REPEATED AND COMBINED PARTICIPATION IN WORKFORCE, EDUCATION AND SOCIAL PROGRAMSERVICES: PRELIMINARY EVIDENCE FROM MARYLAND ADMINISTRATIVE DATA SOURCES[1]

Submitted to:

U.S. Department of Labor
Employment and Training Administration
Office of Policy Development and Research
Division of Strategic Planning and Performance
200 Constitution Avenue NW
Washington DC 20210

Submitted by:

Grace Fendlay
Director, Discretionary Grants
Division of Workforce Development and Adult Learning
Maryland Department of Labor, Licensing and Regulation
1100 North Eutaw Street
Baltimore, MD 21201

March 2012

[1]The authors of this report are Ting Zhang, PhD, Research Assistant Professor, and David Stevens, PhD, Research Professor, The Jacob France Institute, University of Baltimore. John Janak, Sang Truong and Jing Li participated in the data processing conducted for this research. The Institute is a sub-award recipient of Workforce Data Quality Initiative (WDQI) funds received by DLLR from the U.S. Department of Labor, Employment and Training Administration. The authors accept full and sole responsibility for the content of this report. Agreement or disagreement with the views expressed here should not be attributed to any other person or organization. Correspondence should be addressed to tzhang@ubalt.edu.

EXECUTIVE SUMMARY

One justification for extended retention of program administrative data is a resulting capability to study whether and how often repeated, concurrent and sequential appearances are found in multiple program transaction records. The expected return-on-investment from extended retention of administrative data rises when coverage of multiple programmatic interventions is consolidated in an integrated data system.

This report is based on a preliminary exploration of linked historical administrative data files containing recorded instances of individual participation in defined education, workforce and social service programs. Our results to date should be treated as illustrative of decision-relevant insights that future refinements can produce.

To explore the incidence of repeated, concurrent and sequential participation in more than one government program we accessed two workforce, threepublic education, and two social services administrative data files.

We document that many individuals do engage in multiple workforce, education and social services program activities, and in repeated participation within each of the program types.

There are many reasons for sustained interest in reliable measurement of repeated and combined program engagements. Some combinations and sequences are looked upon favorably, while others are viewed with some concern. Care should be exercised to avoid hasty conclusions about the benefits and costs that can be assigned to particular combinations of program engagements.

Our next steps will include renewed attention to dynamic person identification and program engagement methodologies, while continuing the design and implementation of our study of up to 27 years of multiple program engagements by 1984 credit-course enrollees in Maryland's public community colleges.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

INTRODUCTION

One justification for extended retention ofprogram administrative data is a resulting capability to study whether and how often repeated, concurrent and sequential appearances are found in multiple program transaction records. We seek improved understanding of the life-cycle incidence and mix of program engagements.[2]

The return-on-investment from extended retention of administrative data rises when coverage of multiple programmatic interventions is consolidated ina P-20W Longitudinal Data System (LDS).[3] The expectation of a higher return-on-investment reflects awareness that there are complex interdependencies among program participation engagements, and among the immediate and long-term impacts of these encounters on the participants and society.

This report is based on a preliminary exploration of linked historical administrative data files containing recorded instances of individual participation in defined education, workforce and social service programs. The frequency and mix of programmatic engagements that we document here are artifacts of the time and program coverage available to us when the study was undertaken. Our results to date should be treated as illustrative of valuable decision-relevant insights that future refinements can produce.[4]

This report is the second[5] in a planned series of related studies being conducted in partnership with the Maryland Department of Labor, Licensing and Regulation (DLLR) and other Maryland state agencies, using Workforce Data Quality Initiative (WDQI) funds awarded to DLLR by the U.S. Department of Labor, Employment and Training Administration, Office of Policy Development and Research, Division of Strategic Planning and Performance.

---

[2]We use the intuitive word 'engagements' throughout the report. This is intended to convey a clear understanding of our ultimate interest in individual participant involvement in defined services recorded in administrative records.

[3]The commonly used acronym P-20W refers to the full time spectrum from Early Childhood or Preschool (P) through postsecondary education (20) and/or workforce (W) engagement.

[4]We, and some colleagues in the multi-state network of Administrative Data Research and Evaluation (ADARE) state partners, are conducting related statistical analyses sponsored by the U.S. Department of Agriculture, Economic Research Service, concentrating on Supplemental Nutrition Assistance Program (SNAP) and Unemployment Insurance (UI) benefit spells. Daniel Schroeder, Ray Marshall Center, University of Texas, Austin, and Peter Mueser, Economics Department, University of Missouri, Columbia, will be leading this collaborative research.

[5] The first report in the series, Zhang, T. and Stevens, D. (2012), *P-20W Integrated Data System Person Identification: Accuracy Requirements and Methods*, Baltimore, MD: University of Baltimore, The Jacob France Institute, is available at http://www.jacob-france-institute.org/documents/MD-WDQI-Person-Identification-Report.pdf.

BACKGROUND

Researchers have long recognized the occurrence and importance of an individual's participation in multiple programs, and the need to measure impacts on recipients of defined services.[6]The *Intelligence for Social Policy* program[7], sponsored by the John D. and Catherine T. MacArthur Foundation, hosted by the University of Pennsylvania, and co-directed by Penn professors Dennis Culhane and John Fantuzzo, is a focal point for current integrated data system (IDS) research and applied case-management practices that use program engagement transactions data found in site-specific portfolios of administrative data sources.

METHODOLOGY

Our primary data linkage method has been probabilistic matching and subsequent computer-based and manual review to effectively link administrative records and identify potential issues related to SSN or other identifiers. We began by checking SSN validity using the Social Security Administration's monthly SSN issuance schedule. The remainder of thissection introduces detailed data sources, methods and identifier information.

Data Sources

Two workforce, three public education, and two social services administrative data files were available for our authorized use.

The twoworkforce datasets are:

- MD Job Training Partnership Act (JTPA), 1984-2004; and
- MD Workforce Exchange data system (WFE), 2005-2009.

The three education datasets are:

- A single school district, School District A[8] Data (SDAD) student record extracts, 1998 to 2010;

[6]Many research studies have concentrated on immediate and short-term impacts on the recipients of services only. This has usually been a compromise necessitated by unavailability of reliable information needed for longer-term perspectives and consideration of impacts beyond the recipients of services. This reinforces the importance of sustained IDS capabilities.

[7]http://www.ispc.upenn.edu. An earlier, but still useful resource is: *An inventory of research uses ofadministrative data insocial services programsin the United States,* A Report by UC DATA, Berkeley, CA: University of California at Berkeley to the Northwestern University/University of Chicago Joint Center for Poverty Research (February 1, 1999), available at http://ucdata.berkeley.edu/pubs/inventory/entire.pdf.

[8] We use "School District A" to avoid disclosure.

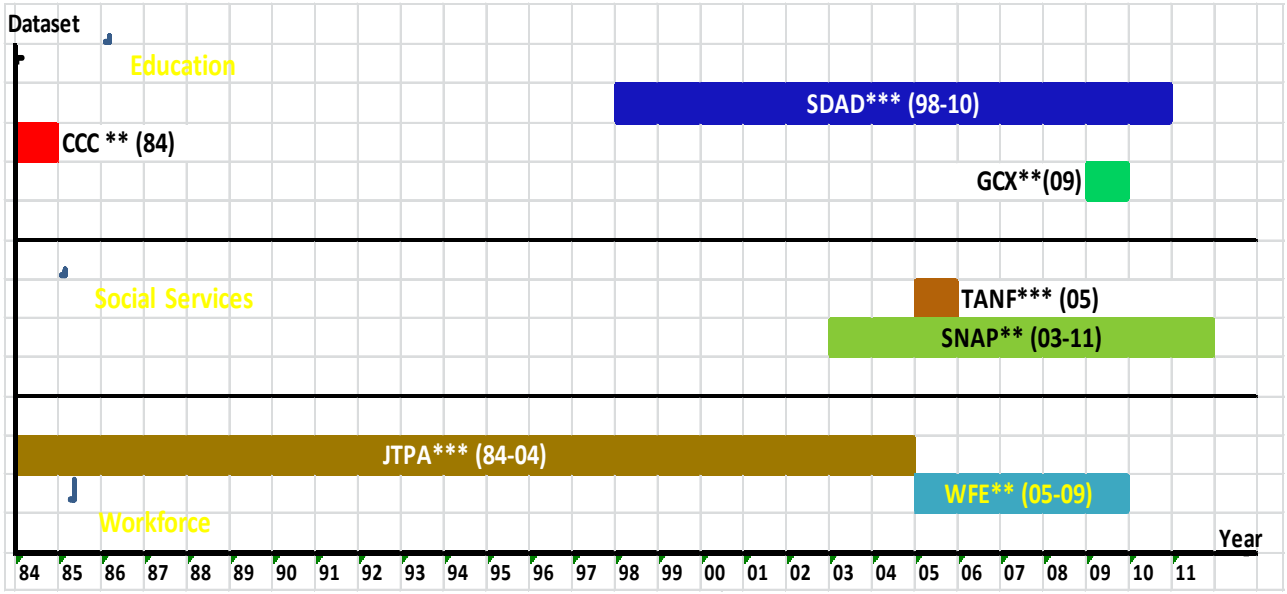
- A single cohort of public community college enrollees, 1984 (CCC); and
- A single cohort of public high school graduates,Graduates Cohort X[9] (GCX) student record extracts, 2009; a smaller extract than SDAD.

The two social services datasets are:

- Temporary Assistance for Needy Families benefit recipients(TANF), 2005; and
- Supplemental Nutrition Assistance Program benefit recipients (SNAP), 2003-2011.

Figure 1 shows the time coverage of each administrative data file.

Figure 1 Profile of Datasets Used



**	SSN + DOB, or SSN + Name, or +gender (or ethnicity or education).
***	SSN + Name + DOB + gender (or ethnicity or education or address).

Record linkage methods

We implemented athree-step approach to study engagementsin the seven programs. We started with the four datasets used for our *P-20W Integrated Data System Person Identification: Accuracy Requirements and Methods*report—SDAD, GCX, JTPA, and WFE. We first identified potential matches based on consistent or

[9] Again, the specific dataset name is not disclosed.

highly similar Social Security Numbers (SSN), date of birth (DOB), full name, race and ethnicity, gender, and education, using Link Plus (a commercial software product).

We carried outseveral deterministic and probabilistic matching diagnostics among the four datasets. Link Plus enables a match between a nine digit SSN in one file and a four digit SSN in another file. If the last four digits of the nine digit number are the same as the four digit number, the comparison pair receives a high score. Link Plus also identifies specialized name and date matches, exact matches, and enables phonetic matching on blocking variables[10], such as person name components, or any variable for which pronunciation versus spelling helps to identify an individual record.

Our second step was to separate the potential matched pairs into categories and then conduct computer-based or manual review to verify true matches. This verification was possible because multiple person identification data fields are found in the available administrative records. Linkage using multiple pairings of candidate data fields supports these cross-checks.

Our third step began with the true matches found from the diagnostics completed in our *P-20W Integrated Data System Person Identification: Accuracy Requirements and Methods*report. We confirmed the appearance of these true matches in the original four datasets and then attempted to find these true matches in each of the other three administrative data sets defined above and shown in *Figure 1*. We compiled the frequency count of matches in each pairing of the seven datasets.

To match records across administrative data sets, combinations of two or more of the followingdata fieldswere used as identifiers when available: Social Security number (SSN), date of birth (DOB), first name, surname, and middle name or middle initial; gender, race and ethnicity; and education status.

To achieve an acceptable linkage, we first scanned within each dataset to locate possible duplicates.After screening for possible duplication,wematchedacross datasets. Between each pairing of two datasets, we either used SSN or DOB as a blocking variable and the other data fields as matching variables.

In the next step we classified the matched pairs into categories based on which of the identifiers were shared. Those categories reflect different likelihoods of being atrue match. This step does not just depend on the matching score; matching score is not necessarily the optimal measure of true match likelihood.

[10] For files with millions of records, the total of all possible comparison pairs is too large for practical computation. Blocking Variables are variables common to the two files that are used to ‘block’ (or partition) the two files. Only within these blocks are matching variables compared between the records. Blocking is a way to reduce the computing cost by partitioning files into mutually exclusive and exhaustive blocks and performing comparisons only on records within each block.

We selectedonly the true matchpairs. If in a matched pair we identified the same SSN and same DOB, this pair identifies a true match and a verified individual. If a matched pair has the same SSN and different DOB and if the DOB is highly similar or missing in one record of the matched pair, but has the same (or basically same) full name, and if availablerace/ethnicity, gender and education information, the pair becomes a true match and the pair identifies a verified individual as well. If a matched pair has the same DOB and same combination of all other identifiers, but different SSN, we are not as sure whether the pair identifies an individual. Therefore, in the last case, no verified individual is identified.

Our last step began with only those records identified as true matches. We deterministically matched these true match records to each of the seven datasets, and counted the number of individuals that appear in combinations of the seven datasets.

FINDINGS

A total of 39,588 true matches were identified. Each true match identifies an individual with verified and consistent SSN, DOB, and if available names and other information. Among the 39,588 pairings, 33,088, or 84%, of the pairswere foundin more than one of the datasets. We concentrate on the distribution of these pairs among the seven datasets here.

The 6,500, 16%, pairs not described further here are individuals that appeared more than once within any one, but only one, of the seven datasets. These are what we described earlier as repeat engagements in a single program's services, which distinguishes them from concurrent and sequential users of multiple program services.

Table 1 introduces the mix of program engagements for the 33,088 true match cases that were identified in our probabilistic matching steps described in the previous section. A brief how-to-read guide for proper understanding of Table 1 follows next.

Start at the upper left number 18,789. This is the count of true match individuals that were found in the SNAP dataset and one or more of the other six datasets covered. The 57% appearing directly below the 18,789 count indicates that 57 percent of the 33,088 true matches include a pairing of SNAP engagement and participation in at least one other program.

Next, beginning to read down the left-hand column, find the 2,897 count followed by the 9% notation. This is a count of the true match individuals that were found in both the SNAP and SDAD datasets, but not necessarily exclusively so, as we will show later in Table 2. The 9% indicates the percentage of 33,088 true match cases that were found in both the SNAP and SDAD datasets.

The explanation offered in the previous two paragraphs should be followed for proper interpretation of the remaining cells in Table 1. The bottom row of zeros is a statistical artifact of the small number and early historical time-span coverage of the community college dataset.

Table 1 Frequency and Percentage of Program Engagement Pairings

% of Total	SNAP	SDAD	GCX	JTPA	WFE	TANF	CCC
SNAP	18789 *57%*						
SDAD	2897 *9%*	6422 *19%*			**Total:**	**33088**	
GCX	413 *1%*	434 *1%*	585 *2%*				
JTPA	16636 ***50%***	1911 *6%*	2 *0%*	28426 *86%*			
WFE	17815 ***54%***	4454 *13%*	152 *0%*	26898 ***81%***	31120 *94%*		
TANF	8836 *27%*	991 *3%*	92 *0%*	9293 *28%*	9687 *29%*	9994 *30%*	
CCC	50 *0%*	3 *0%*	2 *0%*	114 *0%*	114 *0%*	26 *0%*	116 *0%*

We ask that full attention be given to our introductory caution that this preliminary research is illustrative of what can be accomplished and learned when extended time coverage of multiple administrative datasets is available. Figure 1, on page 3, clearly shows when and for how long overlap among the seven datasets occurs. If common extended time coverage had been available for all seven of the datasets a different pattern of multi-program appearances would have emerged.[11] There are many reasons why one might expect, or not expect, particular subpopulations to appear in defined administrative datasets. Future thought about these reasons, followed by appropriate research diagnostics, will serve as the bridge to reach conclusions that have policy and program management relevance.

[11]Common time-span coverage would have been impossible, of course, because the JTPA file ends before the WFE dataset starts in 2005.

We also note here that we intentionally omitted inclusion of Maryland Unemployment Insurance (UI) wage records from this preliminary analysis. Our attention has been limited to program engagement events, not employment affiliations or timing.[12]

Table 2illustrates the distribution of multiple program engagement events among the three program types and seven datasets. Among the 33,088 verified true-match individuals, 39% appeared in two of the seven datasets; 34% appeared in three of the seven; 27% participated in four of the seven; and0.3% participated in five of the seven datasets.Table2 highlights in red font major program engagement combinations.

Counting across the three program types—workforce, public education, and social services—29% of the 33,088 verified individuals participated in only one of the three program types (but at least two defined programs within a classification of program *type*); 63% participated in two of the three program types; and 8% participated in all three program types.

[12]One future improvement in our analysis that will utilize UI wage records is to estimate whether and when an individual is known or highly likely to have resided in Maryland and therefore had local access to program participation opportunities. UI wage records are based on place of work, not residential address.

Table 2Combinations of Program Engagements

# of Data sets	# of Prog-ram Types	Overlapping Participation in over One Data Sources					Edu	Soc -ial	Work - force	Freq.	%	Cum. %
2	2	SDAD	SNAP				*	*		5	0.02	0.02
2	2	GCX	WFE				*		*	120	0.36	0.38
2	2	SDAD	JTPA				*		*	918	2.77	3.15
2	2	SDAD	WFE				*		*	2,309	6.98	10.13
2	1	JTPA	WFE						*	9,551	28.87	39.00
3	1	SDAD	CCC	GCX			*			1	0.00	39.00
3	2	SDAD	GCX	WFE			*		*	1	0.00	39.00
3	2	TANF	GCX	WFE				*	*	3	0.01	39.01
3	2	SDAD	SNAP	TANF			*	*		4	0.01	39.02
3	3	SNAP	GCX	WFE			*	*	*	26	0.08	39.10
3	3	SDAD	TANF	JTPA			*	*	*	28	0.08	39.18
3	3	SDAD	TANF	WFE			*	*	*	30	0.09	39.27
3	2	SDAD	TANF	GCX			*	*		45	0.14	39.41
3	2	CCC	JTPA	WFE			*		*	58	0.18	39.59
3	2	SDAD	JTPA	WFE			*		*	183	0.55	40.14
3	2	SDAD	SNAP	GCX			*	*		345	1.04	41.18
3	3	SDAD	SNAP	JTPA			*	*	*	392	1.18	42.36
3	2	TANF	JTPA	WFE				*	*	1,036	3.13	45.49
3	3	SDAD	SNAP	WFE			*	*	*	1,153	3.48	48.97
3	2	SNAP	JTPA	WFE				*	*	7,878	23.81	72.78
4	3	SDAD	TANF	JTPA	GCX		*	*	*	1	0.00	72.78
4	3	SDAD	SNAP	CCC	WFE		*	*	*	1	0.00	72.78
4	2	SNAP	TANF	GCX	WFE			*	*	2	0.01	72.79
4	3	TANF	CCC	JTPA	WFE		*	*	*	6	0.02	72.81
4	3	SDAD	TANF	JTPA	WFE		*	*	*	8	0.02	72.83
4	3	SNAP	CCC	JTPA	WFE		*	*	*	30	0.09	72.92
4	2	SDAD	SNAP	TANF	GCX		*	*		40	0.12	73.04
4	3	SDAD	SNAP	JTPA	WFE		*	*	*	123	0.37	73.41
4	3	SDAD	SNAP	TANF	JTPA		*	*	*	188	0.57	73.98
4	3	SDAD	SNAP	TANF	WFE		*	*	*	577	1.74	75.72
4	2	SNAP	TANF	JTPA	WFE			*	*	7,937	23.99	99.71
5	3	SDAD	TANF	CCC	JTPA	GCX	*	*	*	1	0.00	99.71
5	3	SNAP	TANF	CCC	JTPA	WFE	*	*	*	19	0.06	99.77
5	3	SDAD	SNAP	TANF	JTPA	WFE	*	*	*	69	0.21	99.98
Total										***33,088***		

CONCLUDING REMARKS

Expression of interest in repeated and combined participation in workforce, education, and social services programs is not new. Accelerated advance of data processing capabilities and use of these capabilities to assemble and analyze integrated data systems of administrative records that document concurrent and sequential program engagement events are relatively recent.

There are many reasons for sustained interest in reliable measurement of repeated and combined program engagements. Some combinations and sequences are looked upon favorably, such as a high rate of transition from high school graduation to postsecondary enrollment and subsequent persistence and attainment of a defined credential. Other combinations are viewed with some concern, such as a high and rising rate of Supplemental Nutrition Assistance Program participation, particularly if this program involvement is found to precede, be concurrent with or come after a recorded Unemployment Insurance benefit program spell.

Care should be exercised to avoid hasty conclusions about the benefits and costs that can be assigned to particular combinations of program engagement events. Detection and analysis of repeated program engagement and complex series of multiple program engagements are promoted by access to extended integrated data system time coverage.

We have shown in our first two reports completed under Workforce Data Quality Initiative auspices that accurate individual identification—Person Identification is the label we prefer—is an essential first criterion for successful progress toward ultimate documentation of life-cycle program interactions. We have presented in this report preliminary evidence about frequencies of defined combinations of program engagement events.

This report documents that many individuals do engage with multiple workforce, education and social services programs, as well as participating in repeated contact within a program type. Almost all of the observed individuals participated in two, three or four of the seven programs studied:

Our next steps will include continuing attention to dynamic person identification and program engagement methodologies, while continuing the design and implementation of our study of up to 27 years of multiple program engagements by 1984 credit-course enrollees in Maryland's public community colleges.[13]

[13]The latter stages of this research and anticipated release of findings remains contingent upon confirmation that our *Memoranda of Understanding* and analysis specifications are in full compliance with the new Family Educational Rights and Privacy*Final Rule* published in December 2011 that became effective in January 2012.