**Memorandum**

**United States Department of Education**
**Institute of Education Sciences**
**National Center for Education Statistics**

DATE: April 26, 2017

TO: Robert Sivinski and E. Ann Carson, OMB

THROUGH: Kashka Kubzdela, OMB Liaison, NCES

FROM: David Richards, BPS:12/17 Project Officer, NCES
Tracy Hunt-White, Team Lead, Postsecondary Longitudinal and Sample Surveys, NCES

SUBJECT: 2012/17 Beginning Postsecondary Students Longitudinal Study (BPS:12/17) OFAC Compliance Change Request (OMB# 1850-0631 v.14).

The 2012/17 Beginning Postsecondary Students Longitudinal Study (BPS:12/17) is conducted by the National Center for Education Statistics (NCES), within the U.S. Department of Education (ED). BPS is designed to follow a cohort of students who enroll in postsecondary education for the first time during the same academic year, irrespective of the date of high school completion. Data from BPS are used to help researchers and policymakers better understand how financial aid influences persistence and completion, what percentages of students complete various degree programs, what are the early employment and wage outcomes for certificate and degree attainers, and why students leave school. The request to conduct the BPS:12/17 full-scale data collection was approved by OMB in December 2016 (1850-0631 v.10) with the latest revisions approved in April 2017 (OMB# 1850-0631 v.11-13).

This request is to modify the BPS:12/17 sample and incentive payment options to comply with the requirements of the Office of Foreign Assets Control (OFAC) of the U.S. Department of the Treasury. This request does not introduce changes to survey content or to the costs to the federal government. Changes to the estimated respondent burden are described below and reflected in the revised Supporting Statement Parts A and B.

As described in the approved study plan (OMB# 1850-0631 v.10-13), PayPal is used as one option for providing incentives to sample members to complete the student interview. In March 2017, during data collection for the BPS:12/17 calibration sample, the PayPal Compliance office notified RTI International (RTI), the BPS:12/17 data collection contractor, that three sample members who were identified to receive prepaid incentives had been flagged as persons possibly sanctioned by OFAC. OFAC administers and enforces economic and trade sanctions based on U.S. foreign policy and national security goals. As part of its enforcement efforts, OFAC publishes a list of individuals and companies called the "Specially Designated Nationals and Blocked Persons" List (SDN). Their assets are blocked and U.S. entities are prohibited from conducting trade or financial transactions with those on the list (https://www.treasury.gov/resource-center/sanctions/Pages/default.aspx).

Once these three sample members were flagged, PayPal requested the sample members' dates of birth in order to clear them from the SDN list. However, due to NCES confidentiality and security requirements, PayPal was notified that such information could not be provided. In order to maintain account activity, RTI

removed these individuals from the PayPal prepayment submission.

To comply with OFAC sanctions and to ensure the BPS:12/17 PayPal account remains in good standing, RTI began implementing methods to identify sample members who may match those listed on OFAC's SDN list. Programmatic matching using the Jaro-Winkler and Soundex algorithms recommended by OFAC, was performed on the entire BPS:12/17 sample (n=33,754). These methods are described in Attachment 1.

This matching process resulted in 345 potential matches between the BPS:12/17 sample and the SDN list. These 345 cases were manually reviewed and 30 cases could not be ruled out as sanctioned individuals. Given that BPS:12/17 data collection is already underway and it would be difficult to establish separate processes for these cases, we plan to exclude the 30 cases from the survey. The remaining 315 individuals were ruled out as matches to individuals on the OFAC SDN list. We recommend these individuals receive incentive payments by check only, and not be offered the PayPal option.

The rationale for excluding 30 sample members from the student interview is based on the inability to offer incentives for participation while complying with OFAC requirements. Incentives are an integral part of the study methodology for this longitudinal cohort, and they have been offered for each of the prior interviews, BPS:12/14 and NPSAS:12. Furthermore, keeping these 30 sample members in the study without offering incentives would require substantial change to study materials and the interview instrument, all of which have been developed and are currently in use including references to the incentive offer. Such changes have cost and risk implications that are disproportionate to the benefit of keeping the 30 sample members in the student interview.

The rationale for removing PayPal as a payment option for the 315 individuals is based upon their initial match to the SDN list using programmatic matching techniques similar to those used by OFAC's Sanction List Search tool. Because PayPal is understood to also use the matching techniques recommended by OFAC, it is expected that these 315 cases would be blocked by PayPal, as the original three were during the calibration sample data collection. Dates of birth, which were essential to RTI's manual review process, cannot be shared with PayPal due to NCES's confidentiality and security requirements. Therefore, PayPal would be unable to "unblock" these individuals. Rather than jeopardize the BPS:12/17 PayPal account and ability to use this popular payment method for the remainder of the sample, the PayPal payment option will not be offered to these 315 cases.

Part A of the approved OMB package has been revised to reflect the reduced estimated burden corresponding to the BPS:12/17 sample members being excluded from the survey. The total burden to respondents for student interview was reduced from 14,459 hours to 14,447, which translates to a reduction in estimated respondent burden time cost from $289,614 to $289,374. The reductions of cost and respondent burden are reflected in Table 2 in Supporting Statement Part A. An explanation of the matching process and results were also added to Part A, section A.9.

Supporting Statement Part B has also been updated to address the exclusion of the 30 sample members. The description of the BPS:12/17 sample has been adjusted, including Table 3. Aspects of the responsive design described in section 4.b have been updated. The targeted sector groups, shown in Table 8, have been revised based on the revised sample. Power calculations have been recalculated with the removal of the 30 cases, and the results are shown in a revised Table 11. Furthermore, tables 8 and 11 were revised slightly to be consistent with other tables that report institutional sector, using institutional sector at the time of sampling for NPSAS:12.

## Attachment 1. Matching methodology

The U.S. Department of the Treasury's OFAC administers and enforces economic and trade sanctions based on U.S. foreign policy and national security goals. As part of its enforcement efforts, OFAC publishes the SDN, a list of individuals and companies whose assets are blocked and whom U.S. entities are prohibited from conducting trade or financial transactions with (https://www.treasury.gov/resource-center/sanctions/Pages/default.aspx), and makes available a set of six other non-SDN sanctions lists in a consolidated set of data files, the "Consolidated Sanctions List." Collectively, we refer to the combined set of the SDN list and the Consolidated Sanctions List, as *SDN+*.

In addition to SDN+, OFAC also provides a web-based application, Sanctions List Search (SLS), designed to facilitate the use of SDN+ for individual cases. SLS makes use of the Jaro-Winkler and Soundex algorithms to determine if individuals are on the SDN list.

The following is a step-wise description of the matching and review processes used by RTI, which parallels the logic of SLS to the extent possible.

### Step 1: Download

The latest versions of the SDN list and Consolidated Sanctions List was downloaded from OFAC's website[1], time-stamped, and stored in a dated repository.

### Step 2: Pre-processing

The SDN list and Consolidated Sanctions List were merged together to form a single list, referred to as SDN+. As noted in OFAC's FAQ, some records on the Consolidated Sanctions List appear on the SDN list, therefore SDN+ is de-duplicated and subset to only individuals.

The SDN+ name field was in the form, "LAST NAME, first name." A second data element was created in the form, "first name last name" and was used for all subsequent matching.

The SDN+ remarks field contains various identifiers and features, including DOB, in a partially-structured text format. We parsed this field to extract DOBs that conformed to the format, "DOB dd Mmm YYYY." Approximately 76 percent of SDN+ cases had a DOB that matched this format.

Sample data were imported. First and last name fields were concatenated into a new data element in the form "first name last name." Given that BPS:12/17 is the second follow-up study, multiple previous names provided by the sample member are sometimes found in the imported data. For example, a single student may have provided names of "John Michael Astor" and "John Astor" previously. Some noise remains in the data, such as a small number of cases where the student's first name is listed as "John" and last name is "John Astor," so the concatenated form would result in "John John Astor."

### Step 3: Name Matching

OFAC's Sanction List Search (SLS) combines two fuzzy-matching algorithms – Jaro-Winkler and Soundex – with two matching techniques – name elements and full name – to calculate a single match score, expressed as a match "percentage." Likewise, in our matching we used both algorithms and two similar matching strategies.

[1] https://www.treasury.gov/resource-center/sanctions/SDN-List/Pages/sdn_data.aspx

## Fuzzy-matching algorithms

**Jaro-Winkler.** The Jaro-Winkler algorithm (Winkler 1990) is an extension of the Jaro distance measure (Jaro 1989) developed to compare names for the U.S. Census. The earlier Jaro measure considers the number of character matches and the ratio of their transpositions to change one word into the other. This difference between two strings is expressed as a continuous value between 0 and 1, with a value of 0 indicating an exact match and 1 indicating no similarity. For our matching, we used Winkler's standard prefix scale of $p = 0.10$.

**Soundex.** The Soundex algorithm is a phonetic index, not a strictly alphabetical one. This algorithm converts a name to a code, where the first letter is the first letter of the word, and numbers represent phonetic parts of latter syllables. Vowels, some consonants, and repeated sounds are ignored.

In our matching, we implemented the Soundex algorithm used by the National Archives and Records Administration (National Archives and Records Administration 2007). This version differs slightly from the original algorithm patented by Russell (Russell, Index 1918) and (Russell, Index 1922). This implementation of Soundex produces a binary distance of the match: 0 if a match occurs and 1 if no match is found.

## Matching strategies

**Name elements.** OFAC describes the name elements strategy as follows:

> "[This] technique involves **[1]** splitting the name string entered into multiple name parts (for example, John Doe would be split into two name parts). **[2]** Each name part is then compared to name parts on all OFAC's sanctions lists using the Jaro-Winkler and Soundex algorithms. The search calculates a score for each name part entered, and **[3]** a composite score for all name parts entered."

We broke this match technique into a sequence of three stages.

**1.** In the first stage, names were split into component parts. OFAC does not specify how the name-splits are determined, and whether symbols or characters other than whitespace, implied by the example, are used as splitting characters. Hyphenated names posed a challenge as they can be clearly the composition of two names, (e.g., 'Emily Harris-Crowley') or a more nuanced composition (e.g., name particles such as 'Bassam al-Hassan'). We chose a more inclusive approach in which names were decomposed into a sequence of elements by splitting on both empty characters (' ') and hyphens ('-').

As a reference system, let $S$ be an ordered list of elements (tuple) containing all names on the SDN+ list and $B$ be the tuple of BPS:12/17 sample member names. $M_i$ and $N_j$ are themselves name tuples within $S$ and $B$, respectively. That is

$$S = (M_1, M_2, \ldots, M_i) \text{ and } B = (N_1, N_2, \ldots, N_j)$$

Let $M_i$ be a tuple containing the decomposed elements, $e_k$, of a particular full name, such that

$$M_i = (e_1, e_2, \ldots, e_k)$$

In expanded form

$$S = \left((e_1, e_2, \ldots, e_k), (e_1, e_2, \ldots, e_k), \ldots, (e_1, e_2, \ldots, e_k)\right)$$

Likewise, $N_j$ is a tuple containing the decomposed elements, $f_l$, of a particular full name within $B$.

As a concrete example:

| **SDN+ name** | | | $M_i$ | $e_1$ | $e_2$ | $e_3$ |
|---|---|---|---|---|---|---|
| ‘lucy sulliman’ | → | $S$ | (‘lucy’, ‘sulliman’) | ‘lucy’ | ‘sulliman’ | |
| ‘ljiljana zelen-karadzic’ | → | | (‘ljiljana’, ‘zelen’, ‘karadzic’) | ‘ljiljana’ | ‘zelen’ | ‘karadzic’ |
| ‘hani al-tikriti’ | → | | (‘hani’, ‘al’, ‘tikriti’) | ‘hani’ | ‘al’ | ‘tikriti’ |

2. In this second stage, each $M_i$ and $N_j$ was compared for similarity using Jaro-Winkler and Soundex. Let J $(x, y)$ be defined as a function which takes as its input two name sequences and outputs the Jaro-Winkler distance between the two names, and let S$(x, y)$ be a function that outputs the Soundex distance between two name sequences. The output of J$(x, y)$ and S$(x, y)$ is an $m \times n$ matrix, $O_J$ and $O_S$ (Jaro-Winkler and Soundex, respectively) where $m = length(x)$ and $n = length(y)$, showing the distance between each pair-wise element within $x$and $y$.

As an example, for a SDN+ name of ‘Lucy Sulliman’ and a sample member named ‘Lu-Chi Su,’ the two tuples would be

| **SDN+ name** | $M_i$ | $e_1$ | $e_2$ |
|---|---|---|---|
| ‘lucy sulliman’ | (‘lucy’, ‘sulliman’) | ‘lucy’ | ‘sulliman’ |

| **Sample member name** | $N_j$ | $f_1$ | $f_2$ | $f_3$ |
|---|---|---|---|---|
| ‘lu chi su’ | (‘lu, ‘chi’, ‘su’) | ‘lu’ | ‘chi’ | ‘su’ |

Therefore, the resulting matrix from a Jaro-Winkler match would be

$$O_J = \begin{bmatrix} J(e_1, f_1) & J(e_1, f_2) & J(e_1, f_3) \\ J(e_2, f_1) & J(e_2, f_2) & J(e_2, f_3) \end{bmatrix}$$

¿
$$\begin{bmatrix} J(\text{‘lucy’}, \text{‘lu’}) & J(\text{‘lucy’}, \text{‘chi’}) & J(\text{‘lucy’}, \text{‘su’}) \\ J(\text{‘sulliman’}, \text{‘lu’}) & J(\text{‘sulliman’}, \text{‘chi’}) & J(\text{‘sulliman’}, \text{‘su’}) \end{bmatrix}$$

¿
$$\begin{bmatrix} 0.167 & 1.000 & 0.417 \\ 0.417 & 0.514 & 0.250 \end{bmatrix}$$

And for Soundex

$$O_S = \begin{bmatrix} S(e_1, f_1) & S(e_1, f_2) & S(e_1, f_3) \\ S(e_2, f_1) & S(e_2, f_2) & S(e_2, f_3) \end{bmatrix}$$

¿
$$\begin{bmatrix} S(\text{‘lucy’}, \text{‘lu’}) & S(\text{‘lucy’}, \text{‘chi’}) & S(\text{‘lucy’}, \text{‘su’}) \\ S(\text{‘sulliman’}, \text{‘lu’}) & S(\text{‘sulliman’}, \text{‘chi’}) & S(\text{‘sulliman’}, \text{‘su’}) \end{bmatrix}$$

$$¿\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

3. In the third stage, a composite score was created for all name elements. OFAC's SLS applies some function to $O_J$ and $O_S$ which maps the two matrices to a single composite score, normalized to a range of 0 to 100. However, the FAQ does not provide a description of this function detailed enough to reconstruct how $O_J$ and $O_S$ are combined—especially given that Soundex returns a binary value of 0 or 1 while Jaro-Winkler returns a continuous value between 0 and 1—which is to be interpreted as a percentage: "A value of 50 will return all names that are deemed to be 50% similar based upon the matching logic of the search tool."[2]

To create our own mapping function, we drew on other criteria mentioned in OFAC's FAQ. In Step 3 of the FAQ on Assessing OFAC Name Matches, OFAC notes that when two or more names are provided, a single name match (e.g., "just the last name") is insufficient to be counted as a valid match. We therefore extract the two lowest elements from $O_J$ and take their mean, which we call $l_J$, and extract the lowest two elements from $O_S$ and take their mean, which becomes $l_S$. It is important to note that, while OFAC provides the example of "just the last name," by extracting the lowest two elements within a matrix we are indifferent to order. This order indifference can be noticed in cases with multipart first names (sometimes indicating a middle name) or multipart last names; for example, "Emily Sarah Crowley" would match with "Emily Sarah," although this could be considered just the first name. We implemented this as it is a more inclusive approach, and later filtering steps trim down the false positive list.

In the matrix $O_J$ provided above in stage 2, this would correspond to

$$\begin{bmatrix} J(\text{'lucy'},\text{'lu'}) & J(\text{'lucy'},\text{'chi'}) & J(\text{'lucy'},\text{'su'}) \\ J(\text{'sulliman'},\text{'lu'}) & J(\text{'sulliman'},\text{'chi'}) & J(\text{'sulliman'},\text{'su'}) \end{bmatrix}$$

$$¿\begin{bmatrix} 0.167 & 1.000 & 0.417 \\ 0.417 & 0.514 & 0.250 \end{bmatrix}$$

Which would be combined into

$$l_J = \frac{0.167 + 0.250}{2}$$

$$¿ 0.209$$

Likewise, this function would map $O_S$ to $l_S = 1$.

Rather than merging both $l_J$ and $l_S$ into a single composite score, as OFAC does, we retained these two elements and later combine them into a composite score which factors in both matching strategies (explained in the **Match score** section, below).

To add a matched-pair, $M_i$ and $N_j$, to a list of *possible name matches*, we used an inclusion filter of $l_S < 0.15$. Only matches in which the mean of the lowest 2 Jaro-Winkler distances is less than 0.15 were considered further. A value of 0.15 was chosen, as initial tests between single names (e.g., $J('markus','mark')$) indicated that a Jaro-Winkler distance $¿ 0.15$ seemed to be an upper ceiling on relatively similar terms, after which match quality degraded heavily.

[2] https://www.treasury.gov/resource-center/faqs/Sanctions/Pages/faq_compliance.aspx#5.

Note that Soundex was not used at this point as an inclusion filter as we had concerns about the accuracy of using Soundex on non-Western (e.g., Arabic and Asian) names (Patman and Shaefer 2003). Issues with using Soundex have motivated the development of several new phonetic matching algorithms. For example, the Daitch-Mokotoff Soundex was developed because of problems encountered while trying to apply the Russell Soundex to Germanic and Slavic surnames. Similarly, Metaphone, Double Metaphone, and Metaphone 3, were all developed to incorporate features such as handling a set of non-Latin characters and accommodating some non-English words to address inadequacies in the Soundex algorithm. We continue to use Soundex, despite these concerns, in order to more closely parallel OFAC's SLS.

**Full name.** OFAC describes its second match technique, the full name match, as follows:

> "[This] technique involves using the Jaro-Winkler algorithm to compare the entire name string entered against full name strings of entries on OFAC's sanctions lists."

Rather than decomposing the full name into elements, as in the first match strategy, this technique applies only the Jaro-Winkler algorithm to each match-pair combination. In our approach, we considered only possible match-pairs identified by the *name elements* strategy. For each match-pair, we applied $J(x, y)$ to the non-decomposed names on SDN+ and in the sample. This individual score was saved as a new data element, $f_J$ – a value between 0 and 1.

**Match score.** To produce a single match score which is output to the user, the FAQ notes that OFAC's SLS returns the higher of the two techniques' score in the score column.[3] That is, either the Jaro-Winkler only score is returned or the Jaro-Winkler/Soundex score is returned.

In our approach, we departed from OFAC's either/or approach and create a single match score, or composite score, which combined all three measures. Our composite score is defined as the mean of the three distance measures, that is

$$c = \frac{l_J + l_S + f_J}{3}$$

This effectively allows each distance measure to cast a "vote" and produced a single score from their resulting combination. An advantage to this approach is that it considers all the distance measures in a single score. In OFAC's SLS, the system would score a reversed name (e.g. 'Ann Thomas' and 'Thomas Ann') a 100 percent, exact match whereas our system would return a less definitive, though still likely, value of 0.10.

### Step 4: Alternative Name Matching

The same stages described in step 3 above were applied to the list of AKA or alternative names included in SDN+. The resulting list of possible names was called the list of *possible alternative name matches*.

As noted in OFAC's FAQ, "OFAC does not expect that [financial institutions and others] will screen for weak AKAs, but expects that such AKAs may be used to help determine whether a 'hit' arising from other information is accurate." Due to their general nature and likelihood of generating a considerable number of "matches," we did not programmatically parse weak AKAs nor include these in the fuzzy match search. However, weak AKAs were taken into later consideration for determining a study exclusion.

[3] Note that the Sanctions List Search produces a score from 0 to 100, where 100 indicates an exact match, while our method produces a distance in which a score of 0 indicates an exact match.

### Step 5: List Consolidation and Sorting

The list of possible name matches was combined with the list of possible alternative name matches (note that at this point, match-pairs may appear on both the possible name match list and the possible alternative name match list). The resulting combined list was then subset to only match-pairs with a composite score, $c$, of 0.10 or less. This filter parameter was established as a maximum score after which match-pairs' quality seemed to degrade.

### Step 6: Manual Review

Files were then prepared for manual review. A new data element was created for the year difference between the DOB extracted from SDN+ remarks field and the DOB on file for the sample member to aid the reviewer in manually examining the quality of a match-pair (the lower the resultant number, the more likely the match). The list was then sorted by the composite score and exported to Microsoft Excel files, which include the following field:

- SDN+ remarks column
- $l_J$: Jaro-Winkler lowest-two score
- $l_S$:Soundex lowest-two score
- $f_J$: Jaro-Winkler full name score
- $c$: Composite score
- SDN+ DOB
- Sample member (e.g., BPS:12/17) DOB
- Year difference between both SDN+ DOB and sample member DOB
- SDN+ name
- Sample name (e.g., BPS:12/17 name)

The file was reviewed moving from top to bottom (i.e., from most likely match to least).

OFAC notes that, in cases with multiple name components, more than just one component must match. Therefore, the file is manually reviewed for multiple name matches; a match on both first and last name is seen as more likely. At this stage, the reviewer follows OFAC's guidance for assessing match quality.[4]

Manual review is subjective, but was guided by the following principles:

- Possible match-pairs should only be considered valid matches if several names appear to be either misspellings (e.g., 'Richard and 'Richaard'), short forms (e.g., 'Andy' and 'Andrew'), alternative spellings (e.g., 'Sayied' and 'Sayeed'), or otherwise appear to share several similarities across SDN and the sample list.
- We now consider potential matches as only those that match on both first (or middle) name and last name. Additionally, we consider inverted names (e.g., 'Adrian Tomas' and 'Tomas Adrian') as a match-pair for further investigation.

[4] https://www.treasury.gov/resource-center/faqs/Sanctions/Pages/faq_compliance.aspx#5.

Particularly for Arabic names, if a first name is composed of three or more names (e.g., 'Abu Sufian al-Salamabi Muhammed Ahmed'), we would consider a match-pair for further investigation if it matched on the first name, but not on the last. Likewise, if the last name is composed of three or more names, we would consider a match-pair further if it matched only on the last name. Given that Arabic names do not conform to strict first name/last name conventions employed in the West, this allows for some variability in how names may have been divided in either SDN+ or our records. RTI will continue to investigate how best to handle non-Western names.Cases identified for further investigation were then moved to the second stage of manual review. In this stage, cases were further winnowed by examining DOB consistency (we consider a DOB match any DOB difference within ±3 years or a match on month and day). The resulting list was then subset to only match-pairs with a composite score, $c$, of 0.10 or less.

## References


Jaro, Matthew A. 1989. "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida." *Journal of the American Statistical Association* 84 (406): 414-420.

National Archives and Records Administration. 2007. *The Soundex Indexing System.* May 30. Accessed March 25, 2017. https://www.archives.gov/research/census/soundex.html.

Patman, Frankie, and Leonard Shaefer. 2003. *Is Soundex Good Enough for You? On the Hidden Risks of Soundex-Based Name Searching.* Herndon, VA: Language Analysis Systems, Inc.

Russell, Robert C. 1918. Index. United States of America Patent US1261167 A.

Russell, Robert C. 1922. Index. United States of America Patent US1435663 A.

Winkler, William E. 1990. "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage." *Proceedings of the Section on Survey Research.* 354–359.