Overview of the Global Longitudinal University Enrolment Dataset (GLUED)

Version 1.0 – December 2022

Overview and Rationale

The Global Longitudinal University Enrolment Dataset (GLUED) is a novel dataset that seeks to provide data on university enrolments around the world and over time. This dataset represents a significant new effort to capture institutional-level data in order to track sector-specific enrolment trends and better understand the rise of private higher education as a global phenomenon.

GLUED systematically compiles and/or estimates student enrolment data from roughly 17,000 universities in 194 countries and territories between 1950 and 2020. Drawing on the definitions used by the primary sources from which the dataset is compiled, its focus is on universities, defined as institutions that offer at least one Bachelor's, Master's, or Doctoral program (i.e., ISCED 6+). The dataset includes information on a set of institutional characteristics and interpolated and estimated data on student enrolments at five-year intervals.

The dataset is published under a Creative Commons licence 4.0. It is free to use and adapt with proper attribution. The suggested citation for the dataset is:

> Buckner, Elizabeth. 2022. "Global Longitudinal University Enrollment Dataset (GLUED)." https://doi.org/10.5683/SP3/P0D1KE, Borealis, V1.

More information on the rationale for the dataset can be found here:

> Buckner, E. (2022). The Global Longitudinal University Enrolment Dataset (GLUED). International Higher Education, (112), 9-11. https://ejournals.bc.edu/index.php/ihe/article/view/15729/11549

Variables and Codebook

country – Name of country. The default is the country name, merged in from the World Development Indicators. The name represents the country the institution is currently located in – many country borders have changed over time and so this does not necessarily represent the name of the country when founded or historically.

countrycode – Three-letter code outlined by ISO 3166 alpha-3 international standard.

region - Region in which the institution is currently situated, merged from the World Bank.

incomegroup – Income group, merged from the World Bank.

iau_id – This is a unique identifier for each university-level institution. The data comes from IAU's World Higher Education Database, which can be found here: https://whed.net/

iau_id1 – This is a unique identifier specific to our dataset, that is used to identify institutions that were once separate but later merged and today all have the same iau_id.

eng_name – Current name of university in English. Some of these names were translated from Google Translate and will need to be cleaned up in a future iteration.

orig_name – Current name of university in Latin characters, as reported to IAU-WHED.

foundedyr – The year in which each respective university recognizes as the year they were founded/established.

yrclosed – Year university closed, if closed and year known.

private - Binary variable for private = 1, or public = 0 ; this is scraped/coded from whed.net, self-reported by the university. We do not distinguish between non-profit or for-profit.

coordinates – GPS coordinates given in decimal degrees format (string).

latitude – GPS coordinate representing the institution's latitude. This data was scraped from Google Maps based on institutional name using a 100m zoom when possible. This may not be 100% accurate if an institution's location is not accurately linked in Google Maps to the name we use in GLUED.

longitude - GPS coordinate (decimal degree) representing the institution's latitude. This data was scraped from Google Maps based on institutional name using a 100m zoom when possible. This may not be 100% accurate if an institution's location is not accurately linked in Google Maps to the name we use in GLUED.

phd_granting – Binary variable for if WHED.net lists the university as offering a PhD degree or not. Data scraped from IAU-WHED in 2020.

m_granting – Binary variable for Masters degree granting or not. Data scraped from IAU-WHED in 2020.

b_granting – Binary variable for Bachelor's degree granting or not – GLUED aims to exclude all non-degree-granting institutions but there are some very specialized institutions that offer doctoral degrees but not bachelor programs.

divisions – Data coded from IAU-WHED based on number of separate divisions listed – it represents the number of different faculties the university has (Arts & Science, Business, Architecture, Education, etc.).

total_fields – This is the number of total degree programs offered by the university as listed on IAU-WHED – if a university offers a BA and an MA in Sociology, that would be counted twice.

unique_fields – This is the number of unique degree programs offered by the university as listed on IAU-WHED – so if a university offers a BA and an MA in Sociology, that would be counted only once in this variable. We include it to capture the extent of curricular breadth offered.

specialized – This is a binary variable we created, to mean if the institution is "specialized" – it is coded as 1 if the institution has a list of words in its name: Arts, Theater, Music, Drama, Institute, or if it has less than three unique programs.

merger – This is a binary variable for if the university today is the result of a merger in the past between multiple/other institutions.

noiau – This is a binary variable (0/1) to identify if the university does not have an IAU ID on WHED.

year – Calendar year

students5_interpolated – Raw or interpolated enrolment data input from original sources

students5_extrapolated – Extrapolated enrolment data (see section on Treatment of Missing Data)

students5_estimated - Estimated enrolment data – this variable includes raw, interpolated, or extrapolated values and in addition to that, fills in an estimated value for institutions that never reported any enrolment data (see section on Treatment of Missing Data)

Data Sources and Process of Dataset Construction

Data Sources

Data on student enrolments by institution were compiled from many sources: The Europa World of Learning (EWOL) and the International Association of Universities' World Higher Education Database (IAU-WHED), the United States' Integrated Postsecondary Education Data System (IPEDS) and, for some institutions, Wikipedia.

For all countries except the United States (U.S.), enrolment data spanning 1950-2015 were collected from two sources: hard copy volumes of The Europa World of Learning (EWOL) and the International Association of Universities' World Higher Education Database (IAU WHED). Both databases have collected basic information on higher education institutions (HEIs) around the world, including sector and enrolment for roughly the past 70 years. Each handbook lists institutional characteristics, including its country, name, sector (i.e., private or public) and founding date, as well as its total student enrolment and total number of faculty. The data is organized by country and new handbooks are published annually, although institutions do not report updated enrolments to the databases annually.

The IAU's WHED reports the year that each enrolment figure was last updated, and reports that each year, they target different institutions, so that enrolment figures are updated roughly every 6-7 years. Therefore, our dataset collected enrolment figures at five-year intervals, although the specific years available for each institution varied. We collected enrolment data from both sources and compared the values during data cleaning to check for consistency and reliability of the data.

Student enrolment data for the United States came from IPEDS data, which is collected by the National Center for Education Statistics. This data is publicly available and includes an exhaustive list of post-secondary enrolment at the institutional level starting in 1970. Enrolment rates for US institutions before 1960 are estimated.

To supplement existing sources of data on student enrolments, particularly for institutions that had no data available in IAU-WHED or EWOL, we also used data from Wikipedia on student enrolments. Wikipedia is a public resource that can be edited by the general public. This means that Wikipedia data may not be officially reported by institutions. Its inclusion may introduce errors in student enrolment data. Nonetheless, given the overriding goal of the dataset to capture as much data on institutions as possible, and the problem that many universities have never reported any student enrolment data to IAU-WHED or EWOL, we decided to add Wikipedia data. Wikipedia data was available for roughly 4,500 institutions. If a year was provided in Wikipedia, the enrolment data was merged at that year. If not, then the year 2022 was used for the Wikipedia data. We know that the 2022 data from Wikipedia may actually represent enrolment a few years earlier.

Data collection focused on bachelor's degree-granting institutions (i.e., ISCED 6+). We excluded institutions that were clearly secondary, vocational, and community college institutions throughout the time period. Some institutions were founded as vocational or technical institutions but gained university status at some point after founding are included, if those institutions are currently listed in the IAU's WHED. Data from the United States comes from IPEDS and some primarily vocational institutions may be included.

Data on institutional characteristics primarily come from the IAU's WHED. We default to the WHED whenever possible for data on founding year, sector, and number of faculties and programs. Data on the degrees offered and the number of programs offered represent only current data, and should not be applied or considered representative of the institution's past offerings.

Geographical location data was scraped from Google Maps based on university name in summer 2022.

Data Collection Process

The raw data for the dataset was compiled between May 2018-November 2020, during which research assistants (RAs) entered data from the physical volumes of EWOL and WHED since no comprehensive amalgamation of the volumes existed online at the time. Each RA entered data on total student enrolment, number of faculty, founding date, and sector for every university-level (i.e., ISCED 6+) institution by country in a master spreadsheet. Hardcopy volumes were used for the data spanning 1951–2015.

This hardcopy data was subsequently merged with data scraped from the WHED website for the most recent year available for all institutions in WHED. We could only use digital data from the WHED for the most recent period because no historical or panel data exists digitally.

In 2022, student enrolment data for all institutions in the dataset was scraped from Wikipedia. This was then checked, cleaned, and merged into GLUED.

Timeline of Dataset Construction

May 2018 – **September 2018** – Research assistants begin documenting enrolments, number of faculty, sector, and year of establishment for higher education institutions, by country, starting with the volumes of IAU-WHED. Each RA is assigned a volume to code (roughly 500 pages) and keeps a memo of their work with more information on specific institutions and their coding process. RAs work 3 hours/day 5 days a week in-person. Data entry was manual and involved both detailed attention as well as some subjective decision-making about whether to include an institution, and which enrolment figure to record. Therefore, coding was conducted in four or five-person teams, with opportunity feedback, questions, and discussion to maintain inter-coder reliability. RAs complete data entry of the IAU-WHED volumes collected roughly every five years from 1950-2015. Data cleaning, catching duplicate entries, and checking inconsistencies continues into the fall.

June 2019 – January 2020 - Research assistants begin documenting enrolments, sector, and year of establishment for higher education institutions, by country, using EWOL volumes starting in 1960.

January 2020 – May 2022 – ongoing revisions and data cleaning, catching duplicate entries, checking inconsistencies.

May 2019 – Research assistants begin documenting enrolments, faculty, sector, and year of establishment for higher education institutions by country starting with the volumes for EWOL following same process. Data entry is completed by October 2019.

September 2020 – Substantial data cleaning occurs to identify duplicate entries, errors in data entry and identify merged institutions, or inconsistencies in data reporting.

November 2021 - **November 2022** – The International Association of Universities (IAU) unveils a unique identifier for each institution in WHED. Therefore, to make future analyses possible, our RAs match IAUID's to the institutions in our dataset.

May - August 2022 – Geographic coordinates of each university was scraped and merged into the dataset.

June - November 2022 – Cleaning and refining dataset, and estimation of missing student enrolment data.

December 2022 – Version 1.0 of GLUED published online.

Concerns and Known Issues

This version of the dataset has many known issues that have complicated data collection and reporting. Many of these known issues affect a relatively small percentage of institutions, but nonetheless introduce possible errors and/or inconsistencies.

1) University Names and Incorrect Translations in English

GLUED provides the name of the dataset in English. For data cleaning and estimating missing values, it is important that names be standardized within IAU-ID. Therefore, although university names may have changed over time, we only report a single name, even if the university has had many names in the past. In addition, for many universities whose data was originally reported to IAU-WHED in a language other than English, we had to translate the university name to English for merging because some data sources were only reported in English and an English name was needed to merge data (i.e., when there is no IAU-ID). To translate university names into English, we used Google Translate. In some cases, English translations are incorrect and need cleaning. We hope to clean these in a future version of the dataset.

2) Concerns with Self-Reported Data

Student enrolment data is self-reported, not independently verified. Institutional definitions for enrolment may vary – we believe that most enrolment data, particularly in earlier decades, refers to persons or individuals enrolled. In recent years, some institutions may be reporting full-time equivalent (FTE) enrolments. In the earliest years (1950s), some institutions disaggregated their total enrolment numbers by gender, domestic or foreign students, and full or part-time status. In all cases, we calculated the total enrolment. Additionally, institutions do not disaggregate data by undergraduate and graduate students; therefore, we assume most reports include both undergraduate and graduate enrolment combined. The proportion of graduate enrolment has likely increased over time, and we cannot distinguish the two.

3) Significant Lags in Reporting

A second concern is that student enrolment figures reported to the sources from which we gathered data were often not up-to-date. This was suspected when if student numbers at the institutional level appeared to stay at the exact same number over two time points, which typically represented 10+ years, but then increased dramatically at a future time point. In order to address this, we identified these cases through the data cleaning and interpolation processes. Specifically, non-updated data points were removed from the dataset before interpolating missing values. That said, our estimation processes may artificially assume enrollment increases occur more gradually than they actually do.

4) Difficulties Defining the University

There are substantial cross-national differences in what institutions are considered universities. Our focus is on all institutions offering at least one BA, MA or doctoral degree. Institutions do not have to offer all three to be included – for example, we include liberal arts colleges in the United States and medical institutes only offering doctoral training in many countries. That said, this introduces complications for estimating student enrolments. GLUED includes enrolment at institutions that offer any ISCED 6+ degrees – not the number of students actually enrolled in those degree programs. This is an important distinction for some institutions that offer primarily vocationally-oriented diplomas and short courses, but also offer at least one BA program. As a result, the dataset includes some institutions that may not be called universities or considered universities by their national societies or regulatory bodies. This includes institutions that are primarily vocational but have recently begun to offer

Bachelor's degrees, such as the public college sector in Canada. We cannot correct for this issue because institutional enrollment data are almost always reported at the institutional level, with no disaggregation by program.

5) Difficulties Defining the Institution as our Unit of Analysis

Another concern with GLUED concerns the unit of analysis, which is the institution. In fact, we faced many difficulties defining the boundaries of "the institution." One issue was whether we should identify distinct branches or campuses as separate institutions or under the umbrella of a single institution. Our default decision was to defer to the unique identifiers produced by IAU (i.e., IAU-ID). This ID identifies branches of a larger university with a separate four-digit identifier after the IAU-ID of the main campus (IAU-######-####). In this model, each campus is considered a separate institution. However, we do not know when and how IAU makes the decision to distinguish between main institutions and branches.

Moreover, IAU's current WHED (Nov. 2022) does not include all branches and affiliated colleges of universities, in countries where universities have many affiliated colleges. This is more of an issue in some countries than others. In two very important countries and sources of private sector growth – China and India – it is possible or even required for private colleges to affiliate with public universities. In both countries, we are concerned that our data may not be accurately capturing enrolments in private colleges. For example, in India, private colleges affiliated with public institutions are classified as public in WHED/IAU, despite being considered for-profit and private by many local officials and academics. We hope to better classify institutions from China and India in subsequent iterations of data collection. Please note, that given our substantial concerns with the data on India, we suggest that all researchers conducting cross-national analyses on the data replicate models with and without the inclusion of India.

6) Mergers and University Splits

The names and organizational boundaries of many universities have changed significantly over time, which poses many challenges for tracking institutions over time. In our dataset, it was much more common for various smaller schools, colleges and universities to merge into larger ones. This posed numerous challenges for us in estimating enrolments. It was unlikely that we had regular enrolment data on all the constituent institutions pre-merge, making it difficult to estimate pre-merger enrolments.

To deal with merged institutions in a way that permits both tracking enrolments at the national level over time while also keeping historical data accurate, we decided to create a new variable: iau_id1, which includes an additional digit to keep track of the number of campuses or distinct universities that today all have the same IAU-ID. The interpolated data retain the original enrolment data for these institutions. However, for the purposes of analysis, in data cleaning, we retained only one estimate of enrolment for each institution, defaulting to the institution that exists in the present, which was determined by matching name and/or founding date. Our code sets estimated student enrolment (students5_estimated) for all other institutions that eventually merged into the present institution. This data likely not be accurate for pre-merger institutions at the institutional level, but we hope that it can best approximate the country's total enrolment in that era across the multiple institutions. Mergers are also identified in the dataset with the variable *merger,* which is a binary variable coded as 1 if the institution experienced a merger in the past.

Universities splitting into others was less common, but some universities were created as spin-offs from others. In these cases, we treated the university's founding date as the founding date of the new institution only. However, spin-offs may result in a steep drop in enrolments in the next time period. In some cases, we treated a university as "closed" – if all the pre-split university no longer exists and only smaller universities continued to exist. In the future, we hope to be able to better track histories of universities by tracking merger dates and split dates.

7) Changes in Institutional Status from Not-University to University

Some institutions were founded as vocational or technical institutions but gained university status at some point after founding are included, if those institutions are currently listed and have unique identifiers in the IAU's WHED. We decided to retain these institutions in the dataset because they are now classified as universities. For this reason, university enrolments in past decades may slightly over-estimate actual university enrolments at the time. We hope that in the future, we can better track changes in university status to identify when exactly they became universities.

8) Difficulties in Establishing a Founding Year or Closed Year

Founding dates for universities can be particularly difficult to establish, as universities are reorganized over time, created by merging other existing and older institutions or upgraded from secondary schools or technical colleges. Our practice is to default to the founding date listed in the World Higher Education Database or Wikipedia, if an institution is not in WHED. During data entry we found that the founding year in which each university was established was not always consistent between sources. For any inconsistencies, when the founding year was not available in WHED, the university website was checked to verify the founding year.

Our dataset to track university closures through the variable (yrclosed). However, as far as we know, there is no cross-national compilation of data on university closures. Because our data comes from compilations of existing universities, occasionally universities disappear from those records, with no explanation. It is difficult to ascertain if a university closed, or if it simply merged into another institution in earlier eras with no record today. Because we began tracking data in IAU's WHED in 1950, our raw data includes many universities that no longer exist under the same name. Many of these institutions have very generic names, such as "Institute of Education" – while we believe it is likely they subsequently were subsumed into more comprehensive universities, we do not know for sure. To deal with these institutions (<800), we manually researched each one to trace its history and determine if we could association it with a past or contemporary university. We drop universities from our dataset that may have existed at some point, but for which we have no closure date, no record of merger, no current institutional information and where there is no historical record of what happened to that institution available online. Our assumption is that these institutions merged into other universities or closed at some point, but we do not know when.

9) Institutional Outliers

Our dataset includes a number of Open and/or Distance universities that are very large. The two largest institutions in our dataset in the recent period is The Open University of China and Indira Gandhi National Open University. With a self-reported enrollment over 2 million students each in 2020, they account for roughly 1% of all student enrollments alone. The inclusion or exclusion of these influential institutions may affect some analyses.

10) Substantial Missing Data

A major concern with our dataset is that there is missing student enrolment data for many institutions. Missing enrolment data refers to both partial missing data (i.e., missing enrolment data in some years for institutions that have data at other time periods), as well as a large number of institutions that we know exist but for which we never have any student enrolment data. Many institutions exist in either WHED or EWOL, but do not report student enrolment data to those databases. We carried out various statistical analyses to estimate enrolments for institutions with missing data in order to create a dataset that captures the vast majority of university enrolments in a given periods, as described in the "Treatment of Missing Data" section.

Treatment of Missing Data

To fill in estimate enrolment data for those with missing values, we first compiled data from all sources (IAU-WHED, EWOL, Wikipedia and IPEDS) and conducted linear interpolations within known values. Compiling enrolment from IAU-WHED and EWOL did not always report the same or reasonably similar student enrolments. We are not sure exactly why the two sources vary; however, we believe they are related to either slightly different definitions of student enrolments or data collection processes, or lags in reporting. Because of these discrepancies between sources, we sought to avoid combining enrolment data from both IAU-WHED and EWOL – as we thought his might introduce more error in student enrolments that came from changing sources.

We then ran a series of commands to determine which raw data source was preferable for each institution. Selection of raw data source defaulted to the source with the most updated data and data over the longest period. IPEDS data was used in the United States. Wikipedia data was merged in to all institutions for which it was available. In general, the final dataset relies more on WHED data than EWOL, because we had access to more volumes of WHED and at more regular intervals. In some cases, when the two data sources created an unclear best source, based on one having more updated data points and another having a greater span of data, we took the means of both. In cases where data for a single institution was only available for one institution in early years and only the second data source in later years, we did combine the two data sources into a single span. Not doing so would have meant estimating student enrollments for years when some raw data was available. IAU WHED data was always preferred in India over EWOL, because IAU WHED data included all affiliated colleges, while EWOL did not. Not including affiliated colleges in India results in sometimes ten-fold differences in reported enrolments.

For all raw data, a five-year mean was calculated corresponding to the two years before and after the reported year (i.e., 2015 refers to the five-year mean of enrollments between 2013-2017). This was necessary to smooth out any changes in enrollments due to lags in reporting.

For institutions with some enrolment data and some years with missing data, we extrapolated from known data to estimate enrollments for missing years. To extrapolate, we classified institutions into three categories based on availability of raw data: institutions with 3+ datapoints, institutions with 2 datapoints and institutions with only 1 datapoint. This was necessary because extrapolation for each of these types of institutions posed distinctive challenges.

For institutions with 3+ interpolated mean datapoints: We conducted two types of longitudinal extrapolation and took the mean of the two values when possible. First, we used linear extrapolations for institutions with 3+ data points and 15 years or less of unknown values forward in time and/or backward in time. The choice of 15 years was subjective, but let us balance the goal of making use of some known data, while not forcing a linear model of extrapolation on university enrollments. We are quite certain from known data that university enrollments do not exhibit a linear growth model over a long period of time, but linear estimates may be reasonably close within short time periods.

The second extrapolation technique we used for institutions with 3+ datapoints was to calculate an institution-specific and national mean growth rate in student enrollments. We took the average of the institution and national mean growth rates over the time period, and used the growth rate to extrapolate enrollments back in time and forward in time.

For institutions with exactly 2 raw or interpolated values (~8,000 observations): Institutions with only two raw or interpolated student enrolment values posed a particular challenge because linear extrapolation from only two datapoints could be highly unreliable if extrapolated over larger spans of time. Institutions that decreased in size proved to be particularly difficult to estimate because a simple linear extrapolation could create estimates that seemed unrealistic, by creating very large institutions in earlier decades (1960s). For institutions with only two raw or interpolated values, we conducted the following missing-value estimations: 1) if institutions grew in size over time and there were less than 20 years of future values to estimate, we conducted a linear extrapolation. 2) We also created an annualized growth rate for each institution and an annualized growth rate for all universities in the country, which we combined to create an estimate of university growth rate by country and time period. A second estimation of missing values was conducted that applied this combined growth rate to known values to extrapolate missing values. The mean of these two estimated values was taken for all institutions with two values.

We did not have validated ways of modeling student enrolments in institutions with only two values and declining enrollments between those two values. Modeling enrolments in institutions with negative growth was difficult because in most countries in our dataset and over the six decades we track enrolments, universities' total enrolments have tended to grow. Declining enrolments are associated with conflict or demographic decline, but the global story is really one of expansion. As a result, it is not a reasonable assumption to model university growth rate as linear decline since founding, as this would assume a university started out "large" – which is not in line with realities that universities build reputations, infrastructure and enrolments over time. Therefore, in the small number of universities with only two datapoints and negative growth, we modified their institutional growth rates with two specific assumptions: 1) we assumed positive growth during the first 20 years after founding; and, 2) we moderated institutional declines over time period to make negative growth subtle.

For institutions with exactly 1 interpolated value: Many institutions had only a single value reported or interpolated from raw data. Of these, most were founded recently and had not been reporting to WHED or EWOL for very long. For these institutions, we carried out a similar estimation as for those with two values: we used the annualized growth rate for all universities in the country. We applied this annualized national growth rate to the single known value to extrapolate missing values.

Institutions with only one interpolated value that was reported over 20 yeas ago were dropped from the dataset. Through our extensive manual cleaning process, we have a strong reason to believe that these institutions were likely merged with other institutions at some point, but we have no way of tracking their merger or their closure. Because we are estimating enrolments for all institutions over time, it is important for GLUED to not include "ghost" institutions that no longer exist as separate institutions, as that makes it much more likely for us to over-estimate enrolments at the country level.

For institutions with 0 raw or interpolated values:

A large number of institutions had no raw data – these are institutions that we know exist, and most of which are found in IAU-WHED or on Wikipedia, but which have never reported enrolment data to any of the primary sources we used for compilation. To estimate missing values in these institutions, we ran a random effects panel regression analysis in Stata on known values with a set of covariates known to be strongly associated with student enrolment data, including: calendar year, logged country population, sector (public or private), PhD-granting, number of raw datapoints, country mean student enrollments, age of the university, university age squared, university age cubed, total number of degree programs in 2022 and a country-specific intercept. Various regression models were tested. This model explained 43% of variance (adjusted R-squared).

External Validation

Throughout the process of interpolation, extrapolation and estimation of missing values, we compared GLUED estimates to other sources of data on enrolments to evaluate our estimations. Figure 1 shows GLUED data in blue, the UNESCO Institute for Statistics (UIS) data on total worldwide enrollments in ISCED (2011) 5+ institutions in red, which include short programs and vocational education. We expect that GLUED data should be significantly less than UIS data on all tertiary enrolments, as we intentionally exclude ISCED 5 programs. The green line in Figure 1 shows UIS data for total enrolments, after subtracting UIS data on ISCED 5. Our goal is for GLUED to mirror the green line (UIS-ISCED6+) to the extent possible. Currently, we believe that GLUED may be over-estimating enrollments in 2000 and under-estimating enrolments in 2010. We are not sure why this is the case, and this may depend to some degree on which institutions are included in GLUED as compared to those included in national government reporting to UIS. The yellow line in the dataset represents the reported values published in Schofer and Meyer (2005)[1] – these data were very useful for earlier time periods. As the Figure shows, in the year 2000, Schofer and Meyer (2005)'s figure is nearly identical to that reported for all ISCED5+ in UIS. Ours is less because we focus specifically on university-level (ISCED 6+) enrollments.

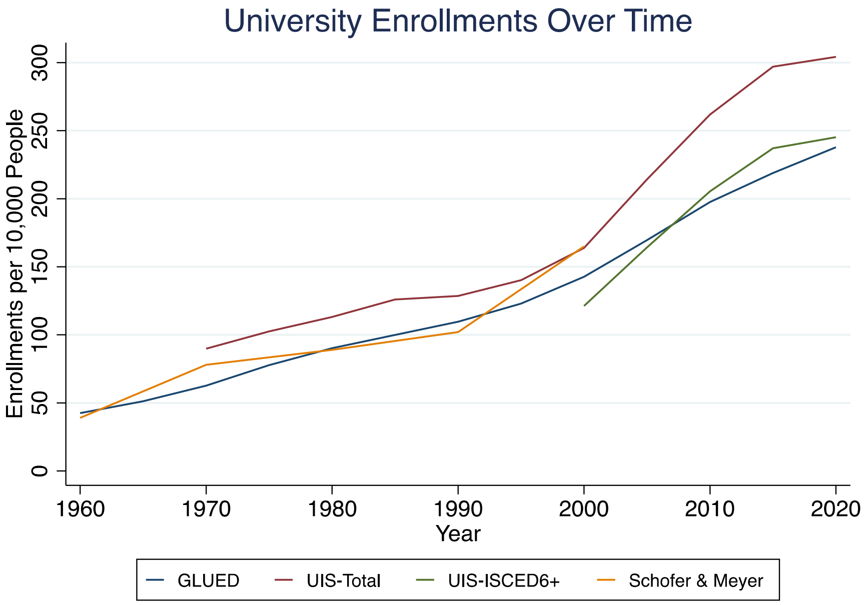

Figure 1: Comparing Data Sources on University Enrolments

[1] Schofer, E., & Meyer, J. W. (2005). The worldwide expansion of higher education in the twentieth century. *American Sociological Review*, *70*(6), 898-920.

Overview of the Country-Level University Enrolment Panel (CLUE)

Version 1.0 – December 2022

CLUE is a country-level panel dataset that seeks to estimate sector-specific university enrolments by country. All data in CLUE is aggregated from the institutional-level dataset, GLUED. The dataset spans 1950-2020 at five-year intervals, and includes 194 countries.

The dataset is published under a Creative Commons licence 4.0. It is free to use and adapt with proper attribution. The suggested citation for the dataset is:

> Buckner, Elizabeth. 2022. Country-Level University Enrolment Panel (CLUE)" https://doi.org/10.5683/SP3/AJGTC9 Borealis, V1.

Creation of CLUE

CLUE is a country-level panel dataset created from GLUED. To create this dataset, we aggregated institutional-level data to the country-level using Stata. The unit of analysis for CLUE is the country-year, and the dataset includes a set of basic country identifiers and data on universities, including the total number of universities, and estimated enrolments by sector.

Overview of Variables

countrycode – Three-letter code outlined by ISO 3166 alpha-3 international standard.

region - Region in which the institution is currently situated, merged from the World Bank.

incomegroup – Income group, merged from the World Bank.

year – Calendar year

privateunis – Total number of private universities

publicunis - Total number of public universities

totalunis - Total number of universities

privateenroll – Raw or interpolated sum of country-level enrolment in private universities

privateenroll_est – Country-level sum of total estimated enrolment in private universities

publicenroll – Raw or interpolated sum of country-level enrolment in public universities

publicenroll_est - Country-level sum of total estimated enrolment in public universities

percest_private – This variable reports the percent of private institutions in the country-year with an extrapolated or estimated value - we expect is positively associated with estimation error.

percest_public - This variable reports the percent of public institutions in the country-year with an extrapolated or estimated value - we expect is positively associated with estimation error.

Important Notes for Country-Level Data Analysis

1) For country-level analyses of worldwide enrolments in higher education over time, we suggesting using students5_estimated, as the interpolated and extrapolated values likely under-estimate total enrolments significantly.

2) Data on India and China likely under-report enrolments. Data from India, where the student enrolment figures for many of the country's large public universities include all students studying at both the university and its many affiliated colleges, may over-estimate enrollments in the public sector. This is because affiliated colleges are considered private because they charge tuition and have independent operations. This approach does not reflect the popular understanding of private and means that any analysis of universities in India may not reflect other percentages. We advise any cross-national analyses be tested for both the inclusion and exclusion of India.

Organization and Management

Research Team

The GLUED project is lead by Dr. Elizabeth Buckner, a professor in the Leadership, Higher and Adult Education (LHAE) Department at the Ontario Institute for Studies in Education (OISE), University of Toronto.

Special thanks go to the lead Research Assistant, who helped with all aspects of data entry and compilation: Ceara Khoramshahi, Leah Gibbins, who was responsible for institution-specific data cleaning, and Dmitriy Prokopchuk, who was responsible for scraping data from digital sources. Other research assistants involved in data entry and manual data cleaning and merging include: You Zhang, Shangcao Yuan, Margaret Paul, Grace Park, Sahra Nikkhah, Nathaniel Heayn, Mathuja Jayakumar, Thayani Jayadev, Sagal Ugas,Nadim Dabbous, Jerry Zhu and Maleeha Iqbal.

Dataset Headquarters

The dataset is housed at the University of Toronto Dataverse, a research data repository.

Funding

GLUED is supported in part by funding from the Social Sciences and Humanities Research Council of Canada and support from the University of Toronto through their University of Toronto Excellence Award (UTEA) program.

GLUED est en partie financé(e) par le Conseil de recherches en sciences humaines.