics of ZIP Code Tabulation Areas, United States, 2008-2017: Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E120462V1-130760*

NaNDA contains other data that might be of interest. Researchers may request approval from ICPSR to have additional NaNDA data uploaded to their VDE space by contacting User Support at ICPSR-help@umich.edu.

9.3.2 Study Design

The NaNDA data cover all U.S. ZIP Codes except those that are large unpopulated places. ZIP Codes assigned to businesses only or single delivery point addresses will not necessarily appear as ZCTAs. Please see the Census documentation for further information.

9.3.3 Data Collection

Data were compiled from the U.S. Census and the American Community Survey (ACS). The study's creators describe the data collection process as follows: "[W]e extracted key census indicators related to race, ethnicity, age, income level, employment, poverty, and home ownership from the ACS 2012 five-year estimate (covering 2008-2012). We merged the variables with the same variables from the ACS 2017 five-year estimate (covering 2013-2017) and with each ZCTA's land area from the 2010 TIGER/Line shapefiles for ZIP code tabulation areas" (Melendez et al., 2020, p. 2).

9.3.4 Data File Descriptions

9.3.4.1 Overview

The NaNDA socio-economic status (SES) ZIP Code-level data were downloaded from OpenICPSR on April 19, 2021. The data are one observation per ZIP Code Tabulation Area and

contain many community and contextual variables (e.g., density, age and indicators of poverty, and overall measure of neighborhood disadvantage). The documentation, data, crosswalk from ZIP Code to ZIP Code Tabulation Area, and Stata do-file that was used to add ZIP Code to the SES data are included in the NaNDA.zip file within ICPSR's VDE. Users are strongly encouraged to examine this documentation before using NaNDA. Additional information can be found in the U.S. Census Bureau's explanation of ZCTA's relationship to ZIP Codes.

9.3.4.2 Data Structure

The data files contain one line per ZCTA.

9.3.4.3 Variable Names

Variable names are shortened words and numbers and are listed in the documentation.

9.3.4.4 Missing Values

Missing values are system missing. There are 16 observations where the ZCTA-level variables are populated but ZIP Code is missing. These will all be system missing when merged to the CBII data.

9.3.4.5 Linking Records

NaNDA can be merged with any CBII file that contains a variable whose values are ZIP Codes. In the NaNDA data, the variable is **ZIP_CODE**. In the CBII Alumni Survey data, the variable **zip_current** is the respondent's ZIP Code at the time of the survey. In the CBII Administrative Data's Student file, the variable **address_zip_derived** is the student's permanent address ZIP Code. See **Table A8** in the Appendix for linkage results associated with these variables. See Section 1.3 for more information about linking data across files in the series.

10.0 References

Chang, M. J., & Yamamura, E. (2006). Quantitative approaches to measuring student body diversity: Some examples and thoughts. In W. R. Allen, M. Bonous-Hammarth, & R. T. Teranishi (Eds.), Higher Education in a Global Society: Achieving Diversity, Equity and Excellence (pp. 369-386). Amsterdam: Elseiver.

Koester, B. P., Fogel, J., Murdock III, W., Grom, G., & McKay, T. A. (2017). Building a transcript of the future. In Proceedings of the Seventh International Analytics & Knowledge Conference (pp. 209-308).

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). New York: John Wiley & Sons.

Little, R. J. A., & Vartivarian, S. (2003). On weighting the rates in non-response weights. *Statistics in Medicine*, *22*(9), 1589-1599. https://doi.org/10.1002/sim.1513

Melendez, R., Clarke, P., Khan, A., Gomez-Lopez, I., Li, M., & Chenoweth, M. (2020). National Neighborhood Data Archive (NaNDA): Socioeconomic status and demographic characteristics of ZIP Code Tabulation Areas, United States, 2008-2017: Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E120462V1-130760

National Student Clearinghouse (2022a). *Clearinghouse fact sheet*. OneStop Resource Center from the National Student Clearinghouse RSS2. Retrieved November 30, 2022, from https://studentclearinghouse.info/onestop/portfolio-item/clearinghouse-fact-sheet/

National Student Clearinghouse (2022b). *StudentTracker® for Colleges & Universities user manual*. OneStop Resource Center from the National Student Clearinghouse RSS2. Retrieved November 30, 2022, from https://studentclearinghouse.info/onestop/portfolio-item/studenttracker-for-colleges-universities-user-manual/

Page, S. E. (2010). *Diversity and complexity.* Princeton University Press.

Valliant, R., Dever, J., & Kreuter, F. (2013). *Practical tools for designing and weighting survey samples* (Vol. 1). New York: Springer.

Zemsky, R. (1989). Structure and coherence: Measuring the undergraduate curriculum. *Curriculum Data Base Series. Publications Desk, Association of American Colleges, 1818 R St., NW, Washington, DC 20009.*

11.0 Appendix

11.1 Additional Information

11.1.1 Academic Calendars at CBII Systems

All systems except System G use a semester-based academic calendar. System G uses a quarter-based academic calendar; Fall, Winter, Spring, and Summer. We designate Fall and Spring as the two standardized semesters for institutions on a semester system. For semester systems, Summer, Spring/Summer, and other terms that are administered after Spring are collapsed into a single “Summer” term. At Systems D, E, and F, a shortened “Winter” term is occasionally found between the Fall and Spring semesters. This is generally called “Winter” as well but is understood to not be a standard semester or an analog to the quarter system of System G. Accommodating this term causes an apparent skip in the coded numbering of terms.

There was concordance among institutions that the Fall term is the beginning of the academic year; the first academic term is generally “Fall XXXX,” but term/semester organization thereafter varied among institutions. This was due mainly to 1) the existence of shortened terms interspersed among primary semesters, and 2) the adherence to either a semester system or a quarter system.

11.1.2 Assigning and Standardizing Major CIP Codes

Four systems – A, C, E, and F – provided six-digit CIP codes for the majority of major descriptions. For systems A, C, E, and F, we took the CIP codes as provided by the systems. For records in which the major description was not missing, but the CIP code was missing, an analyst manually annotated descriptions of majors with the appropriate 2020 CIP codes by comparing the content of majors on institutional websites against the descriptions of CIP codes provided by the National Center for Education Statistics. In rare cases of ambiguity, a second analyst examined the same majors and the two, in concert, decided on the appropriate CIP code.

System E provided some major descriptions with the six-digit CIP code 99.9999. However, an identical major description could also be associated with a different six-digit code at the same institution or in the same term. Because it is unclear why some records were labeled 99.999 when similar majors were labeled differently, we did not correct the limited number of 99.9999 CIP codes in System E.

Systems B and D did not provide CIP codes, and one analyst manually annotated text descriptions of all majors with the appropriate 2020 CIP codes by comparing the content of majors on institutional websites against the descriptions of CIP codes provided by the National

Center for Education Statistics. In rare cases of ambiguity, a second analyst examined the same majors and the two, in concert, decided on the appropriate CIP code.

System G did not provide CIP codes but did maintain a table mapping the majority of major descriptions to six-digit CIP codes on their institutional website. To begin, we scraped this table to get CIP codes for the majority of text descriptions. One analyst then manually annotated text descriptions of all majors with the appropriate 2020 CIP codes by comparing the content of majors on institutional websites against the descriptions of CIP codes provided by the National Center for Education Statistics. In rare cases of ambiguity, a second analyst examined the same majors and the two, in concert, decided on the appropriate CIP code.

11.1.3 Alumni Survey Weight Calculation

First, design weights were calculated for each of the survey sample members as follows:

$$w_{hi.0} = \frac{N_h}{n_h}, i = 1, \ldots, n_h; h = 1, \ldots, 91$$

Where N_h is the number of domestic 2009-2010 college graduates in the sampling frame in stratum *h,* and n_h is the number of graduates in the survey sample in stratum *h*. The 91 strata correspond to the cross-classification of institution, field of major degree, and under-represented minority (URM) status.

Next, these weights were adjusted for nonresponse. If respondents and nonrespondents are systematically different with respect to the study's outcomes, the survey estimates can be subjected to nonresponse bias. If nonresponse follows a missing at random (MAR) mechanism, this nonresponse bias can be mitigated through a weighting adjustment (Little & Rubin, 2019). To that end, a nonresponse weighting class adjustment was performed over the CBII data (Valliant et al., 2013). In this approach, the inverse of the response rate of classes formed by auxiliary variables were used as nonresponse adjustment factors as follows:

$$a_{hi} = \frac{n_k}{r_k}, i = 1, \ldots, n_k$$

Where r_k is the number of survey respondents in weighting class *k*. In order for this adjustment to be effective, that is, reduce nonresponse bias while not decreasing estimate precision, the covariates used to form the weighting classes should be correlated with both the survey response and study outcome (Little & Vartivarian, 2003). In this case, we used the stratification variables (institution, field of major degree and under-represented minority status) as covariates.

Three out of the 91 strata had no survey respondents. Because every weighting class should have at least one survey respondent, these strata were collapsed with three others to form

weighting classes with at least one respondent. Therefore, in total, there were 88 nonresponse weighting classes. The nonresponse-adjusted weight was then computed as the product of the design weight and the nonresponse adjustment factor.

Finally, the adjusted weights underwent calibration. In this final weighting step, the nonresponse-adjusted weights went through a calibration adjustment, in which the weighted distributions were matched on a set of auxiliary variables to the population distribution obtained from the sampling frame. We evaluated through regression models the main and two-way interaction effects on fourteen selected study outcomes and indexes. We then decided to use in the calibration only the dimensions that were significant in the regression models at a 5% level to at least three study outcomes or scale indexes.

Calibration adjustment was performed using a raking procedure through the rake function in R survey package. The nonresponse adjusted weight was used as the input weight for this procedure.

11.1.4 Details on Annotating Department CIP Codes

At Systems A and C, this annotation process was informed by CIP codes that were provided by the systems. At all other systems, this annotation was done by two CBII team members on the basis of the course catalog descriptions associated with that department drawn from publicly available course catalogs and the course titles associated with that department drawn from the variable **course_catalog_title** in the Administrative Data's Course file. In the following paragraphs, we describe the annotation process for each system.

System A's raw course files provided a CIP code for departments. System C identifies each course with an eight-digit code in which the first six digits correspond to CIP classifications. To identify the appropriate CIP code for each subject, we identified the most common six-digit CIP code associated with each **course_catalog_subject**.

The raw course files at Systems B, D, E, and F did not provide a CIP code for departments. At Systems B, D, E[7], and G, several CBII team members manually annotated each unique **course_catalog_subject** with the CIP code that most closely matched the subject of the course.[8]

Annotation of **course_catalog_titles** followed a process with several steps. First, a study team member searched for **course_catalog_subject** in the system's course catalog and identified the closest CIP code to that subject's content. If the study team member was unable to find the **course_catalog_subject** in the system's course catalog or was unable to make a determination regarding the appropriate CIP code on the basis of the course catalog, a second team member examined the associated values of **course_catalog_title** present in the

[7] Excluding Institution 4014, which doesn't comprise a large number of enrollments.
[8] At System F, the majority of **course_catalog_subject** values are distinguished by a macro general topic and a more specific subtopic, with the macro and specific topic distinguished by a "-" character. A team of CBII team members annotated each unique macro topic if the **course_catalog_subject** had a "-" character, and each subject if it did not. For instance, at System F, the **course_catalog_subject** values of CSCI, CSCI-A, and CSCI-OS were all annotated to the same CIP code, 11.0701 - Computer Science.

Administrative Data Course file. If the closest CIP code was still unclear after examining the course titles in the administrative data, the **department_cipcode** was labeled as “99.9999 – Other/Unable to Annotate.”

11.2 Tables

Table A1: Unique Keys in CBII Data Files

Data File	Unique Key
Administrative Data	
Student	id_person
Term	id_person + id_college + term_code
Course	id_person + id_college + id_course + course_code_comp + course_grade_basis[a]
Alumni Survey	
Participation	id_person
Main	id_person
Alumni Survey Open Response	
–	id_person
Enrollment & Awards	
Enrollment	Raw data from NSC may contain duplicate observations
Awards	Raw data from NSC may contain duplicate observations
Derived	id_person
Course Content	
–	id_course + id_college + course_catalog_title
Advanced Placement	
–	id_person + ap_test_name
Student Experience Analytics	
Transcript of the Future	id_person
Course Diversity	id_course + term_descr + course_catalog_number
Contextual Data	
IPEDS	unitid
NaNDA	ZIP_CODES

[a]See Technical Appendix’s ‘Data Errors’ tab for exceptions.

Table A2: IPEDS Comparison Data for the Administrative Data

Term	System A			System B			System C			System D			System E			System F			System G		
	CBII	IPEDS	Ratio	CBII	IPEDS	Ratio	CBII	IPEDS	Ratio	CBII	IPEDS	Ratio	CBII	IPEDS	Ratio	CBII	IPEDS	Ratio	CBII	IPEDS	Ratio
Fall 2000[a]	5,229	5,418	96.5%	884	817	108.2%	2,559	3,135	81.6%	–	–	–	6,369	13,881	45.9%	4,425	6,936	63.8%	–	–	–
Fall 2001[a]	5,323	5,540	96.1%	939	881	106.6%	2,897	3,475	83.4%	–	–	–	8,252	14,570	56.6%	6,079	6,815	89.2%	–	–	–
Fall 2002[a]	4,981	5,187	96.0%	1,029	983	104.7%	3,117	3,457	90.2%	–	–	–	9,452	15,210	62.1%	6,261	7,080	88.4%	–	–	–
Fall 2003[a]	5,331	5,550	96.1%	1,058	1,031	102.6%	3,107	3,325	93.4%	1,310	1,317	99.5%	11,346	15,879	71.5%	6,000	6,784	88.4%	3,248	4,043	80.4%
Fall 2004[a]	5,763	6,037	95.5%	950	923	104.0%	3,305	3,367	98.2%	1,479	1,480	99.9%	12,667	17,194	73.7%	5,749	6,352	90.5%	3,548	3,629	97.8%
Fall 2005[a]	5,800	6,113	94.9%	1,061	1,036	102.4%	3,451	3,445	100.2%	1,441	1,444	99.8%	13,031	17,498	74.5%	6,362	6,949	91.6%	3,924	4,338	90.5%
Fall 2006	5,765	6,128	94.1%	1,530	1,138	134.5%	5,871	3,435	170.9%	1,468	1,367	107.4%	19,761	31,856	62.0%	7,447	8,183	91.0%	6,494	6,431	101.0%
Fall 2007	6,344	6,747	94.0%	1,638	1,629	100.5%	6,116	6,099	100.3%	1,513	1,516	99.8%	21,016	32,724	64.2%	7,510	8,189	91.7%	6,430	6,366	101.0%
Fall 2008	6,107	6,542	93.4%	1,587	1,595	99.5%	7,311	7,109	102.8%	1,489	1,510	98.6%	21,904	34,094	64.3%	7,809	8,499	91.8%	6,037	5,929	101.8%
Fall 2009	6,479	6,949	93.2%	1,657	1,661	99.8%	6,584	6,301	104.5%	1,472	1,496	98.4%	22,511	35,172	64.0%	7,738	8,347	92.7%	5,876	5,763	102.0%
Fall 2010	6,927	7,424	93.3%	1,645	1,622	101.4%	7,506	7,808	96.1%	1,575	1,601	98.4%	21,403	31,123	68.8%	7,398	8,021	92.2%	6,399	6,272	102.0%
Fall 2011	6,659	7,124	93.5%	1,614	1,611	100.2%	7,682	7,793	98.6%	1,558	1,588	98.1%	23,418	34,042	68.8%	7,854	8,390	93.6%	7,005	6,843	102.4%
Fall 2012	6,583	7,076	93.03	1,667	1,656	100.7%	7,632	7,756	98.4%	1,437	1,455	98.8%	23,738	32,992	72.0%	8,035	8,556	93.9%	6,906	6,781	101.8%
Fall 2013	6,669	7,166	93.1%	1,762	1,750	100.7%	6,678	6,828	97.8%	1,465	1,479	99.1%	24,424	33,580	72.7%	8,163	8,555	95.4%	7,567	7,458	101.5%
Fall 2014	6,983	7,434	93.9%	1,840	1,814	101.4%	8,587	8,720	98.5%	1,467	1,484	98.9%	25,751	35,470	72.6%	8,311	8,630	96.3%	7,585	7,459	101.7%
Fall 2015	6,561	6,959	94.3%	1,801	1,782	101.1%	8,931	9,014	99.1%	1,415	1,435	98.6%	26,057	36,397	71.6%	8,585	8,774	97.9%	7,668	7,711	99.4%
Fall 2016	7,334	7,749	94.6%	1,758	1,726	101.9%	9,377	9,464	99.1%	1,400	1,416	98.9%	25,947	36,454	71.2%	8,461	8,571	98.7%	9,118	9,016	101.1%
Fall 2017	7,458	7,934	94.0%	1,822	1,756	103.7%	9,801	10,159	96.5%	1,482	1,492	99.3%	28,178	37,851	74.4%	8,720	8,771	99.4%	9,640	9,535	101.1%
Fall 2018	7,451	7,941	93.8%	1,801	1,755	102.6%	10,159	10,473	97.0%	1,218	1,230	99.0%	29,011	38,353	75.6%	8,721	8,777	99.4%	8,572	8,457	101.4%
Fall 2019	7,633	8,108	94.1%	1,761	1,728	101.9%	10,031	10,277	97.6%	994	1,003	99.1%	14,879	38,746	38.4%	8,894	8,953	99.4%	8,826	9,173	96.2%
Fall 2020	7,702	8,184	94.1%	–	–	–	–	–	–	–	–	–	–	–	–	4,256	8,518	95.0%	8,436	8,500	99.3%
Fall 2021	8,230	8,865	92.8%	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

[a]IPEDS does not have transfer student data for Fall 2000 - Fall 2005 so numbers reflect only first-time freshmen in these years for both IPEDS and CBII. The table also does not count students that enter in terms other than Fall, for which IPEDS counts are not available.

Table A3: CBII Sample Allocation Across Strata for the Alumni Survey

Institution	Total	Physical & Biological Sciences		Social Sciences		Humanities		Other Liberal Arts & Sciences		Engineering		Business		Other Prof.	
		URM	Non URM	URM	Non URM	URM	Non URM	URM	Non URM	URM	Non URM	URM	Non URM	URM	Non URM
4002	2,755	**48**	251	**259**	500	**157**	307	**38**	49	**96**	337	**34**	124	**121**	434
4006	2,344	**51**	422	**129**	688	**43**	161	**28**	139	**27**	159	**13**	164	**50**	270
4008	2,393	**29**	145	**52**	270	**55**	356	**47**	88	0	0	**25**	368	**137**	821
4013	**1,076**	**§**	**154**	**§**	**126**	**19**	**255**	**§**	**§**	**0**	**0**	**§**	**169**	**29**	**289**
4016	**4,215**	**§**	**354**	**36**	**812**	**31**	**531**	**§**	**214**	**§**	**336**	**64**	**1,179**	**25**	**614**
4029	**§**	**70**	**94**	**206**	**102**	**148**	**86**	**50**	**§**	**75**	**94**	**29**	**16**	**126**	**99**
4030	**§**	**§**	**67**	**24**	**137**	**14**	**147**	**0**	**15**	**0**	**0**	**14**	**207**	**44**	**340**
Total	15,000	**218**	1,487	**§**	2,635	**467**	1,843	**172**	**520**	**§**	926	**§**	2,227	**532**	2,897

Note: URM = Underrepresented Minority. Bold numbers indicate certainty strata. § indicates suppressed due to small *n.*

Table A4: Sample Prioritization for the Alumni Survey

Start Date	End Date	# of Individuals	Intervention Type	Effort Type	Cases Selected
2/25/2021	5/20/2021	665	Prioritization	Reminder Calling	All non-final URM sample from institution 4029
2/24/2021	3/3/2021	3,930	Hold		All non-final 4016 sample put on hold from reminder calling and locator efforts due to natural disaster.
3/06/2021	3/23/2021	129	Prioritization	Reminder Calling	4016 Prioritized due to low RR
4/14/2021	6/20/2021	2,495	Prioritization	Reminder Calling	All URM sample
4/28/2021	5/20/2021	231/441	Prioritization	Reminder Calling	4029 & 4016
5/09/2021	6/24/2021	113	Prioritization	Locating	4029
5/20/2021	7/26/2021	411	Prioritization	Reminder Calling	4030
5/30/2021	7/26/2021	349	Prioritization	Text Messaging	4030 & 4029

Table A5: Indexes in Alumni Survey

Index	Component Variables		Scale
index_impact[a]	impact_academ impact_multic impact_pub impact_service impact_perform impact_governm impact_greek impact_athlet impact_clubsp impact_spirit impact_extrac_other impact_residen impact_livingl	impact_classes impact_studyab impact_honors impact_senior impact_divers impact_commun impact_faculty impact_interns impact_writejob impact_onjob impact_offjob impact_highimp_other	Mean index of up to 25 variables on a 5-point scale
index_integr	integr_pastev integr_career	integr_lifeev integr_action	Mean index of 4 variables on a 5-point scale
index_belong	belong_valued belong_belong	belong_enroll	Mean index of 3 variables on a 5-point scale
index_challen[a]	challen_mental challen_physic challen_family	challen_money challen_onjob challen_offjob	Mean index of 4-6 variables on a 5-point scale
index_divers	divers_ideas divers_cultur divers_challen	divers_perspec divers_backgr	Mean index of 5 variables on a 5-point scale
index_plural	plural_multipl plural_cooper plural_negot	plural_challen plural_toleran	Mean index of 5 variables on a 5-point scale
index_artdone	artdone_creativ artdone_writing artdone_read	artdone_music artdone_show	Mean index of 5 variables on a 4-point scale
index_artappr	artappr_think artappr_see	artappr_leave artappr_suffer	Mean index of 4 variables on a 5-point scale
index_generat	generat_skill generat_role generat_told generat_influen	generat_plan generat_pastexp generat_unique	Summative index of 7 variables
index_eudaim_positive	eudaim_positive_01 **eudaim_positive_02** **eudaim_positive_03** eudaim_positive_04	eudaim_positive_05 **eudaim_positive_06** eudaim_positive_07	Mean index of 7 variables on a 6-point scale
index_eudaim_autonom	eudaim_autonom_01 eudaim_autonom_02 **eudaim_autonom_03** **eudaim_autonom_04**	eudaim_autonom_05 **eudaim_autonom_06** eudaim_autonom_07	Mean index of 7 variables on a 6-point scale
index_eudaim_environ	eudaim_environ_01 **eudaim_environ_02** **eudaim_environ_03** eudaim_environ_04	**eudaim_environ_05** **eudaim_environ_06** eudaim_environ_07	Mean index of 7 variables on a 6-point scale

index_eudaim_person	**eudaim_person_01** eudaim_person_02 **eudaim_person_03** eudaim_person_04	**eudaim_person_05** eudaim_person_06 **eudaim_person_07**	Mean index of 7 variables on a 6-point scale
index_eudaim_purpose	**eudaim_purpose_01** **eudaim_purpose_02** **eudaim_purpose_03** eudaim_purpose_04	eudaim_purpose_05 **eudaim_purpose_06** eudaim_purpose_07	Mean index of 7 variables on a 6-point scale
index_eudaim_selfacc	eudaim_selfacc_01 eudaim_selfacc_02 **eudaim_selfacc_03** eudaim_selfacc_04	**eudaim_selfacc_05** **eudaim_selfacc_06** eudaim_selfacc_07	Mean index of 7 variables on a 6-point scale
index_eudaim_total	index_eudaim_positive index_eudaim_autonom index_eudaim_environ	index_eudaim_person index_eudaim_purpose index_eudaim_selfacc	Mean index of 6 variables on a 6-point scale
index_adapt	adapt_01 adapt_02 adapt_03 adapt_04 adapt_05 adapt_06	adapt_07 adapt_08 adapt_09 adapt_10 adapt_11 adapt_12	Mean index of 12 variables on a 5-point scale
index_civic_politic	civic_politic_friends civic_politic_agree civic_politic_disagree	**civic_politic_avoid** civic_politic_action civic_politic_read	Mean index 6 variables on a 5-point scale
index_discrim_freq	discrim_freq_01 discrim_freq_02 discrim_freq_03	discrim_freq_04 discrim_freq_05	Mean index of 5 variables on a 6-point scale

Note: Bold indicates variables that were reverse coded.
[a]Includes variables with coded skip logic.

Table A6: IPEDS Example from the 2011-12 Academic Year

Survey	Year Coverage	Table Name	TableTitle
Institutional Characteristics	Academic year 2011-12	HD2011	Directory information
Institutional Characteristics	Academic year 2011-12	FLAGS2011	Response status for all survey components
Institutional Characteristics	Academic year 2011-12	IC2011	Educational offerings, organization, admissions, services and athletic associations
Institutional Characteristics	Academic year 2011-12	IC2011_AY	Student charges for academic year programs
Institutional Characteristics	Academic year 2011-12	IC2011_PY	Student charges by program (vocational programs)
Institutional Characteristics	Academic year 2011-12	DRVIC2011	Frequently used derived variables (IC): Total cost of attendance and selectivity and admissions yield
Institutional Characteristics	Academic year 2011-12	IC2011MISSION	Mission statement
Institutional Characteristics	Academic year 2011-12	CustomCGids2011	Custom comparison groups
Fall Enrollment	Fall 2011	EF2011	Gender, attendance status, and level of student: Fall 2011
Fall Enrollment	Fall 2011	EF2011A	Race/ethnicity, gender, attendance status, and level of student: Fall 2011
Fall Enrollment	Fall 2011	EF2011B	Age category, gender, attendance status, and level of student: Fall 2011
Fall Enrollment	Fall 2011	EF2011C	Residence and migration of first-time freshman: Fall 2011(optional)
Fall Enrollment	Fall 2011	EF2011D	Total entering class, retention rates, and student-to-faculty ratio: Fall 2012
Fall Enrollment	Fall 2011	DRVEF2011	Frequently used derived variables (EF): Fall enrollment 2011
Completions	July 1, 2010 and June 30, 2011	C2011_A	Awards/degrees conferred by program (6-digit CIP code), award level, race/ethnicity, and gender: July 1, 2010 to June 30, 2011
Completions	July 1, 2010 and June 30, 2011	DRVC2011	Frequently used derived variables (C): Completions, July 1, 2010 to June 30, 2011
Finance	Fiscal year 2011	F1011_F1A	Public institutions - GASB 34/35: Fiscal year 2011
Finance	Fiscal year 2011	F1011_F2	Private not-for-profit institutions or Public institutions using FASB: Fiscal year 2011
Finance	Fiscal year 2011	F1011_F3	Private for-profit institutions: Fiscal year 2011
Finance	Fiscal year 2011	DRVF2011	Frequently used/derived variables Finance (F): Fiscal year 2011
Student Financial Aid	July 1, 2010 and June 30, 2011	SFA1011_P1	Student financial aid: 2010-11

Student Financial Aid	July 1, 2010 and June 30, 2011	SFA1011_P2	Student financial aid and net price: 2008-09, 2009-10, and 2010-11
Graduation Rates	Status of student as of August 31, 2011.	GR2011	Graduation rate data, 150% of normal time to complete - cohort year 2005 (4-year) and cohort year 2008 (2-year) institutions
Graduation Rates	Status of student as of August 31, 2011	GR2011_L2	Graduation rate data, 150% of normal time to complete - cohort year 2008 (less-than-2-year institutions)
Graduation Rates	Status of student as of August 31, 2011	GR200_11	Graduation rate data, 200% of normal time to complete - cohort year 2003 (4-year) and cohort year 2007 (less-than-4-year) institutions
Graduation Rates	Status of student as of August 31, 2011	DRVGR2011	Frequently used derived variables (GR) 150% of normal time to complete - cohort year 2005 (4-year) and cohort year 2008 (2-year) institutions
Human Resources	Fall 2011	EAP2011	Number of staff by occupational category, faculty and tenure status: Fall 2011
Human Resources	2011-12	SAL2011_A	Number and salary outlays for full-time nonmedical instructional staff, by gender, and academic rank: Academic year 2011-12
Human Resources	2011-12	SAL2011_FACULTY	Faculty status of full-time instructional staff in 4-year institutions, by contract length, gender, and academic rank: Academic year 2011-12
Human Resources	2011-12	SAL2011_A_LT9	Number of full-time instructional faculty with less than 9-month contracts, by gender and academic rank: Academic year 2011-12
Human Resources	Fall 2011	S2011_ABD	Full- and part-time staff by occupational category, race/ethnicity, and gender: Fall 2011
Human Resources	Fall 2011	S2011_F	Full-time instructional/research/public service staff, by faculty and tenure status, academic rank, race/ethnicity, and gender (Degree-granting institutions): Fall 2011
Human Resources	Fall 2011	S2011_G	New hires by occupational category, race/ethnicity, and gender (Degree-granting institutions): Fall 2011
Human Resources	Fall 2011	S2011_CN	Employees by primary occupation, race/ethnicity, and gender (Degree-granting institutions with less than 15 full-time employees and all nondegree-granting institutions): Fall 2011
Human Resources	Fall 2011	DRVHR2011	Frequently used/derived variables Human resources (HR): Fall 2011
12-month Enrollment	July 1, 2010 - June 30, 2011	EFFY2011	12-month unduplicated headcount: 2010-11
12-month Enrollment	July 1, 2010 - June 30, 2011	EFIA2011	12-month instructional activity: 2010-11
12-month Enrollment	July 1, 2010 - June 30, 2011	DRVEF122011	Frequently used derived variables (E12): 12-month enrollment, 2010-11

Table A7: CBII Alumni Survey IPEDS Linkage Results

Variable name	Frequency	Percent
col_first_lookup_00_ID		
All observations		
Missing in Survey Data	2,249	80.29
Non-missing in Survey Data	552	19.71
	2,801	**100.00**
Non-missing in Survey Data		
No longer in IPEDS[a]	31	5.62
Matched to IPEDS	521	94.38
	552	**100.00**
col_serious_lookup_01_ID		
All observations		
Missing in Survey Data	1,083	38.66
Non-missing in Survey Data	1,718	61.34
	2,801	**100.00**
Non-missing in Survey Data		
No longer in IPEDS[a]	41	2.39
Matched to IPEDS	1,677	97.61
	1,718	**100.00**
col_serious_lookup_02_ID		
All observations		
Missing in Survey Data	1,470	52.48
Non-missing in Survey Data	1,331	47.52
	2,801	**100.00**
Non-missing in Survey Data		
No longer in IPEDS[a]	36	2.70
Matched to IPEDS	1,295	97.30
	1,331	**100.00**
col_serious_lookup_03_ID		
All observations		
Missing in Survey Data	1,952	69.69
Non-missing in Survey Data	849	30.31
	2,801	**100.00**
Non-missing in Survey Data		
No longer in IPEDS[a]	12	1.41
Matched to IPEDS	837	98.59
	849	**100.00**

Note: IPEDS file used for merge is HD2019.
[a]Unit IDs that became defunct, or were absorbed into another, or new, system.

Table A8: CBII Administrative Data and Alumni Survey NaNDA Linkage Results

Variable name	Frequency	Percent
address_zip_derived		
All observations		
Missing in Administrative Data	662,007	50.46
Non-missing in Administrative Data	649,811	49.54
	1,311,818	**100.00**
Non-missing in Administrative Data		
Invalid or not in NaNDA	6,222	0.96
Matched to NaNDA	643,589	99.04
	649,811	**100.00**
zip_current		
All observations		
Missing in Survey Data	112	4.00
Non-missing in Survey Data	2,689	96.00
	2,801	**100.00**
Non-missing in Survey Data		
Invalid or not in NaNDA	9	0.33
Matched to NaNDA	2,680	99.67
	2,689	**100.00**

11.3 Survey Communications

Informed Consent

UNIVERSITY OF MICHIGAN
CONSENT TO PARTICIPATE IN SURVEY RESEARCH

Study title: College and Beyond II: Outcomes of a Liberal Arts Education
Principal Investigator: Paul N. Courant, Professor, University of Michigan
Study Sponsor: Andrew W. Mellon Foundation

PURPOSE OF THIS STUDY
The purpose of this study is to improve the college experience by learning about the long-term impacts of undergraduate education. The survey will ask questions about your life today including your work experiences, community involvement, and well-being. College graduates from the class of 2010 are being invited to participate.

INFORMATION ABOUT STUDY PARTICIPATION
Participation in the study is voluntary, and you may skip any questions or exit the survey at any time. The survey should take approximately 30 minutes to complete. You will receive $30 for completing the survey. If you choose to participate in this survey, your responses will be linked to data about you collected by [institution name] and other organizations. The purpose of doing so is to understand more about the experiences of college graduates. Appendix 1 [clickable link that opens additional window containing Appendix 1] describes the types of data we will collect and examples of that data.

**Note: this survey is intended for participants who reside in the United States. If you are currently outside of the U.S., please do not take the survey.

INFORMATION ABOUT STUDY RISKS AND BENEFITS
Because this study collects information about you, the primary risk of this research is a loss of confidentiality. We will minimize these risks by: 1) Separating information that can identify you such as your name from the research data, encrypting it, and storing it securely; 2) Using research data only in protected environments; and 3) Presenting study results so that individuals cannot be identified. While you may not receive any personal benefits from being in this study, society will benefit from the knowledge gained.

SHARING RESEARCH INFORMATION
We will store this study's research data at the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan. Before being made available to researchers, this data will be reviewed to make sure it cannot be used to identify you in any way. The data will be available for research use for at least fifty years.

CONTACT INFORMATION
Please contact the researcher listed below to obtain more information about the study or express a concern about the study.

Principal Investigator: Paul N. Courant, Ph.D.
Email: collegeandbeyond2study@umich.edu Phone: (877) 556-1542

If you have questions about your rights as a research participant, or wish to obtain information, ask questions or discuss any concerns about this study (HUM00173324) with someone other than the researcher(s), please contact the following:

University of Michigan
Health Sciences and Behavioral Sciences Institutional Review Board (IRB-HSBS)
2800 Plymouth Road
Building 520, Room 1169 Ann Arbor, MI 48109-2800
Telephone: 734-936-0933 or toll free (866) 936-0933
Fax: 734-936-1852 E-mail: irbhsbs@umich.edu

O Yes, I consent to participate in this survey. Click to enter
O No, I do not consent to participate in this survey. Click to exit.

Appendix 1

Types of data	**Examples of information in these records**
Postsecondary transcript data	Courses taken: course subject, level, credits attempted, credits earned, grade received, honors, standardized test scores.
Postsecondary learning management systems	Written work, assessments, time spent reviewing course material.
Postsecondary extramural engagement	Participation in campus housing, clubs, study abroad, internships.
Additional survey data	Responses from other surveys you have taken (e.g., CIRP Freshman survey).
Postsecondary education records from the National Student Clearinghouse	Educational enrollment, the beginning and ending date that a student is enrolled during each term, whether a student is enrolled full or part-time, private or public school, type of school or college enrolled in, student's major, whether a student has earned a degree, and date degree is earned.
Education records from K-12 and postsecondary institutions	Educational enrollment, educational attainment, achievement test scores, absenteeism, type of school enrolled in (example: high school, middle school), grades, application information, and graduation dates.
Earnings and employment	Quarterly earnings, employment, unemployment benefits, taxes paid by the employee and the employer(s) to the state, income from different sources, disability income, and employer(s).
Credit reports from credit bureaus	Information about loans, bankruptcies, credit card debt, unpaid medical bills, and credit scores. Sharing this information *will not* add a hard inquiry to your credit report and will not affect your credit.
Tax data	Number of dependents in the household, marital status, homeownership, income and sources of income, employers, taxes, receipt/eligibility of Earned Income Tax Credit, eligibility for other government programs (e.g.,: Medicaid).
Voting records	Information about election participation and voter registration (not who you voted for)

Invitation--Letter

Dear [FIRST NAME],

My name is Paul Courant and I am a professor at the University of Michigan. A team of researchers and I are conducting a large-scale study of the experiences of college graduates called College and Beyond II. You are being invited to participate as a graduate of [institution name]. We hope you will choose to participate and help us improve higher education.

The benefits of a college education to individuals and society are numerous and well-documented. Exactly *how* college transforms students' lives, and which aspects of the college experience are critical to this transformation, is less well understood. Please help us by participating in this online study. The survey will ask questions about your college experience and your life today related to work, family, and well-being. We want to know how you are doing.

The survey will take approximately 30 minutes to complete and your answers will remain completely confidential. You will receive a $30 check as a token of appreciation for completing the survey. This study is funded by The Andrew W. Mellon Foundation and is endorsed by the leadership of [institution name].

Please visit the following website to begin the survey: **[tinyURL LINK]**
Enter the following PIN when prompted: **[random 5-digit alphanumeric code]**

We realize that you have many demands on your time now more than ever. Thank you for contributing to knowledge about the impact of higher education.

Sincerely,
Paul N. Courant, Ph.D.
University of Michigan
collegeandbeyond2study@umich.edu
ph: (877) 556-1542

Invitation--Email

(Subject) College and Beyond Survey Invitation

Dear [FIRST NAME],

My name is Paul Courant and I am a professor at the University of Michigan. A team of researchers and I are conducting a large-scale study of the experiences of college graduates from across the nation called College and Beyond II. We hope you will choose to participate and help us improve higher education. By learning about your undergraduate experiences and life today, you will help college leaders understand how college impacts students' lives over the long-term.

The survey will ask questions about your life today related to work, family, and well-being. The survey will take approximately 30 minutes to complete and your answers will remain completely confidential. You will receive a $30 check as a token of appreciation for completing the survey. This study is funded by The Andrew W. Mellon Foundation and is endorsed by the leadership of [institution name].

[Click here to start the survey]

We realize that you have many demands on your time. Thank you for your contribution to this important research.

Sincerely,
Paul N. Courant, Ph.D.
University of Michigan

collegeandbeyond2study@umich.edu
ph: (877) 556-1542

Telephone Script

Hello, my name is [IWER NAME], calling from the University of Michigan Survey Research Center. May I speak with [FIRST/LAST NAME]?

You may remember receiving a letter and/or email from Professor Paul Courant at the University of Michigan. Professor Courant and a team of researchers are conducting a large-scale study of the experiences of college graduates and you were selected to participate.

You will receive a $30 check as a token of appreciation for completing the survey. This study is funded by The Andrew W. Mellon Foundation and is endorsed by the leadership of [institution name]. Your answers will remain completely confidential

We realize that you have many demands on your time. Thank you for your contribution to this project.

[IF NECESSARY] May we re-send you the link to the survey by email?

[IF NECESSARY] Can I confirm your email address as [EMAIL]?

[IF NECESSARY] If you have questions about the survey, please contact collegeandbeyond2study@umich.edu or 1-**877 556-1542**.

[IF NECESSARY] The benefits of a college education to individuals and society are numerous and well-documented. Exactly *how* college transforms students' lives, and which aspects of the college experience are critical to this transformation, is less well understood.

[IF NECESSARY] Please help us by participating in this online study. The survey will ask questions about your life today related to work, family, and well-being.

[IF NECESSARY] The survey will take approximately 35-45 minutes to complete and your answers will remain completely confidential.

[IF NECESSARY] Did you know that there is very little information available about what happens to college students after they graduate? A team of researchers here at the UM are conducting a study so that educational leaders can better understand how college impacts graduates lives. You can help by participating in the online study.

[IF NECESSARY] The survey will **close *soon* on [date].**

[IF NECESSARY] If you have already started the survey, you will be able to return to where you left off.

[IF CALL GOES TO VOICEMAIL] Hello this is [IWER FULL NAME] calling from the University of Michigan about the College and Beyond study. You might remember receiving an 8x11 packet in the mail or an email from (us/the Andrew Mellon Foundation) about this.

We can offer you $30 in appreciation for completing this survey on the web.
Please give us a call if you no longer have the letter or email, so we can provide you with access information to the secure survey.

You can call us back 7 days a week at the 734 University of Michigan# that appears on your caller ID - which is: 734-647-7757.

Thanks in advance for your participation!

Email Reminder 1

Subject: Your College and Beyond II survey

Dear [FIRST NAME],

Would you like to help college leaders improve the undergraduate experience at [name of institution]? I'm leading a team of researchers in conducting a study so that educational leaders can better understand how college impacts graduates lives. **You can help by participating in the online survey.** Here are the details:

- The survey asks for your reflections on your college experience and how you are faring today
- The survey will take approximately 30 minutes to complete
- Your answers will remain completely confidential
- You will receive a $30 check for completing the survey

Click here to enter the survey

If you have already started the survey, you will be able to return to where you left off.

Thank you for your time.

Sincerely,
Paul N. Courant, Ph.D.
University of Michigan

collegeandbeyond2study@umich.edu
www.icpsr.umich.edu/collegeandbeyond2
ph: (877) 556-1542

Email Reminder 2

Subject: Re: College & Beyond II graduates survey

Dear [FIRST NAME],

Recently, I sent you an invitation to participate in an important large-scale study of college graduates that is being sponsored by the Mellon Foundation. The study is called College and Beyond II and it is endorsed by the leadership of [institution name]. Many members of your graduating cohort have completed the survey; **We hope you will add *your* perspective too**.

You will receive $30 as a token of appreciation for completing the survey. The survey will take approximately 30 minutes to complete and your answers will remain completely confidential.

If you have already started the survey, you will be able to return to where you left off.

Click here to enter the survey

If you have questions about the survey, please contact collegeandbeyond2study@umich.edu or 1-**877 556-1542**.

Sincerely,
Paul N. Courant, Ph.D.
University of Michigan

Email Reminder 3

Subject: Your college experience: College and Beyond II study

Dear [FIRST NAME],

I'm writing to make sure you know about College and Beyond II, a large-scale study of college graduates from across the nation. You were randomly selected to participate.The information you provide in this survey will help college leaders better understand which aspects of the undergraduate experience are most important for long-term success.

If you haven't had a chance to complete the survey yet, I urge you to do so now. If you have already started the survey, you will be able to return to where you left off.

You will receive $30 as a token of appreciation for completing the survey. The survey will take approximately 30 minutes to complete and your answers will remain completely confidential.

Click here to start the survey

Sincerely,
Paul N. Courant, Ph.D.
University of Michigan

collegeandbeyond2study@umich.edu
www.icpsr.umich.edu/collegeandbeyond2
ph: (877) 556-1542

Email Reminder 4

Subject: Reminder: Your input is needed now!

Dear [FIRST NAME],

I'm getting in touch to make sure you know that the College and Beyond II study-- which you were selected to participate in-- closes soon! **If you haven't had a chance to complete the survey yet, I urge you to do so now.** If you have already started the survey, you will be able to return to where you left off.

You will receive $30 as a token of appreciation for completing the survey. The survey will take approximately 30 minutes to complete and your answers will remain completely confidential.

Click here to start the survey

If you have questions about the survey, please contact collegeandbeyond2study@umich.edu or (877) 556-1542.

Sincerely,
Paul N. Courant, Ph.D.
University of Michigan

Incentive Increase Email

Subject Line: **College and Beyond II Study, Still Time to Participate!**

Dear [FIRST NAME],

We are quickly approaching the end of the College and Beyond II study. As a graduate from [SCHOOL], you are an irreplaceable part of this research.

In fact, your participation is so important that we have increased the token of appreciation by $20, and are now offering you **$50** to complete the online survey.

We are interested in learning more about the ways an undergraduate education affects post-graduates further into adult life, directly from graduates like you.

Click here to enter the survey

Sincerely,

Paul Courant, PhD

University of Michigan
Collegeandbeyond2study@umich.edu
Ph: (877) 556-1542
More information about the study can be found
here: https://www.icpsr.umich.edu/web/pages/about/cbII/index.html

Text Messages

NO CONTACT [Reason to believe # does NOT belong to R]

Hello, [FIRST NAME ONLY] - we're hoping you can participate in this important study about your experience since graduation from [UNIVERSITY NAME].: [SURVEY LINK] We will send you a check for [**$30]** in appreciation. Please call us at the University of Michigan with any questions: [STUDY PHONE# OR IWER UM CELL#]. **Please let me know if this phone does NOT belong to [FIRST NAME ONLY].**

==========

SURVEY IN PROGRESS [Ph# confirmed]

Hello, [FIRST NAME ONLY] - thank you for starting the College and Beyond II survey about your experience since graduating from [UNIVERSITY NAME]. We hope you will complete this soon. [SURVEY LINK]. We will send you a check for [**$30]** in appreciation. Please call us with any questions at the Univ of Michigan : [STUDY PHONE# OR IWER UM CELL#].

==========

SURVEY IN PROGRESS [Ph# NOT confirmed]

Hello, [FIRST NAME ONLY] - thank you for starting the survey about your experience since graduating from [UNIVERSITY NAME]. We hope you will complete this soon. [SURVEY LINK]. We will send you a check for [**$30]** in appreciation. Please call us with any questions at the Univ of Michigan: [STUDY PHONE# OR IWER CELL#].

==========

REQUEST FOR LINK BY TEXT [Ph# confirmed]

Hello, [FIRST NAME ONLY] - this is [IWER NAME] following up on your request for the link to the College and Beyond II study [SURVEY LINK] : Thank you so much for agreeing to participate. We will send you a check for [**$30]** in appreciation. Please call us with any questions at the Univ of Michigan: [STUDY PHONE# OR IWER UM CELL#].

==========

FRIENDLY REMINDER [Prior Ph Contact, Ph# Confrmed:]

Hello again, [FIRST NAME ONLY] - thanks for agreeing to help with the College and Beyond II study. We know you are busy and hope you can carve out time to complete this survey. We will send you a check for [**$30]** in appreciation. We have just resent your secure survey link to your email address. Please call us with any questions at the Univ of Michigan: [STUDY PHONE# OR IWER UM CELL#].

==========

Email Reminder 5

Subject: Last chance: Survey closing on [date]!

Dear [FIRST NAME],

I hope you are doing well. I wanted to get in touch one last time to let you know the College and Beyond II study closes soon! If you haven't had a chance to complete the survey yet, I urge you to do so now**. Your participation is important for understanding the experiences of college graduates from all walks of life.** We want to hear *your* reflections on your college experience and how *you* are doing now.

Remember, you will receive $30 as a token of appreciation for completing the survey. The survey will take approximately 30 minutes to complete and your answers will remain completely confidential.

If you have already started the survey, you will be able to return to where you left off.

Click here to enter the survey

Sincerely,
Paul Courant, PhD
University of Michigan

collegeandbeyond2study@umich.edu
www.icpsr.umich.edu/collegeandbeyond2
ph: (877) 556-1542