Matching Educational and Criminal Records at the Individual Level in Trinidad and Tobago

Methodology and Implementation

Diether Beuermann
Sabine Rieble-Aubourg
Tatiana Zarate-Barrera

Country Deparment
Caribbean Group
and Social Sector

TECHNICAL
NOTE Nº
IDB-TN-1170

December 2016

Matching Educational and Criminal Records at the Individual Level in Trinidad and Tobago

Methodology and Implementation

Diether Beuermann
Sabine Rieble-Aubourg
Tatiana Zarate-Barrera

December 2016

Inter-American Development Bank

Cataloging-in-Publication data provided by the Inter-American Development Bank
Felipe Herrera Library
Beuermann, Diether.
Matching educational and criminal records at the individual level in Trinidad and Tobago: methodology and implementation / Diether Beuermann, Tatiana G. Zarate-Barrera, Sabine Rieble-Aubourg.
p. cm. — (IDB Technical Note ; 1170)
Includes bibliographic references.
1. Student records-Trinidad and Tobago. 2. Criminal records-Trinidad and Tobago. 3. Education and state-Social aspects-Trinidad and Tobago. I. Zarate-Barrera, Tatiana G. II. Rieble-Aubourg, Sabine. III. Inter-American Development Bank. Country Department Caribbean Group. IV. Title. V. Series.
IDB-TN-1170

http://www.iadb.org

The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the Inter-American Development Bank, its Board of Directors, or the countries they represent.

CET@iadb.org

Diether Beuermann: dietherbe@iadb.org; Sabine Rieble-Aubourg: sabinea@iadb.org
Tatiana Zarate-Barrera: tatianaza@aidb.org

Abstract

This technical note standardizes and merges, at the individual level, several cohorts of academic evaluations undertaken during primary and secondary school in Trinidad and Tobago. In addition, it merges historical criminal records with educational databases. The resulting product is an individual-level panel dataset covering academic and criminal trajectories of citizens of Trinidad and Tobago from 1995 to 2015. This dataset is a powerful tool for evidence-based policymaking that is being used to investigate diverse issues in educational policy. This technical note describes the merging process of these databases, demonstrating the richness and practical usefulness of (sometimes under-exploited) administrative records to shape policy decisions.

JEL classifications: H4, I2
Keywords: Trinidad and Tobago, Administrative Records, Evidence-Based Policymaking

1. Introduction

Reliable and accurate information is essential for the design, monitoring, and evaluation of public programs. However, ensuring the availability, quality, and utilization of administrative records remains a challenge for policymakers, especially in developing countries. In particular, compared to North America and Europe, Latin American and Caribbean (LAC) countries lag behind in terms of organizing, sharing, and exploiting administrative data for monitoring and evaluation purposes. This fact poses a great challenge to governments, limiting the extent to which they can make evidence-based policy decisions.

While the observed administrative data gaps in LAC countries are the result of several factors, they are mainly due to the lack of well-functioning management information systems. This situation creates difficulties in accessing administrative information, particularly in the education sector. That is why tracking student performance over time has usually been stymied, despite the fact that, in most cases, data has been collected.

In the case of Trinidad and Tobago, there are four main examinations that regularly assess student performance at the end of primary and during secondary education: Secondary Entrance Assessment (SEA), National Certificate of Secondary Education (NCSE), Caribbean Secondary Education Certificate (CSEC), and Caribbean Advanced Proficiency Exam (CAPE). These examination results, which are available in databases at the individual level and include student cohorts across years, yield an important opportunity for the analysis of education policy in the country. However, the absence of unique student identifiers across examinations has created difficulties in tracking individual-level student achievement over time and has thwarted a proper analysis of educational outcomes and its determinants in Trinidad and Tobago.

To address this problem, we consolidated a database linking student records prior to secondary education (SEA) to national examination data taken whilst attending secondary school: NCSE, CSEC, and CAPE. To do this, we matched the examinations datasets at the individual level using the students’ personal information, such as name, last name, gender, and date of birth. To tackle the typical challenges of using non-numeric variables to identify students across examinations, we applied a matching methodology that helped us to combine student information in several ways, so that we could pair individual examination results even if there were spelling inconsistencies in their name, sex, or date of birth between datasets.

In addition, we cleaned up and consolidated Trinidad and Tobago’s Police records to match the offenders to the students who sat for the SEA. We followed a matching methodology similar to the one applied in the examination data.

This process yielded a final matched database that allows individual student performance to be tracked over time as well as criminal activity within a panel approach. This dataset is a powerful source to evaluate diverse public programs and educational issues, such as the effects of peers, different school administrative regimes, and single-sex vs coeducational schools on both educational and behavioral outcomes over time. This technical note, therefore, describes the process followed to build this harmonized dataset from administrative records.

The remainder of this document is as follows: Section 2 describes the education system in Trinidad and Tobago and the examination and criminal offence raw administrative data. Section 3 lays out the methodology applied to clean and match individual records across the different administrative datasets. Section 4 presents the main variables created from the harmonized dataset. Section 5 presents a discussion and conclusion.

2. The Education System in Trinidad and Tobago and the Data

2.1. The Education System

Trinidad and Tobago's education system has been influenced by the British model. This is particularly visible in its current examination system, which has emerged from the English Ordinary Level (O-Level) examinations. Primary and secondary public education has no cost for students, and school attendance is compulsory for all children aged 6–12. However, schooling begins for many children between the ages of 3 and 4 because it is expected that children have basic reading and writing skills when they start primary education.

There are eight school districts within Trinidad and Tobago and around 895 schools, including public, assisted, and private. The education system is divided into five levels: Pre-primary, primary, secondary, post-secondary, and tertiary education. Pre-primary education is guided by the Early Childhood Care and Education (ECCE) program, and it consists of two grades: first year and second year (Infant 1 and 2) targeting 3-4 year olds. Primary schooling for most children starts after they turn 5 years old in Standard 1 (Year 1 in the United Kingdom) and continues until Standard 5. The basic subjects included in the national curriculum are Mathematics, English, Science, Social Studies, Physical Education, and Geography. After completion of Standard 5, students must sit for the Secondary Entrance Assessment (SEA) to become eligible and obtain a placement in secondary school. This examination currently tests students in Mathematics, Language, Arts, and Creative writing. For details on the secondary school allocation process see Beuermann, Jackson and Sierra (2015) and Jackson (2010; 2012; 2013; 2016).

Secondary education begins in First Form (6th grade in the United States) and ends at Fifth Form following a standard academic curriculum. There are 150 secondary schools in Trinidad and Tobago, and 137 of them are government-funded schools, either in the form of public schools (fully funded and operated by the government), or government-assisted schools (funded by the government but managed by private bodies which are usually religious boards). For an assessment of relative academic effectiveness between these types of schools see Beuermann, Jackson and Sierra (2015). There are also private secondary schools, but they enroll less than 5 percent of the student population and mainly receive students who obtained low scores on the SEA examinations (Jackson, 2010; 2012; 2013).

Secondary school students sit for two national examinations that award certifications based on successful performance on these tests. At the end of Form 3, students have to sit for the National Certificate of Secondary Education (NCSE), in which students are awarded certification in the following eight subjects: Mathematics, Language Arts, Science, Social Studies, Visual and Performing Arts, Spanish, Technology Education, and Physical Education. Then, after five years of secondary school, students can take the Caribbean Secondary Education Certificate (CSEC). The CSEC examinations, administered by the Caribbean Examinations Council (CXC), include three proficiency schemes (Basic, General, and Technical) and are given in 31 subjects. Students obtain a certificate if they pass five or more subjects, including mathematics and English language. Obtaining a CSEC certificate is one of the basic requirements to be admitted to universities not only across the Caribbean, but also to universities in Canada, the United States, and the United Kingdom.[1]

Post-secondary education starts for students who have successfully completed their secondary education with a CSEC certificate and seek to continue their studies at the tertiary level. Students who choose this path have can stay at school for two additional years and sit for the Caribbean Advanced Proficiency Examinations (CAPE). These students have a greater chance of being admitted into tertiary education because this test is accepted not only in Trinidad and Tobago but also in recognized by selective colleges and universities in most nations. CAPE examinations evaluate academic, technical, and vocational skills and include, among others, the following subjects: Caribbean Studies, Communication Studies, Functional French, Functional Spanish, Information Technology, and Statistical Analysis.[2]

Students who want to attend local and regional institutions for tertiary education receive tuition subsidies through the Government Assistance for Tuition Expenses (GATE). In addition,

[1] See http://www.cxc.org/examinations/csec/ for more details on this examination.
[2] See http://www.cxc.org/examinations/cape/ for more details on this examination.

each year the Ministry of Education offers the Advanced Level Scholarship to students who have achieved academic excellence in the CAPE in ten subject groups: Business, Environmental Science, Languages, Mathematics, Modern Studies/Humanities, Natural Science, Technical Studies, Technological Studies, General Studies, and Visual and Performing Arts. People who are awarded this scholarship are allowed to study abroad with payments which contribute to the cost of tuition and compulsory fees, personal maintenance, and books, among others.

2.2 The Data

To track the students' performance through secondary education and enable long-term panel data analysis, we linked the SEA data with the NCSE, CSEC, and CAPE at individual level across various years. However, the original format of these databases does not lend itself to the identification of students across examinations because they do not have a unique individual identifier. Therefore, it was necessary to match the datasets at the individual level using student personal information such as last name, first name, middle name, date of birth, and gender.

The examination databases come from different sources: SEA, NCSE, CSEC, and CAPE (Table 1). The SEA data contains the test scores of 404,252 students from 1995 to 2012, their personal information (name, gender, date of birth, religion code), cohort, primary school district, and secondary school choices, as well as the secondary school to which the students were assigned by the Ministry of Education. In the case of the NCSE databases, they contain the test scores of 108,229 students from 2009 to 2015, their personal information (name, gender, and date of birth), cohort, exam results, and secondary school name. The CSEC databases include the test scores of 473,988 students from 1993 to 2015. The CAPE data contains the scores of 47,884 students from 2005–15. These examinations files contain information on students (name, gender, and date of birth), their grades on each CSEC/CAPE subject, and the secondary school where the students took the test.

The criminal offence data from Trinidad and Tobago was available from 1990 to 2015. It contained around 400,000 criminal records with a detailed description of the offence, including the date, the type of offence, and the chapter and section of the National Criminal Act, as well as the offender's name, gender, and date of birth.

3. Database Consolidation and Matching Process

3.1. Educational Records

3.1.1. Consolidation

The SEA and NCSE examination raw data were available in different file formats and grouped in various years (Table 2). The first step to consolidate these databases was to standardize the file formats and variables. We changed all the files to a standard format and as the databases in each group of years had a considerable number of differences across variables, we modified their attributes, such as name, type, length, and labels.

The second step to organize the data was to consolidate the available years into a single file. To do this, we appended all the years and searched for duplicate records in all of the examination data (SEA and NCSE), creating a student identifier with the full name, sex, and year of birth. We used different definitions of full name (Figure 1) to create this identifier, which later were also useful to match students between databases:

- **a.** Full name 1: uses the full name with no spaces between words
- **b.** Full name 2: uses all the components of a student's name but it only considers the middle name initial not the whole word (if the middle name is not the same as first name)
- **c.** Full name 3: uses the first and last name
- **d.** Full name 4: uses the last name, first name and two middle name initials

The CSEC and CAPE examination databases were also available in different file formats and grouped in various years (Table 3).

We consolidated these databases by standardizing the file formats and variables, applying a similar approach as in the previous ones. However, as each observation in these raw datasets corresponded to a subject taken by a student, we also had to identify all the subjects that a student took in any given year. Since there was no unique identifier for each student, the next step was to build an identifier comprising student's full name, sex, and year of birth. This identification process was developed in the following phases:

- **a.** Create a unique identifier to group the subjects taken by the student
- **b.** Standardize student information for a group of subjects
- **c.** Search for duplicate records in the database
- **d.** Reshape the database to create a new set of data with only one entry per student including all the subjects the student has taken as separate variables

The final output is a database where the unit of observation is the student, not the subject taken by the student, and each column contains information about each subject taken over one or several years (Figure 2).

In addition to the identification process described above, we also standardized the school names using the official school codes assigned by the MOE. However, as a student can sit for the CSEC more than once, the resulting file could also contain duplicate records. To address this situation, we selected the highest grade obtained by the student in each subject, the number of times the student tried until he/she achieved the highest grade, the year in which it was achieved, the first year in which the student took the exam, and the year in which the student passed the test.

3.1.2. Matching Process

There are several challenges when using non-numeric variables to identify students across examinations. In addition to the usual spelling inconsistencies that can be found on students' full names, in the SEA and NCSE examinations, students have only provided their middle name initials, while in the CSEC and CAPE they have provided their complete middle names. On the other hand, we also found problems with the dates of birth, such as differences between months and days of birth, either because they were swapped or had obvious typos.

To overcome those challenges, a matching methodology was developed in two phases: direct merge and fuzzy matching. The direct merge phase prepared the individual databases and created the main matching results using conservative routines: (i) exact matches, (ii) exact matches swapping student first and last names, (ii) exact matches swapping student month and day of birth, and (iv) exact matches swapping students' first and last names and month and day of birth (Table 4).

The "fuzzy matching" phase used a customized algorithm that combines the databases in different ways, changing the order of variables and looking for the best matches. The procedures applied in this phase changed the following variables in order to find matches between examination databases:

- **a.** Match for the date of birth and sex
- **b.** Match for the date of birth and sex when month and day of birth were swapped
- **c.** Match only for the date of birth
- **d.** Match only for the date of birth when month and day of birth were swapped
- **e.** Match for year and month of birth and the first name initial

The quality of the matches, because of the previous changes in the variables, was assessed using the Levenshtein distance. This concept is defined as the minimum number of changes that are required to change one word into another. Then, according to this definition, the best match will be the one with the minimum Levenshtein distance between names.

The following step was to check if a student was matched more than once; if so, the best match among those was kept. This procedure was applied until there were no repeated observations as best matches. In addition, we defined a cutoff "k" to classify how poor the matches can be. Therefore, only matched students with Levenshtein distance of less than "k" were considered good matches.

The final step was to use an additional algorithm to seek extra matches within the unmatched students. This algorithm takes the first position of every letter in each arrangement of letters and assigns to it a number. For instance, in "ab" the position of "a" is 1 and the position of "b" is 2. It compares each observation between examination datasets and calculates a score based on the sum of the squared difference of positions.[3] At the end, it keeps the best match that is the one with the minimum score.

We also created additional variables to assess the quality of the new matches. One of them was generated to identify the type of match. It uses information from the previous score (sum of the square difference of the positions), the Levenshtein distance between names, and whether there were matches on the date of birth. This variable can take five values:

a. **One**: if there is a match for the date of birth and the score is less than 10 or the individual Levenshtein distances for both last and first names are at most 4
b. **Two**: if there is a match for the date of birth and the Levenshtein distance for the last name of the student is at most 2
c. **Three**: if there is a match for the date of birth and at least one of the first or last names has a Levenshtein distance of at most 1
d. **Four**: if the score is less than 5 or the Levenshtein distances for the last and first names of the student are both at most 1
e. **Five**: Anything else

We also created a variable to identify the students matched in each phase (either direct merge or fuzzy matching).[4] In addition, we created another indicator to identify the cases within each of the phases.

[3] In this case for example, "ab" vs "ac" would have a score of $(1-1)^2+(2-0)^2+(0-2)^2=8$.
[4] Binary variable.

Finally, to increase the match rate, we created some manual rules to seek additional matched students who were not addressed by the previous phases. We only applied these manual revisions to the matched students from phase two or fuzzy.

3.2. Criminal Records

3.2.1. Consolidation

The raw criminal offence data contains 474,304 records from 1990 to 2015. The dataset includes information about the offender's personal data, offence type, offence date, and sentence (if any). It also includes the official chapter and section of the Criminal Offences Act of Trinidad and Tobago and a short description of the offence. However, these offences were not standardized because they were entered as text. Therefore, we first created categories for each type of offence.

To classify criminal offences, we explored the data to find the most common felonies. According to our analysis, we defined the following categories: (i) Drugs, (ii) Non-violent Sexual, (iii) Violent Crime, (iv) Property Crime, (v) Illegal Possession of Weapons, (vi) Driving Offences, (vii) Kidnapping, (viii) Resisting Arrest, (ix) Abusive Language, (x) Selling without Official Permission, and (xi) Others. To group the offences in line with the previous classification, we searched and compared the chapter and section from the Criminal Act to the description of the offence and created a new variable to classify the crime.

Similarly, we followed the previous routine to classify the sentences in categories and we defined eight groups: (i) Dismissed, (ii) Discharged, (iii) Hard Labor, (iv) Community Service, (v) Payment (Fine, Bail, Bond, etc.), (vi) Prison, (vii) Amnesty, (viii) Sentenced to Death, and (ix) Other.

Furthermore, as each observation in the raw data corresponds to an offence committed by a citizen, we had to identify all the crimes that a person committed over the years. Since there was no unique identifier for each individual, the next step was to build an identifier composed of the person's full name, sex, and year of birth. This identification process was developed in the following phases:

a. Create a unique identifier to group the offences committed by each person

b. Standardize the offender's personal data by group of offences

c. Search for duplicate records in the dataset

d. Reshape the database to create a new set of data with only one entry per individual including all the offences the person has committed.

The final output is a database where the unit of observation is the individual, not the offence committed, and each column contains information of each crime with its date and sentence (Figure 3).

3.2.2. Matching Process

To find the students who committed one or multiple offences, we followed a similar matching routine to the one used on the education data but between the SEA examination data and the criminal records. We used personal information such as name, last name, date of birth, and gender to pair the students with their offences (if any). This matching methodology was also applied using the direct merge and fuzzy matching phases as in the previous cases.

4. Summary Statistics

4.1. Educational Records

The SEA data (1995–2012) is linked to the official NCSE (2009–15), CSEC (2000–15), and CAPE (2005–15) examinations. As the NCSE, CSEC, and CAPE examinations are usually taken three, five, and six years after the SEA, respectively, a great number of students from the SEA 2012 cohort have sat for the NCSE[5] by 2015 but not yet for the CSEC, and CAPE. Thus, we were not able to match many students who took the SEA after 2010 with the CSEC[6] and CAPE[7] records. Similarly, the match rate between SEA and NCSE or between SEA and CAPE was low[8] when the students took the SEA before 2006 for NCSE or before 1999 for CAPE.

We estimated the match rate per year of birth between examinations data and compared it with the theoretical rate, that is, the match rate that we would obtain if all the students in the available data were matched. Table 5 summarizes the match rates. As the table shows, the obtained match rates are very close to the maximum theoretical match rates. This evidences that the matching algorithms work adequately.

Descriptive statistics are presented in Table 6. For each database, we report the mean, median, range, and number of test takers available in the row data. As expected from standardized tests, in most of the cases the mean score per component corresponds to its

[5] Match rate SEA-NCSE: 84 percent
[6] Match rate SEA-CSEC after 2010: 3 percent
[7] Match rate SEA-CAPE after 2010: 0 percent
[8] Match rate SEA-NCSE before 2006: 0 percent. Match rate SEA-CAPE before 1999: 1.5 percent.

median score. However, the results of the tests are not perfectly comparable across datasets because of differences in either the methodology of examination or the grading scales established over the years.

SEA data covers cohorts who took the examination between 1995 and 2012, a total of 404,252 individuals with a balanced gender profile. The NCSE examination began in 2009. Therefore, only cohorts from 2009 to 2015 are covered with scores reported in standard deviations.[9] There are two facts from the NCSE that are worth highlighting. First, a relatively lower number of students took the components of Physical Education, Arts, and Technical Skills. Second, we observe extreme positive outliers in Science, Social Studies, and Technical Skills that were 2.6, 2.4, and 3.4 standard deviations above their means.

It is apparent that the sex composition favors women in the more advanced examinations (CSEC and CAPE), while gender is balanced among SEA and NCSE takers. This reflects the well-known relative academic underperformance of men in the Caribbean. On average, CSEC test takers write six different subjects. However, the average number of subjects passed is 3.89. Note that obtaining a CSEC certificate requires passing five different subjects, including English and math. English average passing rate is 63 percent, while math is only 50 percent. Overall, 42 percent of CSEC test takers between years 2000 and 2015 obtained a CSEC certificate.

CAPE takers have a clear female majority, accounting for 62 percent. However, on average, only one subject per test taker achieves the top mark. This is important as the criteria for obtaining a scholarship for tertiary studies demand obtaining top marks on at least eight subjects, including the core subjects of Communication Studies and Caribbean Studies. In fact, only 4 percent of the 47,198 CAPE takers between 2010 and 2015 met the required criteria to obtain a scholarship.

4.2. Criminal Records

Criminal records (1990–2015) were linked to the official SEA (1995–2012) examination records. Criminal records include all individual offences registered by the Trinidad and Tobago Police Service. The SEA databases include people who were mainly born between 1983 and 2000 given that the SEA is taken at the end of primary school when students are around 11 or 12 years old.

[9] Raw scores have been standardized by subtracting the mean and dividing by the standard deviation of each subject-year distribution. This results in a distribution with zero mean and standard deviation equal to unity for each subject-year.

We estimated the match rate per year of birth and per gender between the crime and SEA databases, and we compared it with the theoretical rate; that is the match rate that we would obtain if all students in the available data were matched. Table 7 summarizes the match rates.

Descriptive statistics are presented in Table 8. We found that 29,259 students (out of 404,252) who took the SEA have committed one or multiple offences, with 86 percent of them being males. Approximately 14 percent of the offenders had violated the law by the age of 16.

5. Conclusions

During the last decades, the Ministry of Education of Trinidad and Tobago has made a great effort to evaluate student performance through the application of national examinations at different stages of primary and secondary education. This effort has accumulated rich and diverse databases of students' achievement over time. However, these databases did not lend themselves to analysis. Therefore, they were stand-alone databases that were not being used to inform the design or evaluate education policy within the country.

The absence of a unique student identifier across examinations has stymied the analysis of these databases because there is no straightforward technique to pair students between tests. Nevertheless, there have been some studies that have tried to evaluate different characteristics of the education system of Trinidad and Tobago using samples of the examination databases.

One example is Jackson (2016), who identified the causal effect of single-sex schooling on student attainment in national examinations using administrative SEA (2007–15), NCSE (2009–15), and CSEC (2012–15) data. According to the results, single-sex education can improve boys' and girls' outcomes and can increase their chances of completing secondary education and of earning the requirements to continue to tertiary level. Similarly, in another study using similar data, Beuermann, Jackson, and Sierra (2015) estimated the effects on academic outcomes of attending privately managed public secondary schools (assisted schools) relative to traditional public secondary schools in Trinidad and Tobago. They found little evidence of any relative benefit in attending an assisted school between ages 10 and 15 in terms of dropout rates or examination performance at the age of 15.

The Ministry of Education has worked on organizing its existing administrative records with the objective of improving the MOE's Sector Management capabilities. They have compiled raw datasets on examination results, and we have applied a matching methodology to pair students between examinations and years. It has helped us to consolidate a final database that is useful to track students who took the SEA across different examinations during secondary schools. As a result, it will be the first time the Ministry has the chance to perform long-term data analysis on their students' achievements.

References

Beuermann, D. W., C. K. Jackson and R. Sierra. 2015. “Privately Managed Public Secondary Schools and Academic Achievement in Trinidad and Tobago: Evidence from rule-based student assignments.” IDB-WP-637. Washington, DC: Inter-American Development Bank.

Jackson, C. K. 2010. "Do Students Benefit from Attending Better Schools? Evidence from Rule-based Student Assignments in Trinidad and Tobago." *Economic Journal, Royal Economic Society* 120(549): 1399–1429.

_____. 2012. "Single-sex Schools, Student Achievement, and Course Selection: Evidence from Rule-based Student Assignments in Trinidad and Tobago." *Journal of Public Economics* 96(1): 173–187.

_____. 2013. "Can Higher-achieving Peers Explain the Benefits of Attending Selective Schools? Evidence from Trinidad and Tobago." *Journal of Public Economics* 108: 63–77.

_____. 2016. "The Effect of Single-Sex Education on Academic Outcomes and Crime: Fresh Evidence from Low-Performing Schools in Trinidad and Tobago." NBER Working Paper No. 22222. Cambridge, MA: National Bureau of Economic Research.

Figure 1. Full Name Definitions

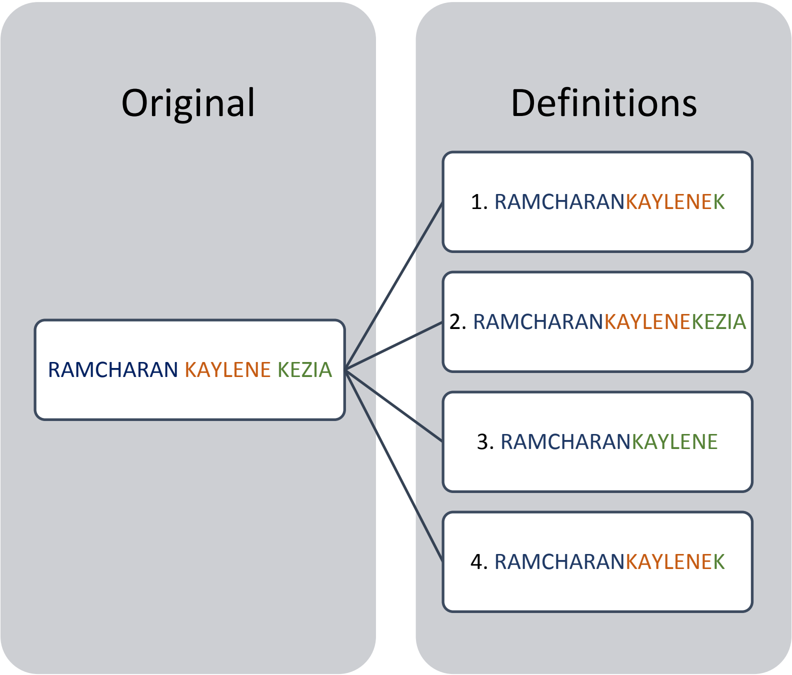
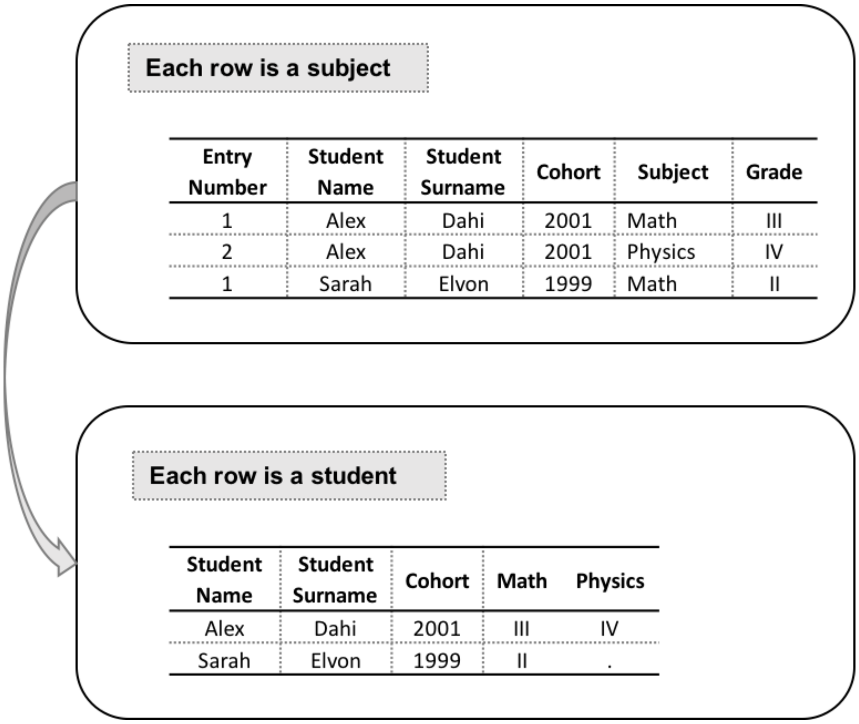


Figure 2. Reshaped Examination Data



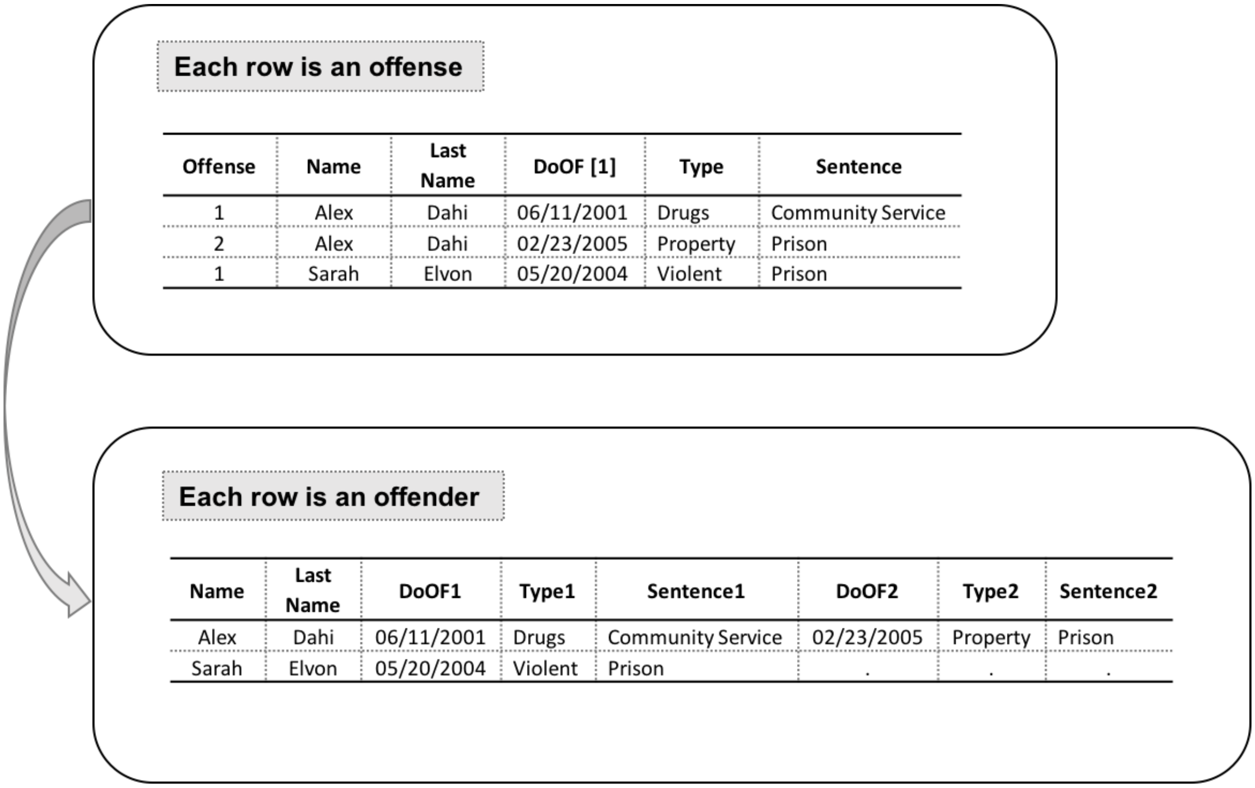
Each row is a subject

Entry Number	Student Name	Student Surname	Cohort	Subject	Grade
1	Alex	Dahi	2001	Math	III
2	Alex	Dahi	2001	Physics	IV
1	Sarah	Elvon	1999	Math	II

Each row is a student

Student Name	Student Surname	Cohort	Math	Physics
Alex	Dahi	2001	III	IV
Sarah	Elvon	1999	II	.

Figure 3. Reshaped Crime Dataset



Each row is an offense

Offense	Name	Last Name	DoOF [1]	Type	Sentence
1	Alex	Dahi	06/11/2001	Drugs	Community Service
2	Alex	Dahi	02/23/2005	Property	Prison
1	Sarah	Elvon	05/20/2004	Violent	Prison

Each row is an offender

Name	Last Name	DoOF1	Type1	Sentence1	DoOF2	Type2	Sentence2
Alex	Dahi	06/11/2001	Drugs	Community Service	02/23/2005	Property	Prison
Sarah	Elvon	05/20/2004	Violent	Prison	.	.	.

[1] Date of Offense

Table 1. Raw Datasets

	SEA	NCSE	CSEC	CAPE
Years	1995–2012	2009–2015	1993–2015	2005–2015
Observations	404,252	108,229	473,988	47,844
Variables	Last name First name Middle names initials Sex Date of Birth Cohort Subjects' grades School name	Last name First name Middle names initials Sex Date of Birth Cohort Subjects' grades School name	Last name First name Middle names Sex Date of Birth Cohort Term Subjects' grades School name	Last name First name Middle names Sex Date of Birth Cohort Subjects' grades School name

Table 2. SEA and NCSE Datasets: Formats and Years

Examination	Years	Format
SEA		
	1995–2001	Excel (.xls)
	2002–2009	Stata (.dta)
	2010–2012	Excel (.xls)
NCSE		
	2009–2014	Stata (.dta)
	2015	CSV (.csv)

Table 3. CSEC and CAPE Datasets: Formats and Years

	Years	Format	Period
CSEC			
	2005 – 2009	Text file (.txt)	January
	1995 – 2006	Stata (.dta)	June
	2011 – 2012	Stata (.dta)	June
	2015	Stata (.dta)	June
	2007 – 2010	Excel (.xlsx)	June
	2013 – 2014	Excel (.xlslx)	June
CAPE			
	2005 – 2009	Text file (.txt)	
	2014 – 2015	Text file (.txt)	
	2010 – 2012	Acces (.acccdb)	
	2013	Excel (.xlsx)	

Table 4. Direct Merge Phase: Cases

Case	SEA Name	CSEC Name	SEA Sex	CSEC Sex	SEA Year	CSEC Year	SEA Month	CSEC Month	SEA Day	CSEC Month
ii	**ALEXANDER AKINS**	**AKINS ALEXANDER**	M	M	1996	1996	9	9	27	27
iii	ALBERT SABRINA	ALBERT SABRINA	F	F	1989	1989	**11**	**9**	**9**	**11**
iv	**ALEXANDER SHENICE**	**SHENICE ALEXANDER**	F	F	1995	1966	**4**	**9**	**9**	**4**

Table 5. Match Rates by Year of Birth

Match	Year of Birth	Theoretical Match	Current Match
SEA-NCSE	1992	8.20%	6.80%
	1993	29.60%	27.00%
	1994	77.40%	77.60%
	1995	83.90%	85.90%
	1996	86.30%	87.00%
	1997	87.90%	88.50%
	1998	88.10%	89.50%
	1999	90.50%	90.60%
	2000	92.00%	89.40%
SEA-CSEC	1988	78.8%	76.5%
	1989	82.2%	76.6%
	1990	83.2%	78.0%
	1991	82.7%	79.1%
	1992	83.4%	80.2%
	1993	84.1%	81.2%
	1994	81.9%	80.8%
	1995	80.7%	80.6%
	1996	80.4%	80.1%
	1997	75.7%	74.8%
	1998	56.1%	54.7%
	1999	9.9%	9.4%
	2000	1.9%	1.5%
SEA-CAPE	1988	17.41%	17.53%
	1989	17.96%	17.79%
	1990	19.82%	19.54%
	1991	20.75%	20.70%

1992	21.82%	21.85%
1993	22.92%	22.75%
1994	23.54%	22.32%
1995	22.99%	22.04%
1996	22.16%	22.04%
1997	19.23%	19.02%
1998	3.30%	3.21%
1999	0.05%	0.05%

Table 6. Summary Statistics: Examinations

Variable	N	Mean	Std. Dev.	Median	Min	Max
SEA – Female	404,252	0.51	0.5	1	0	1
NCSE – Female	107,292	0.51	0.5	1	0	1
NCSE - Physical Education	98,754	0.01	1	0	-1.91	1.84
NCSE - English	107,173	0.02	0.99	0	-1.88	1.88
NCSE – Spanish	106,239	0.01	0.99	0	-1.66	2.03
NCSE - Science	106,987	0.01	0.99	-0	-1.98	2.66
NCSE - Social Studies	105,539	0.02	0.99	0	-1.97	2.44
NCSE - Math	107,155	0.02	0.99	-0	-1.91	1.97
NCSE – Arts	92,247	0.01	1.00	0	-1.67	1.96
NCSE – Technical	74,282	0.00	1.00	-0	-1.53	3.45
CSEC – Female	278,742	0.55	0.50	1	0.00	4.00
CSEC - Number of subjects written	278,742	6.25	2.04	7	0.00	17.00
CSEC - Subjects ranked in I, II or II [1]	278,742	3.89	2.91	4	0.00	16.00
CSEC - English ranked in I, II or III	278,742	0.63	0.48	1	0.00	1.00
CSEC - Math ranked in I, II or II	278,742	0.50	0.50	0	0.00	1.00
CSEC - Certificate [2]	278,742	0.42	0.49	0	0.00	1.00
CAPE – Female	47,198	0.62	0.49	1	0.00	1.00
CAPE - Subjects ranked in I - grouped by Unit	47,198	1.10	2.13	0	0.00	14.00
CAPE - Scholarship	47,198	0.04	0.19	0	0.00	1.00

Notes: [1] III is the minimum passing grade; [2] A certificate is obtained if the student passes five subjects including Math and English

Table 7. Match Rate Crime

Year of Birth	**Total**		**Female**		**Male**	
	Theoretical Match	*Current Match*	*Theoretical Match*	*Current Match*	*Theoretical Match*	*Current Match*
1983	16.7%	11.0%	4.4%	3.0%	29.1%	19.3%
1984	13.7%	10.5%	3.7%	3.0%	24.3%	18.5%
1985	12.6%	10.4%	3.5%	3.0%	22.0%	18.0%
1986	11.9%	10.3%	3.1%	2.8%	21.3%	18.2%
1987	11.4%	9.3%	3.1%	2.6%	20.3%	16.6%
1988	12.6%	9.4%	3.4%	2.8%	21.8%	16.0%
1989	11.9%	8.5%	3.2%	2.3%	20.6%	14.8%
1990	11.1%	8.4%	2.7%	2.2%	19.5%	14.8%
1991	9.9%	7.6%	2.6%	2.1%	17.1%	13.2%
1992	8.8%	6.7%	2.3%	1.8%	15.5%	11.7%
1993	7.2%	5.5%	1.9%	1.6%	12.5%	9.4%
1994	5.8%	4.8%	1.5%	1.2%	10.0%	8.3%
1995	4.4%	3.6%	1.1%	0.9%	7.6%	6.1%
1996	3.1%	2.5%	0.8%	0.5%	5.2%	4.3%
1997	2.0%	1.6%	0.7%	0.5%	3.3%	2.7%
1998	1.2%	0.7%	0.4%	0.2%	1.9%	1.2%
1999	0.8%	0.4%	0.3%	0.1%	1.3%	0.7%
2000	0.4%	0.1%	0.1%	0.1%	0.7%	0.2%

Table 8. Summary Statistics: Criminal Offences

		N	Mean	Std. Dev.	Median	Min	Max
Total	Arrested	404,252	0.07	0.26	0	0	1
	Female	29,259	0.14	0.35	0	0.00	1.00
	Arrested at the age of 15	29,255	0.07	0.25	0	0.00	1.00
	Arrested at the age of 16	29,255	0.14	0.34	0	0.00	1.00
	Total number of arrests	29,259	2.86	4.47	2	1.00	395.00
Female	Arrested at the age of 15	4,111	0.07	0.26	0	0.00	1.00
	Arrested at the age of 16	4,111	0.13	0.33	0	0.00	1.00
	Total number of arrest	4,112	2.20	8.03	1	1.00	395.00
Male	Arrested at the age of 15	25,144	0.07	0.25	0	0.00	1.00
	Arrested at the age of 16	25,144	0.14	0.34	0	0.00	1.00
	Total number of arrest	25,147	2.97	3.55	2	1.00	77.00