A Framework for Sharing Confidential Research Data, Applied to Investigating Differential Pay by Race in the U. S. Government

Supplement: Synthetic Data Validation

Duke University Synthetic Data Project

March 2, 2018

The following graphs and tables are excerpted from work done by the Synthetic Data Team at Duke University to validate the DIBBS synthetic federal employee data set with corresponding authentic data supplied by the U.S. Office of Personnel Management (OPM).[1] The selection here highlights two and three level covariate relationships, especially involving important research variables such as sex, race, age, education, agency, occupation, year, and pay. In assessing similarity of the data sets, emphasis is placed on utility, or the degree to which answers to meaningful research questions obtained from use of synthetic data agree with those from use of corresponding authentic data. Graphs and tables representing synthetic data contain the text "DIBBS" while those for authentic data contain either "OPM" or "JdF."[2]

TWO VARIABLE CORRELATION

Figure 1 shows, for pay plan GS, full-time observations, correlations between 1.) the variable indicated in the title bars and 2.) all levels of all other variables in title bars. Synthetic variable pair correlations are plotted (y-axis) against corresponding pair correlations in the authentic data (x-axis). Points lying near the reference line (slope of 1.0) indicate equality between data sets. Agency and occupation are truncated to the first two positions. Note that correlations involving categorical variables, or fixed effects, effectively measure the association of proportion of observations with levels of the second variable. Missing counts are the number of variable level combinations that appear in the other data set but not in the one indicating a count. For instance, JdF=2 in the agency panel would indicate that observations exist in the synthetic data, but not in the authentic data, for two combinations of agency and some level of a second variable. Note that all missing counts are a multiple of three. This is due to agencies AL, CP, and GD missing in the synthetic data.

Observation: Correlation of pairs of levels of variables within synthetic data are very near those of corresponding pairs in the authentic data. This is indicated by the near proximity of all plotted points to the reference line of slope 1.0, including those for extreme correlation values.

[1] A complete description of both data sets and sources is available in the main document that the current document supplements.

[2] "JdF" is nomenclature for a particular FOIA request that resulted in receipt of authentic data from OPM, which was used to generate synthetic data.

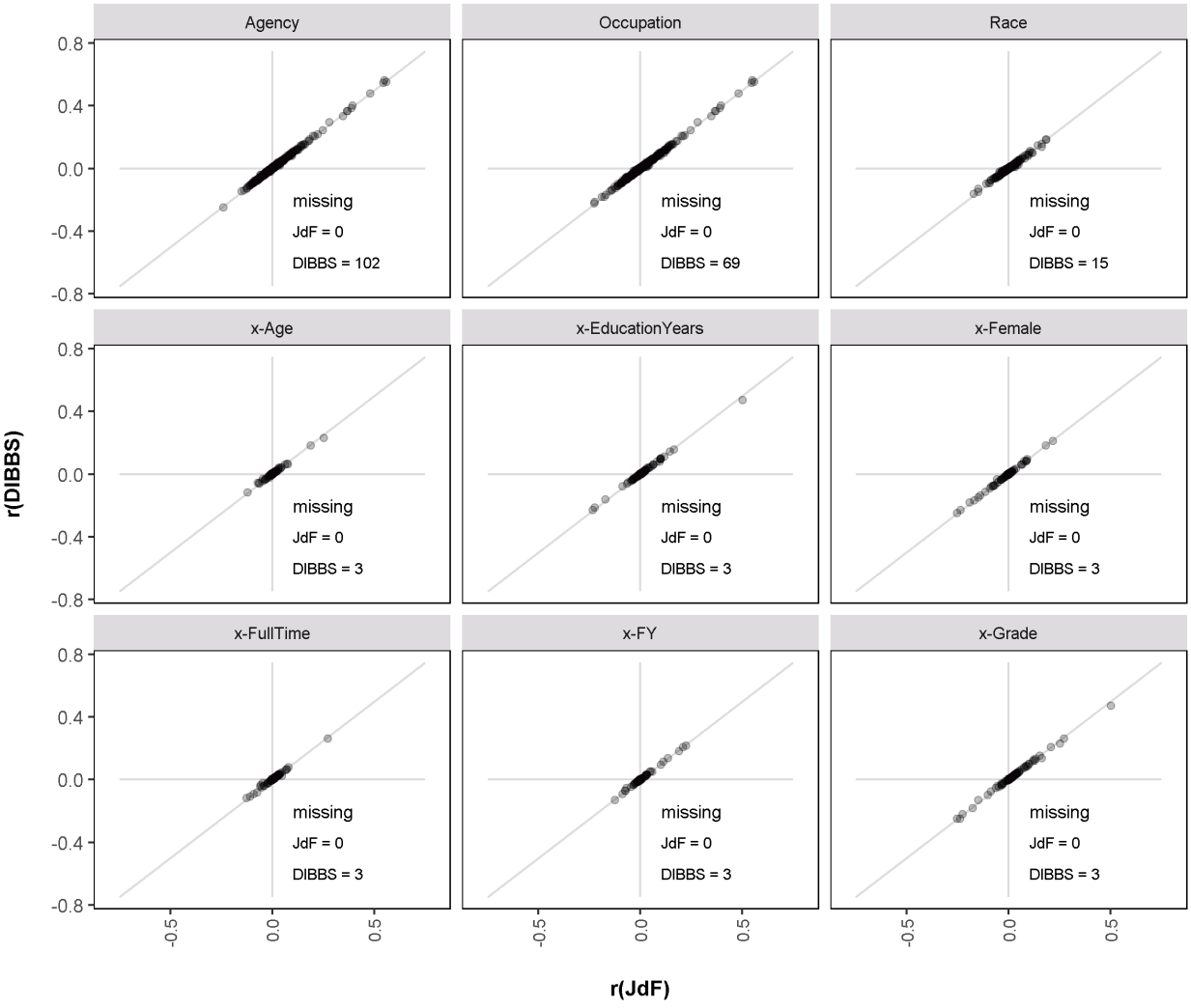
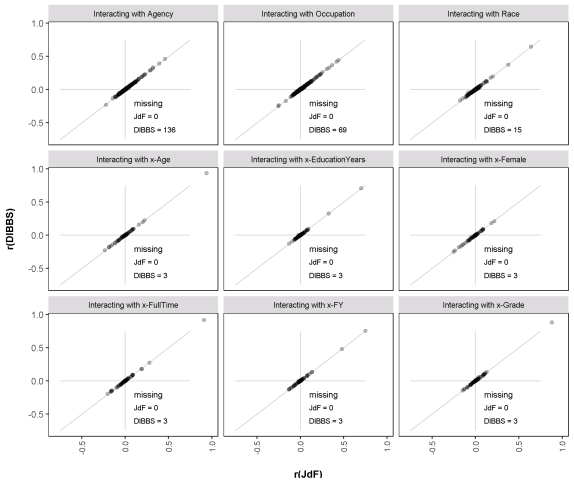


Figure 1: Two variable correlations of corresponding levels of synthetic and authentic data. Synthetic level correlation on y-axis, corresponding authentic correlation on x-axis.

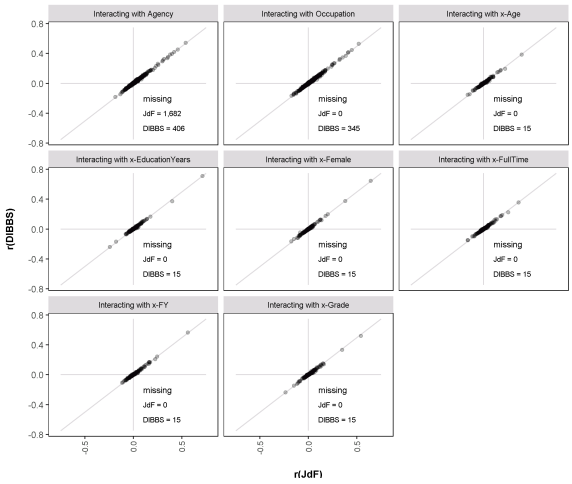
CORRELATION OF PRIMARY VARIABLES WITH TWO-VARIABLE INTERACTIONS

Figures 2 and 3 show, for pay plan GS, full-time observations, correlations between 1.) the variable indicated in the graph title, 2.) all combinations of levels of the variable listed in a title bar, and 3.) all levels of other variables appearing in the title bars. These constitute correlation of main variables with two variable interactions. In the case of categorical variables, or fixed effects, this is the association of a primary variable with the proportion of observations in interacting level combinations of two other variables. Agency and occupation truncated to first two positions.

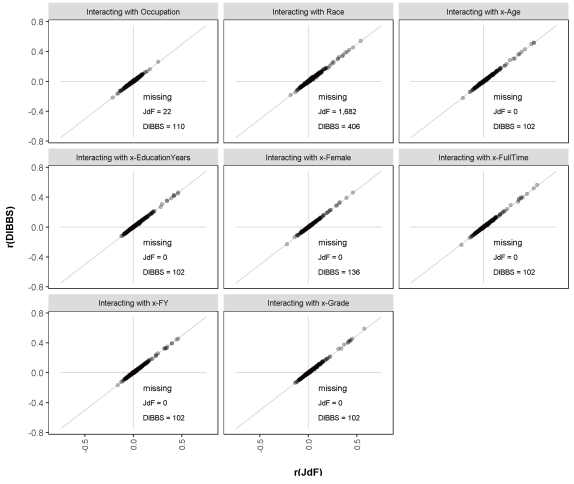
Observation: Proximity of all points to slope 1.0 reference line indicates agreement of three-variable associations between data sets and implies depth of utility beyond simple pairwise relationships.



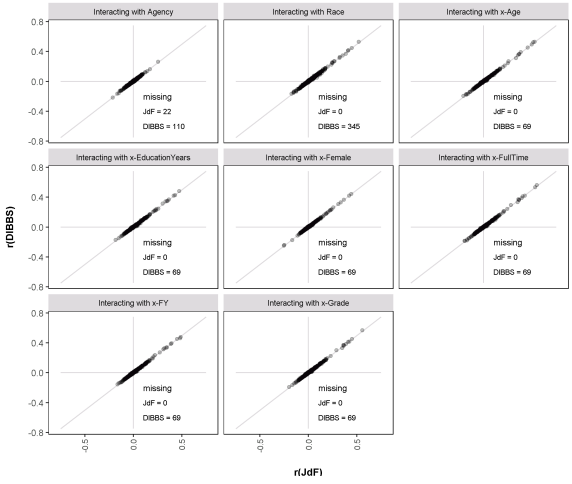
(a) Correlations involving Sex



(b) Correlations involving Race

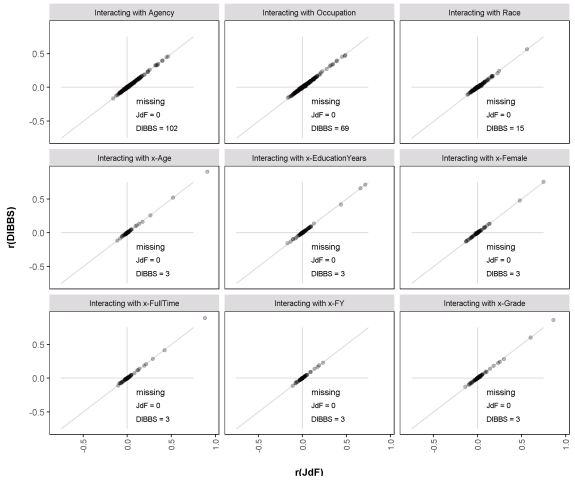


(c) Correlations involving Agency

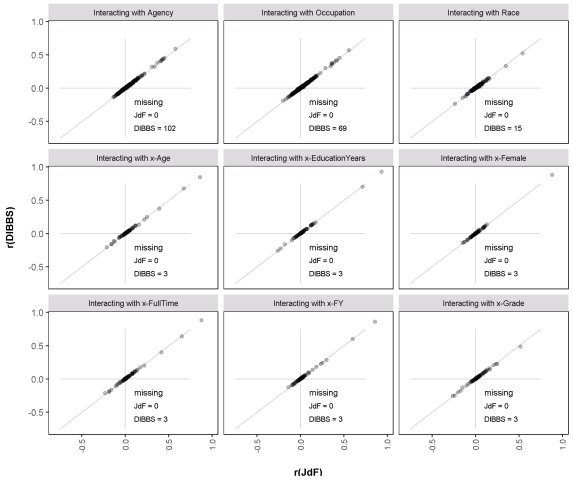


(d) Correlations involving Occupation

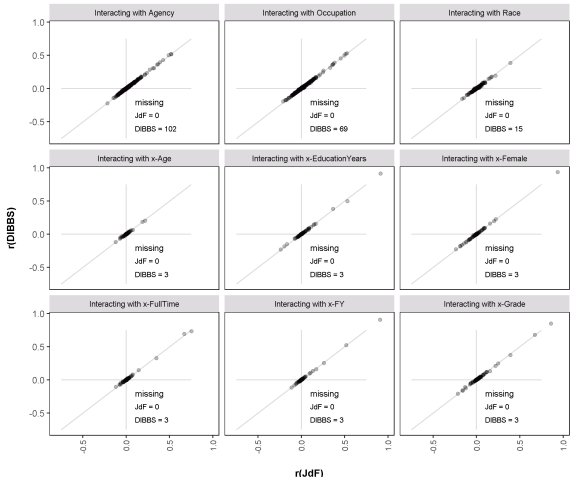
Figure 2: Correlations of primary variables with two variable interactions. Variable set one. Synthetic level correlation on y-axis, corresponding authentic correlation on x-axis.



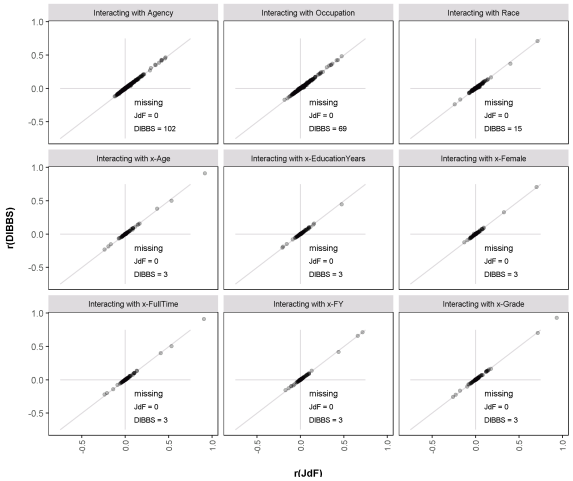
(a) Correlations involving Fiscal Year



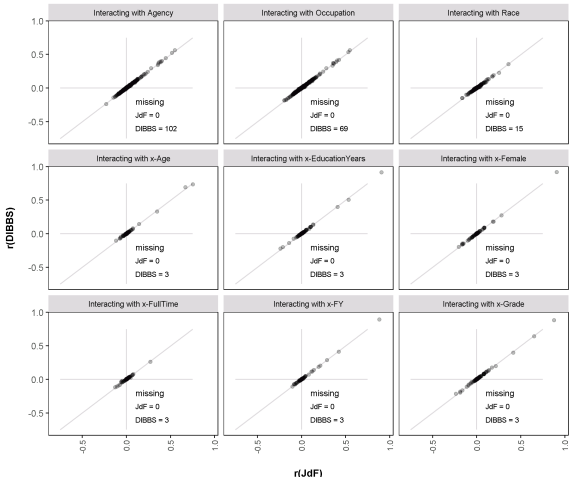
(b) Correlations involving Grade



(c) Correlations involving Age



(d) Correlations involving Education



(e) Correlations involving Work Schedule

Figure 3: Correlations of primary variables with two variable interactions. Variable set two. Synthetic level correlation on y-axis, corresponding authentic correlation on x-axis.

CUMULATIVE MASS (PROPORTION OBSERVATIONS) BY PAY PLAN AND OCCUPATION

Figures 4 and 5 contain example CMF plots of pay plan and occupation combinations. All occupations within each pay plan are represented. Solid line for authentic data, dashed line for synthetic. Overlapping or nearness of lines indicates equality of cumulative mass for corresponding levels of occupation within pay plan. "nJ" indicates observation count in authentic data, "nD" indicates synthetic data observation count. Near identical distribution is observed for high frequency pay plans GS, WG, GM, and VN, which account for more than 95% of observations, indicating overall close agreement between data sets. Increasing departure observed as number of observations decreases.

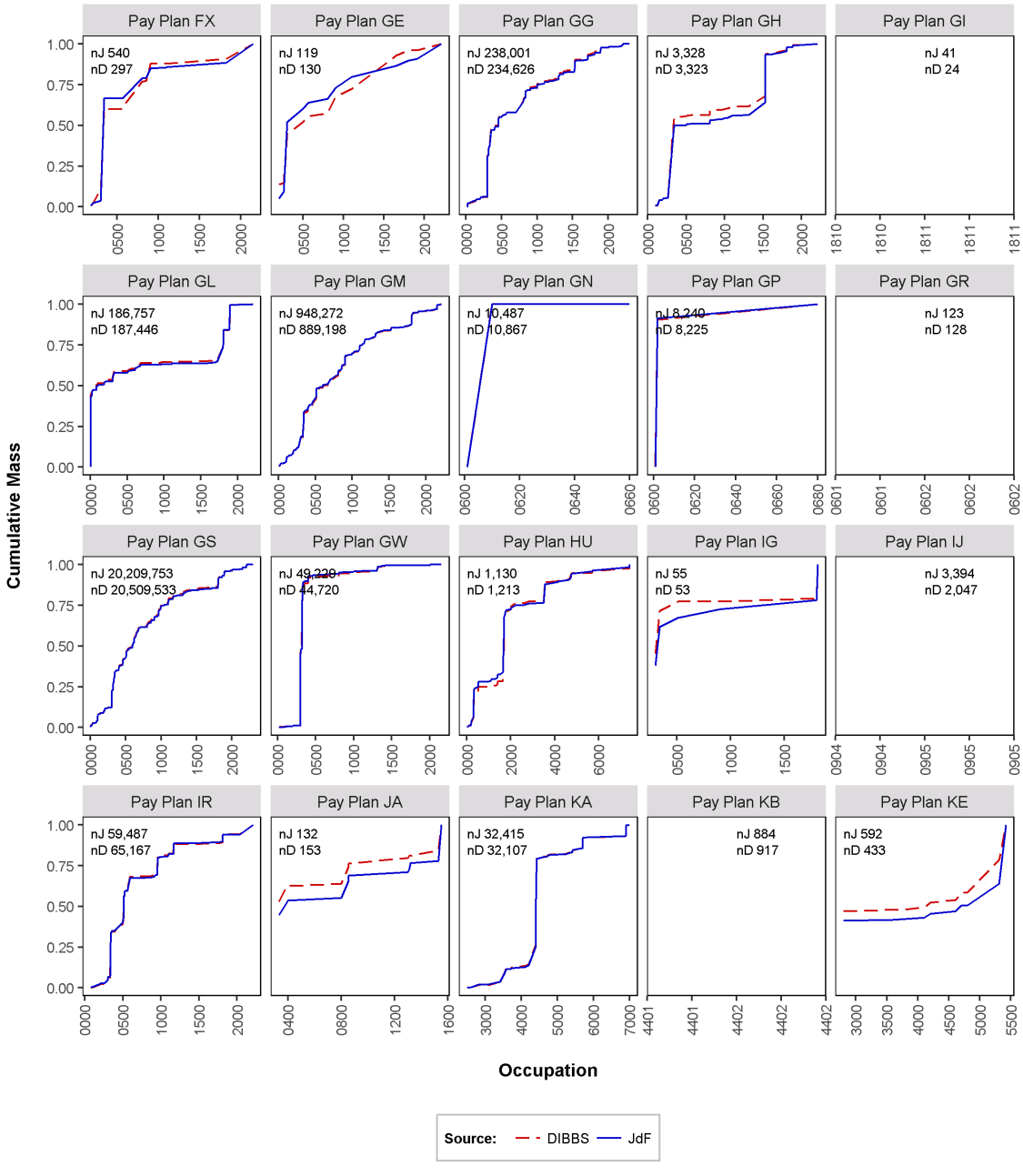


Figure 4: Cumulative mass by occupation within pay plan. Pay plan set one.

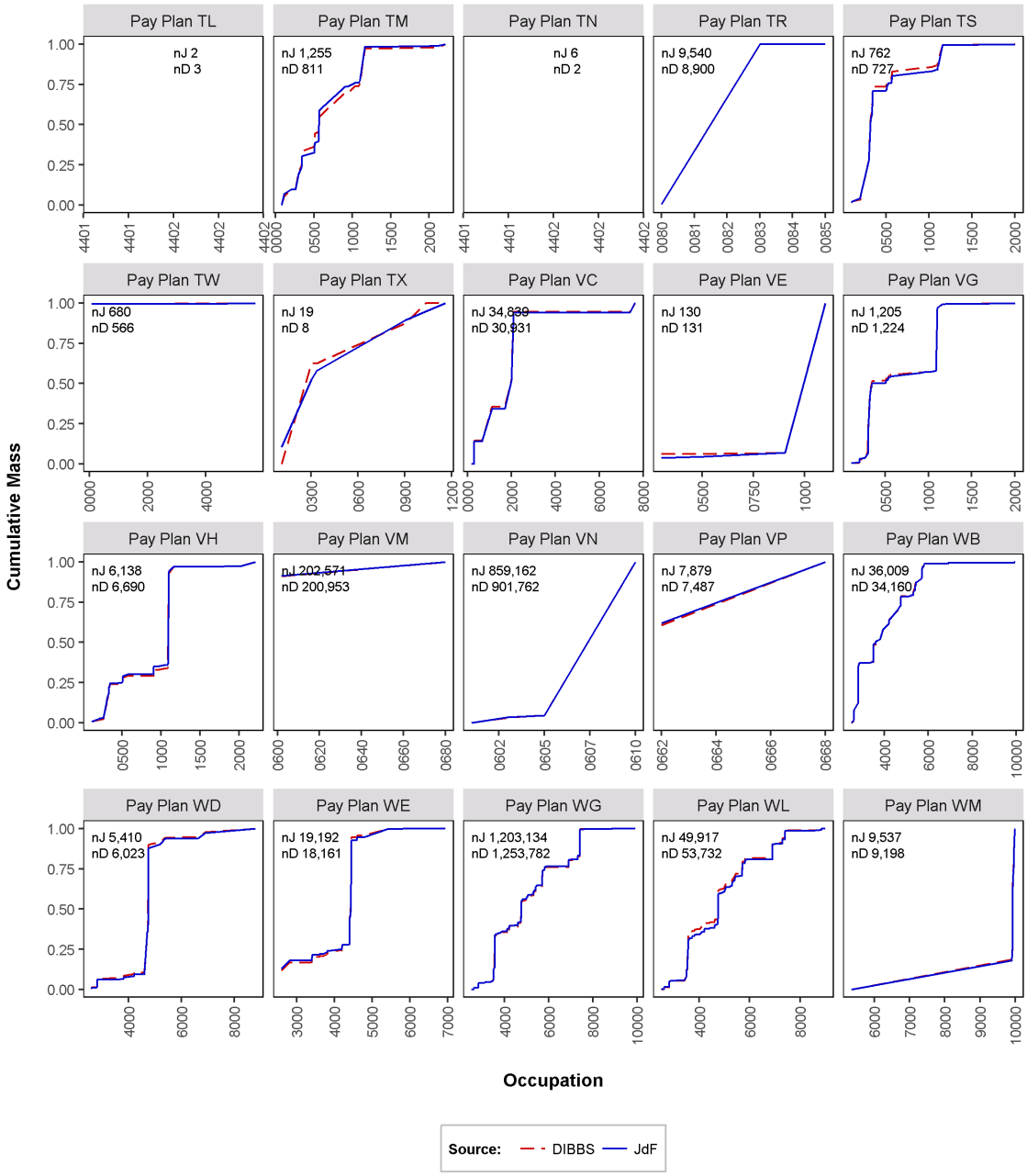


Figure 5: Cumulative mass by occupation within pay plan. Pay plan set 2.

MARGINAL DISTRIBUTION OF BASIC PAY

Basic pay is an important dependent study variable and the distribution of pay values in the authentic data must be maintained in the synthetic data. Figure 6 plots the distribution of basic pay for the top eight frequency agencies (first two positions): Department of Agriculture (AG), Department of Justice (DJ), Department of Health and Human Services (HE), Department of Homeland Security (HS), Department of Interior (IN), Department of Transportation (TD), Department of Treasury (TR), and the Department of Veterans Affairs. These agencies account for approximately 85% of observations. Synthetic distribution (dashed line) is overlayed on authentic distribution (solid line). "n(D)" indicates synthetic data observation frequency, "n(J)" indicates authentic data observation count.

Observations: Although one line appears in each graph for each data set, a single striped-appearing line is visible, due to identical frequency proportions at each pay level.

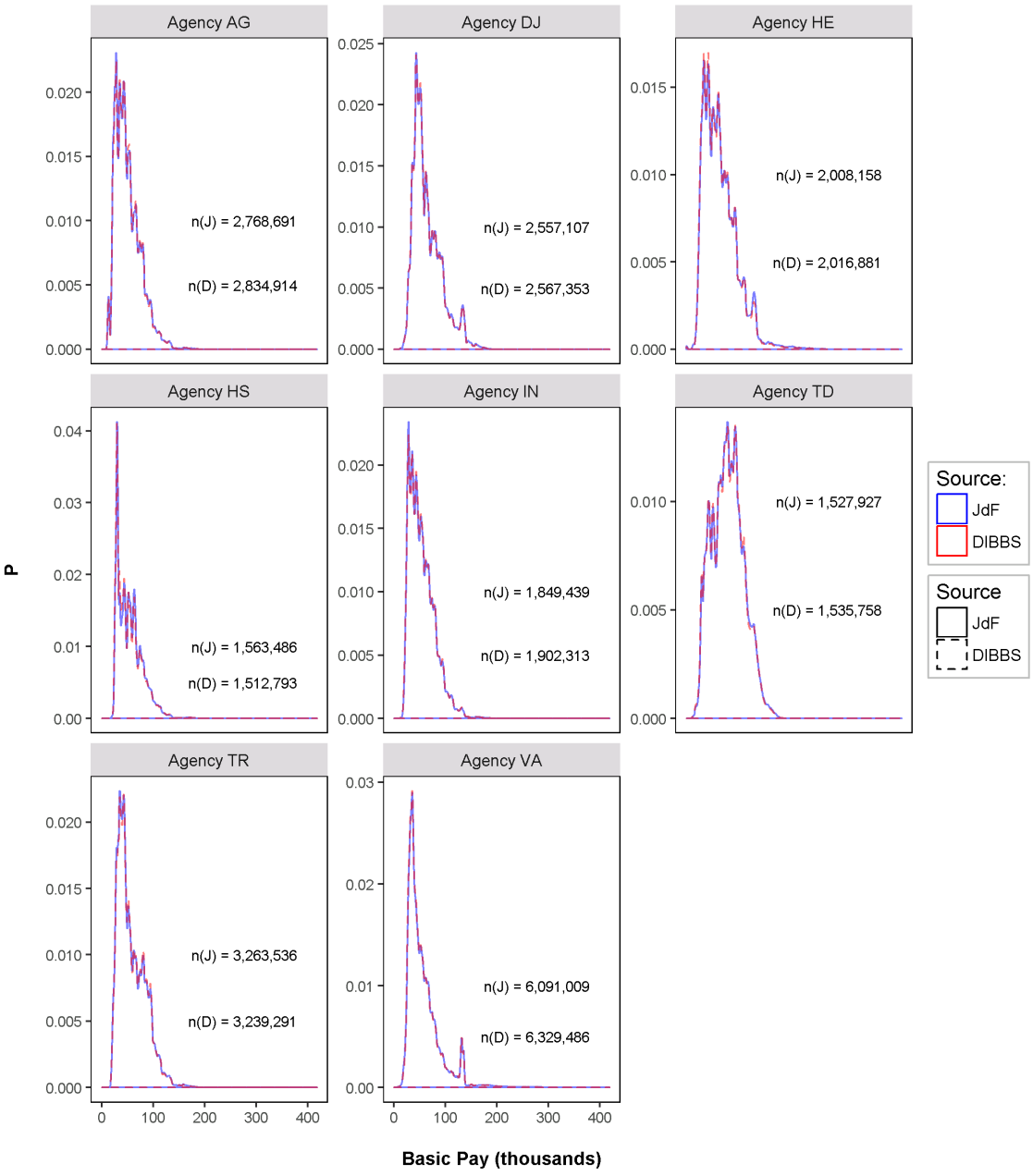


Figure 6: Basic pay marginal distribution for top eight agencies. Dashed line for synthetic data, solid line for authentic.

DISTRIBUTION OF BASIC PAY BY PROFESSIONAL, SUPERVISORY, COLLEGE EDUCATION, AND WORK SCHEDULE CATEGORY

Professional classification, supervisory status, and college education are important independent variables in human capital research. Figures 7 through 14 plot, for the top eight frequency agencies, the distribution of basic pay by these independent variables and work schedule code. One column for each professional, supervisory, college combination (column code position one equals "P" if occupational category is administrative or professional, position two equals "S" if supervisory status is enabled, position three equals "C" if education level at or above college). One row for each work schedule code [significant codes are full time (F), full time seasonal (G), intermittent (I), intermittent seasonal (J), and part time (P)]. Synthetic distribution indicated by dashed line, authentic distribution by solid line. "n(D)" indicates synthetic data observation frequency, "n(J)" indicates authentic data observation count.

Observations: There exists near identical distribution for high frequency combinations, as indicated by striped, single line appearance due to overlay of synthetic on authentic lines. Slight differences in distribution are observed for small frequency combinations.

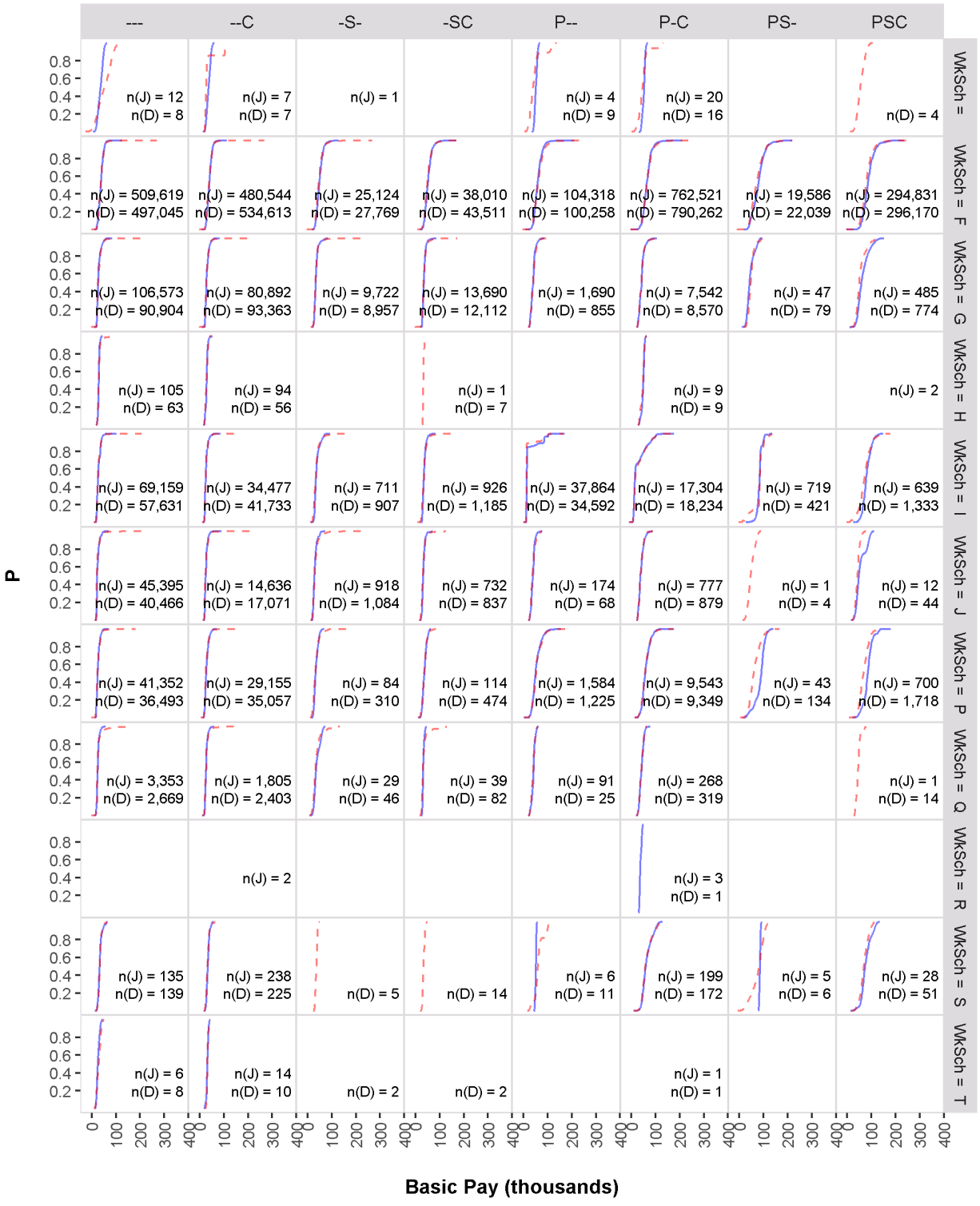


Figure 7: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Agriculture (AG). Dashed line for synthetic data, solid line for authentic.

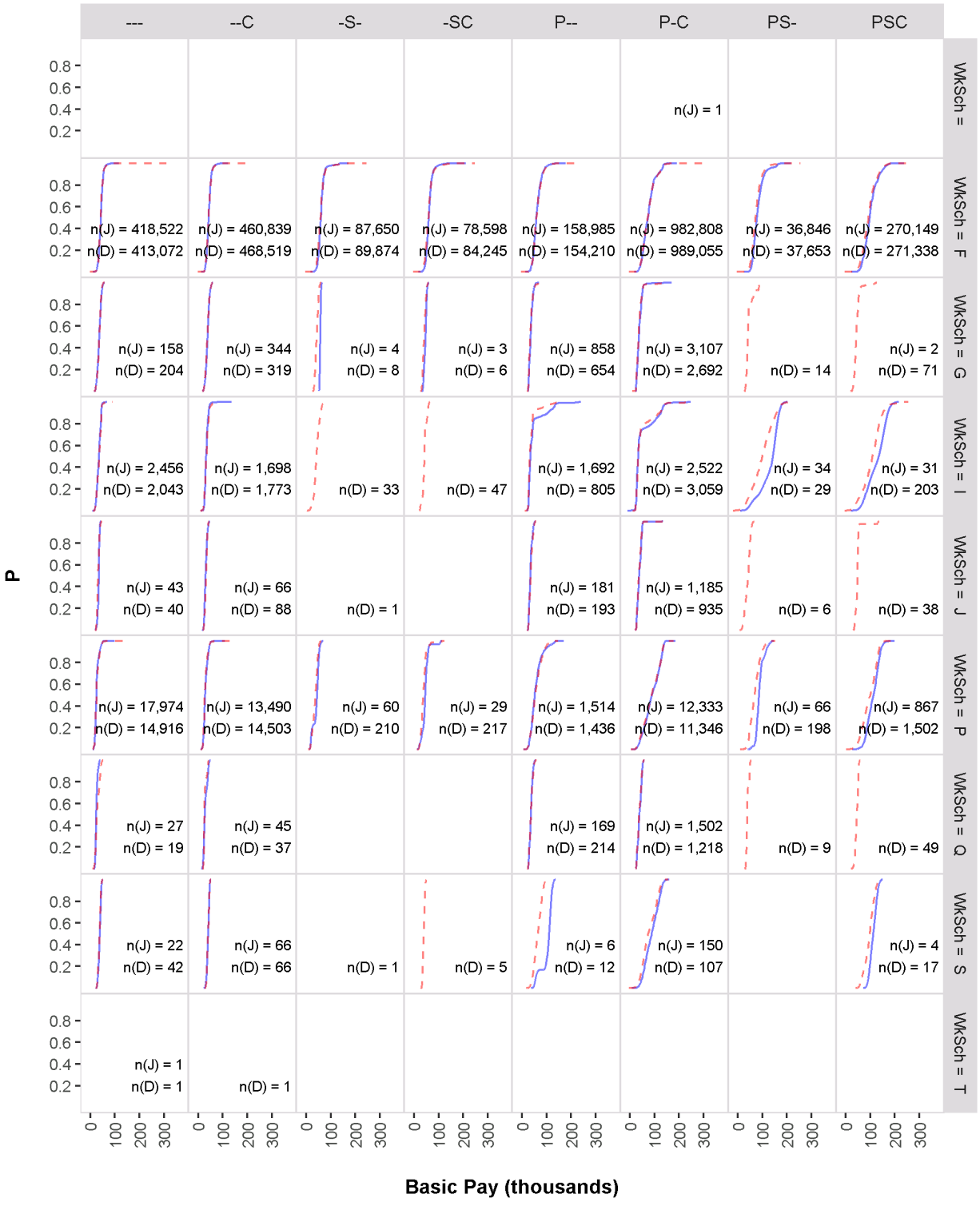


Figure 8: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Justice (DJ). Dashed line for synthetic data, solid line for authentic.

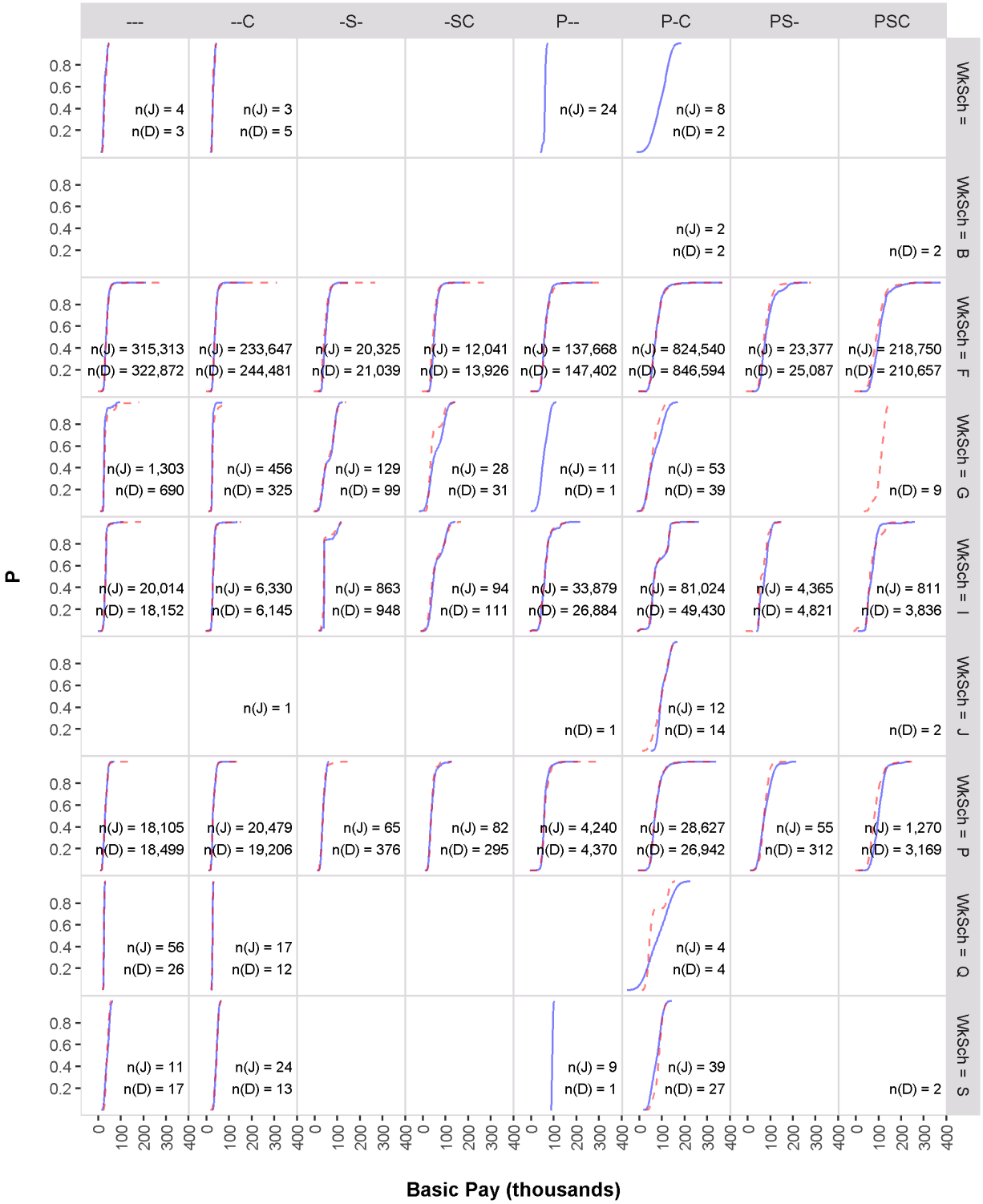


Figure 9: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Health and Human Services (HE). Dashed line for synthetic data, solid line for authentic.

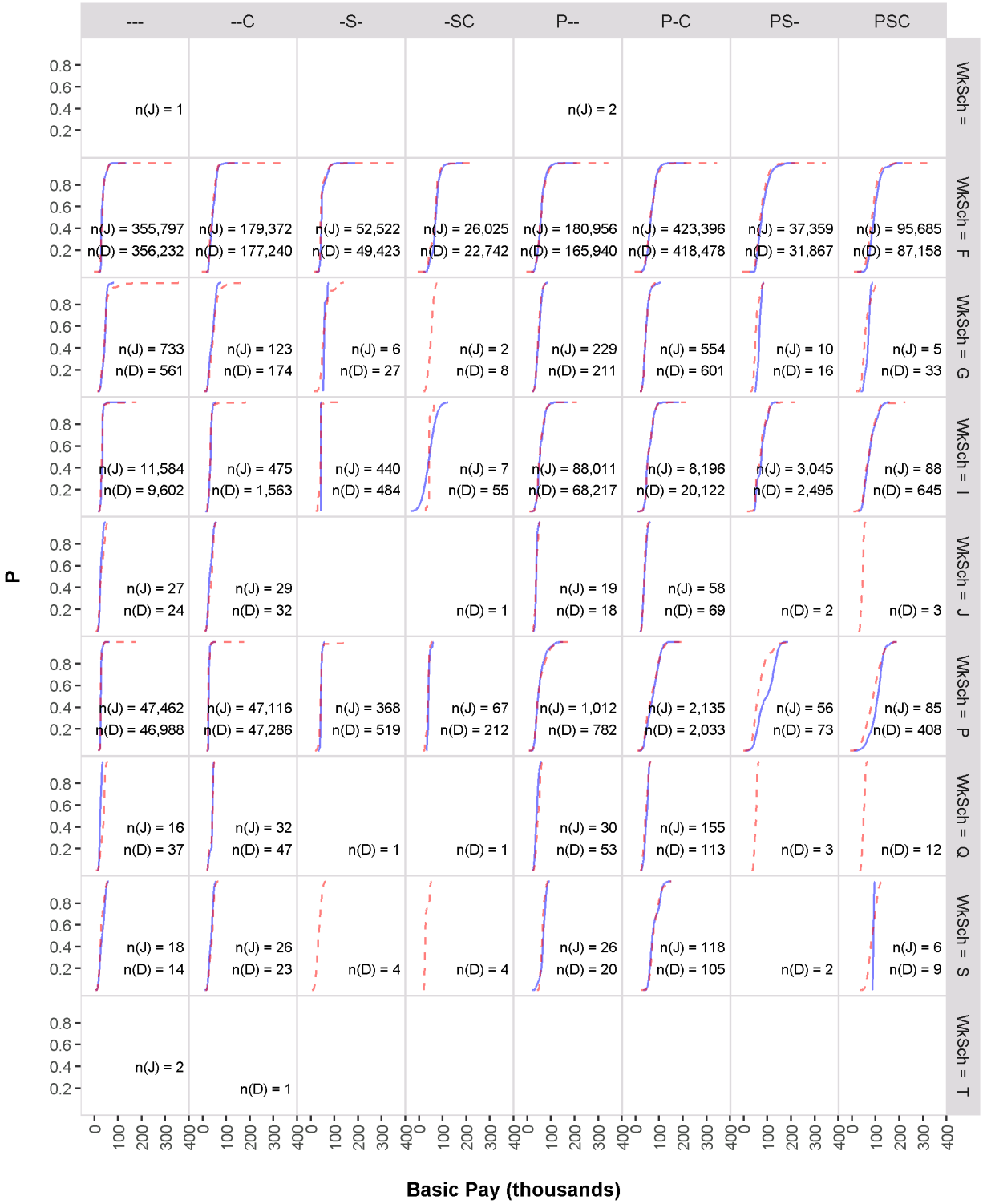


Figure 10: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Homeland Security (HS). Dashed line for synthetic data, solid line for authentic.

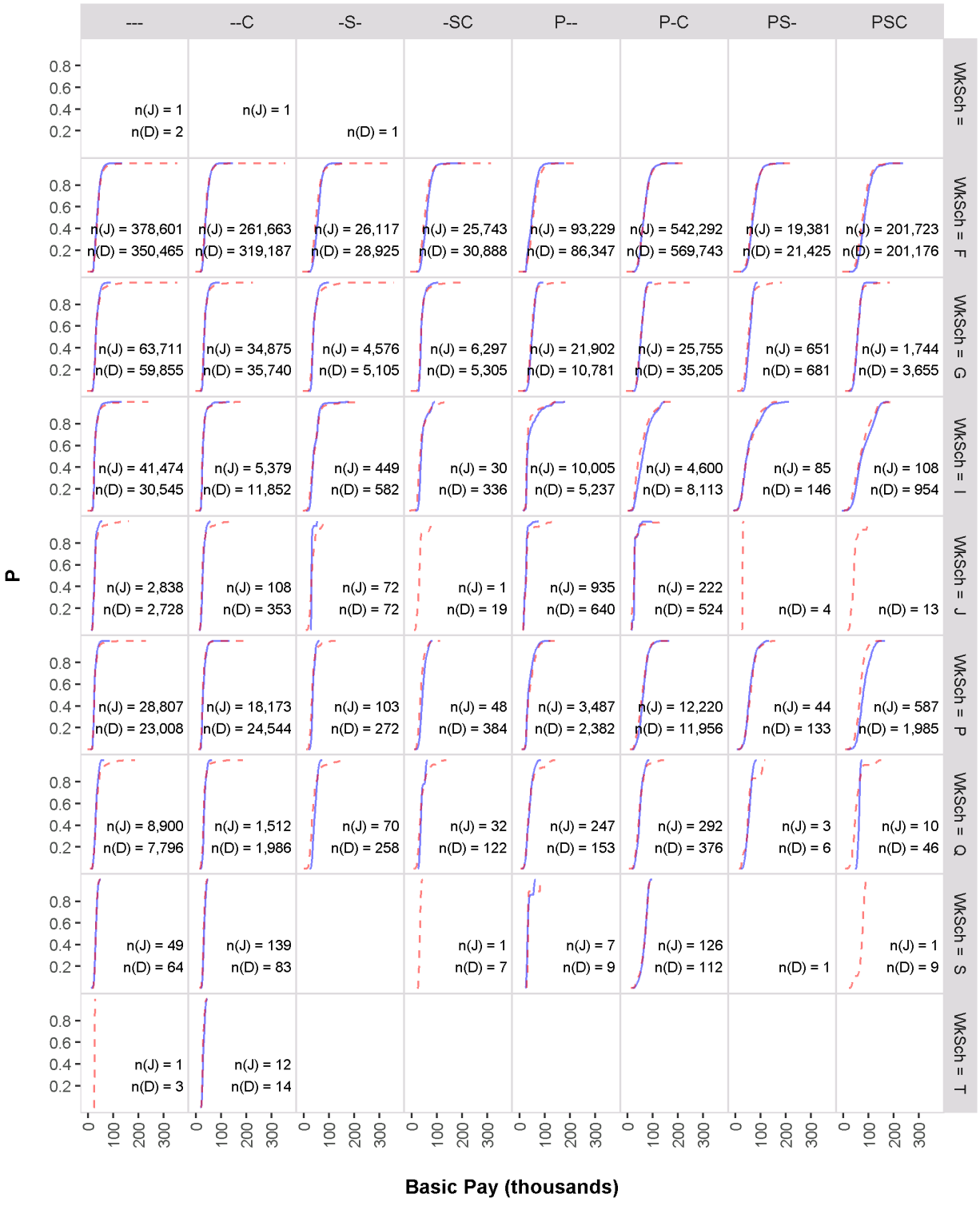


Figure 11: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Interior (IN). Dashed line for synthetic data, solid line for authentic.

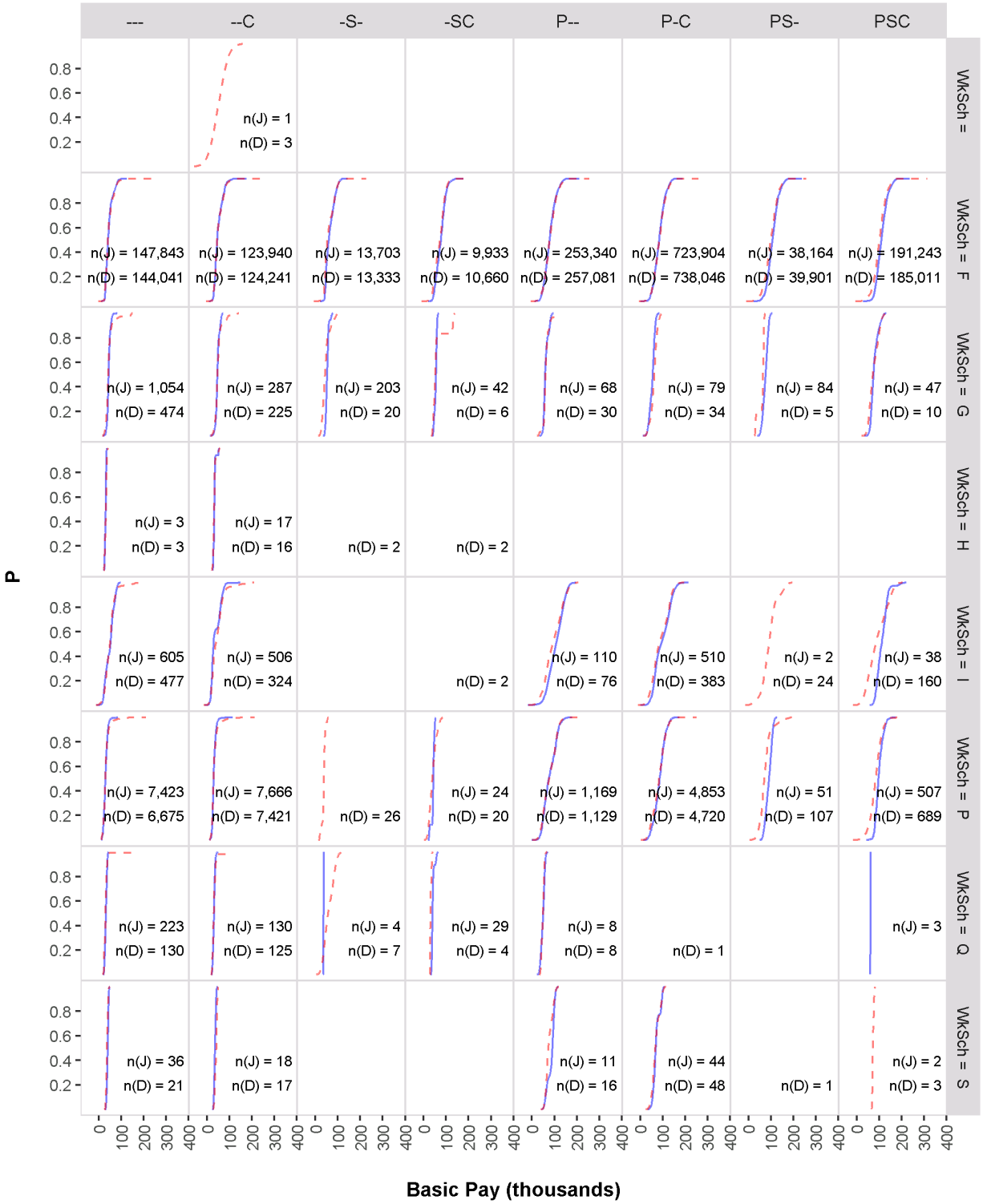


Figure 12: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Transportation (TD). Dashed line for synthetic data, solid line for authentic.

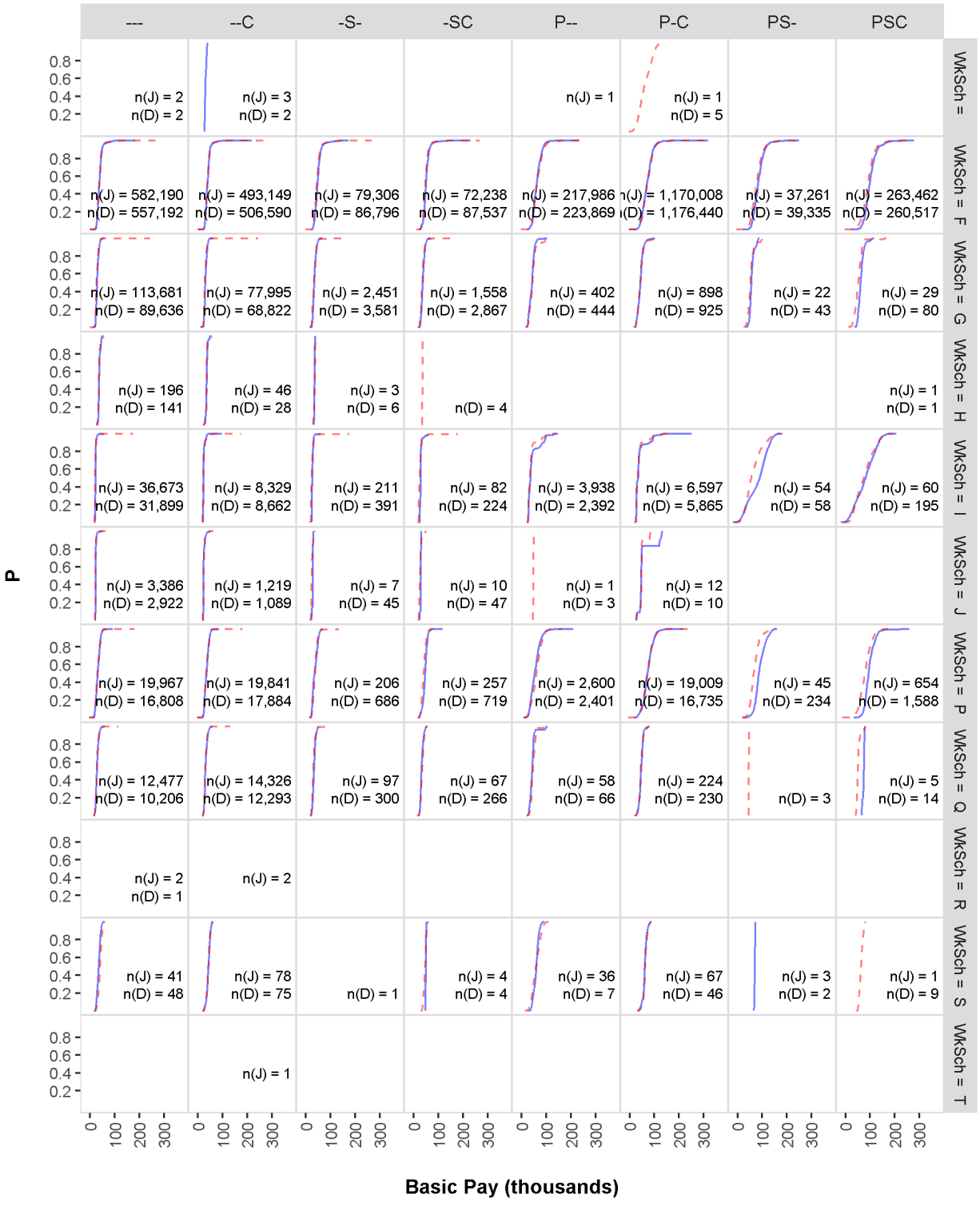


Figure 13: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Treasury (TR). Dashed line for synthetic data, solid line for authentic.

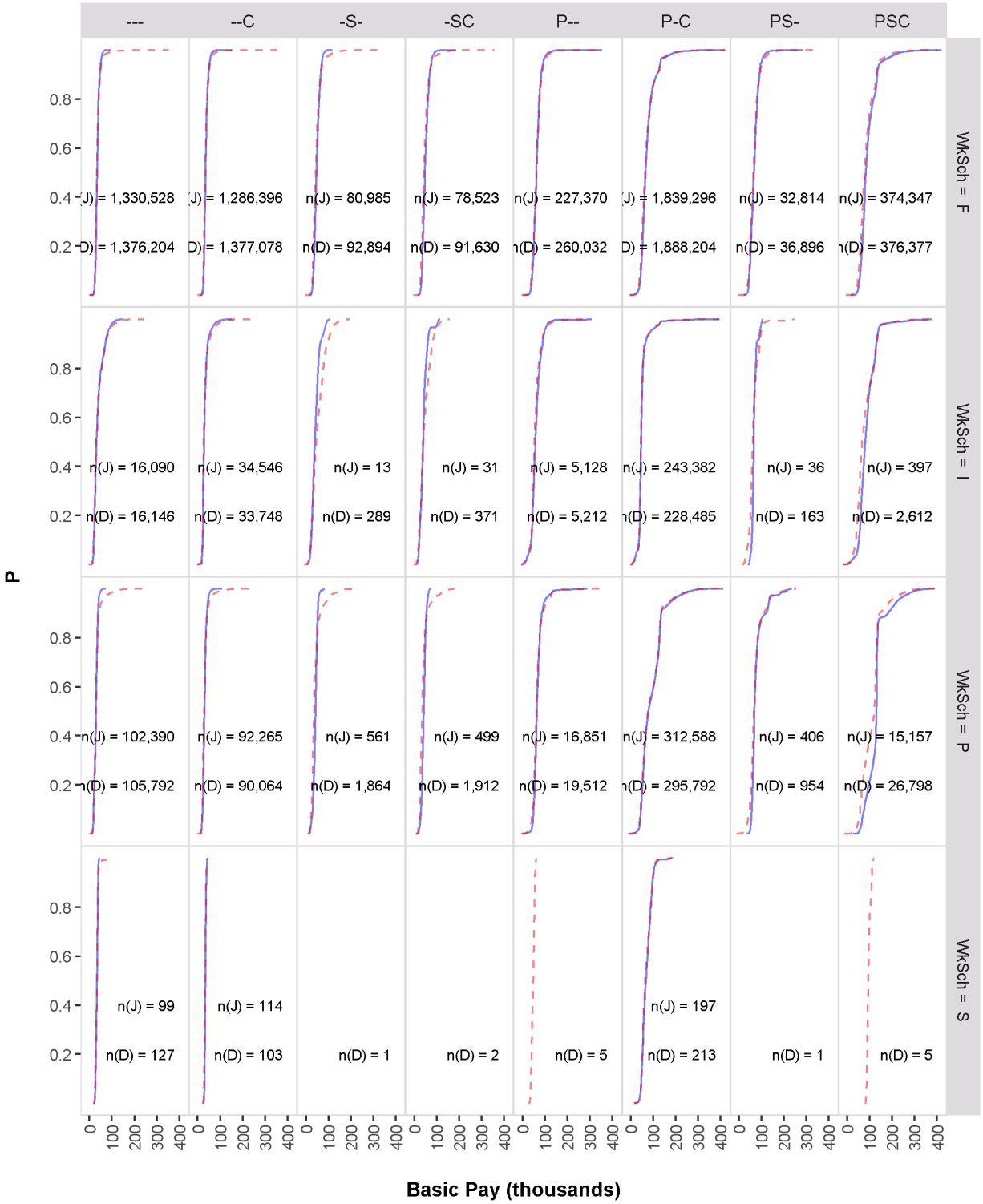


Figure 14: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Veterans Affairs (VA). Dashed line for synthetic data, solid line for authentic.

MEAN LOG(BASIC PAY) BY GENDER, RACE, AND YEAR

The relationship of mean basic basic pay to joint combinations of sex, race, and year is important in human capital research and must be maintained in the synthetic data. Figure 15 plots mean log(pay) by year for each sex and race. Dashed lines for synthetic data, solid lines for authentic.

Observation: Differentiating colors may not be visible, but apparent pairings of lines (dashed near solid following similar trends) form race pairs. Although some systematic difference appears between data sets, inter-year and overall trends are very similar.

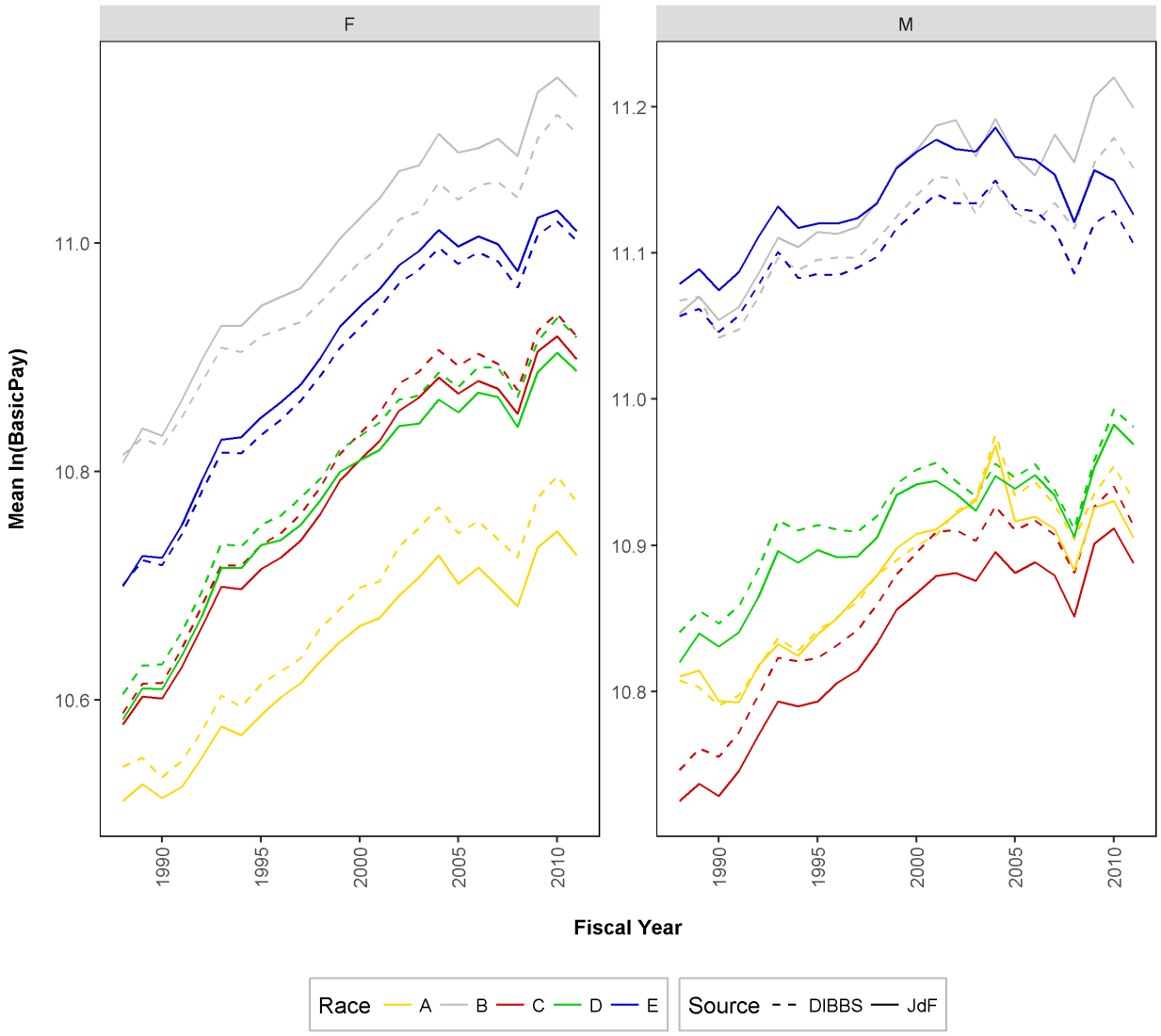


Figure 15: Mean basic pay by sex, race, and year. Left figure for female (F), right figure for male (M). Race codes: A = native American, B = Asian, C = black, D = Hispanic, E = white. Dashed line for synthetic data, solid line for authentic.

GENDER PAY DIFFERENTIAL FIXED EFFECTS QUANTILE REGRESSION MODEL

Figure 16 plots the 0.1, 0.5, and 0.9 quantile estimates of difference in log(basic pay) between federal employee women and men by year, controlled for race, age, education, agency, and occupation.

Observation: Although some systematic separation appears between data sets, trends indicate similar proportion change with respect to time.

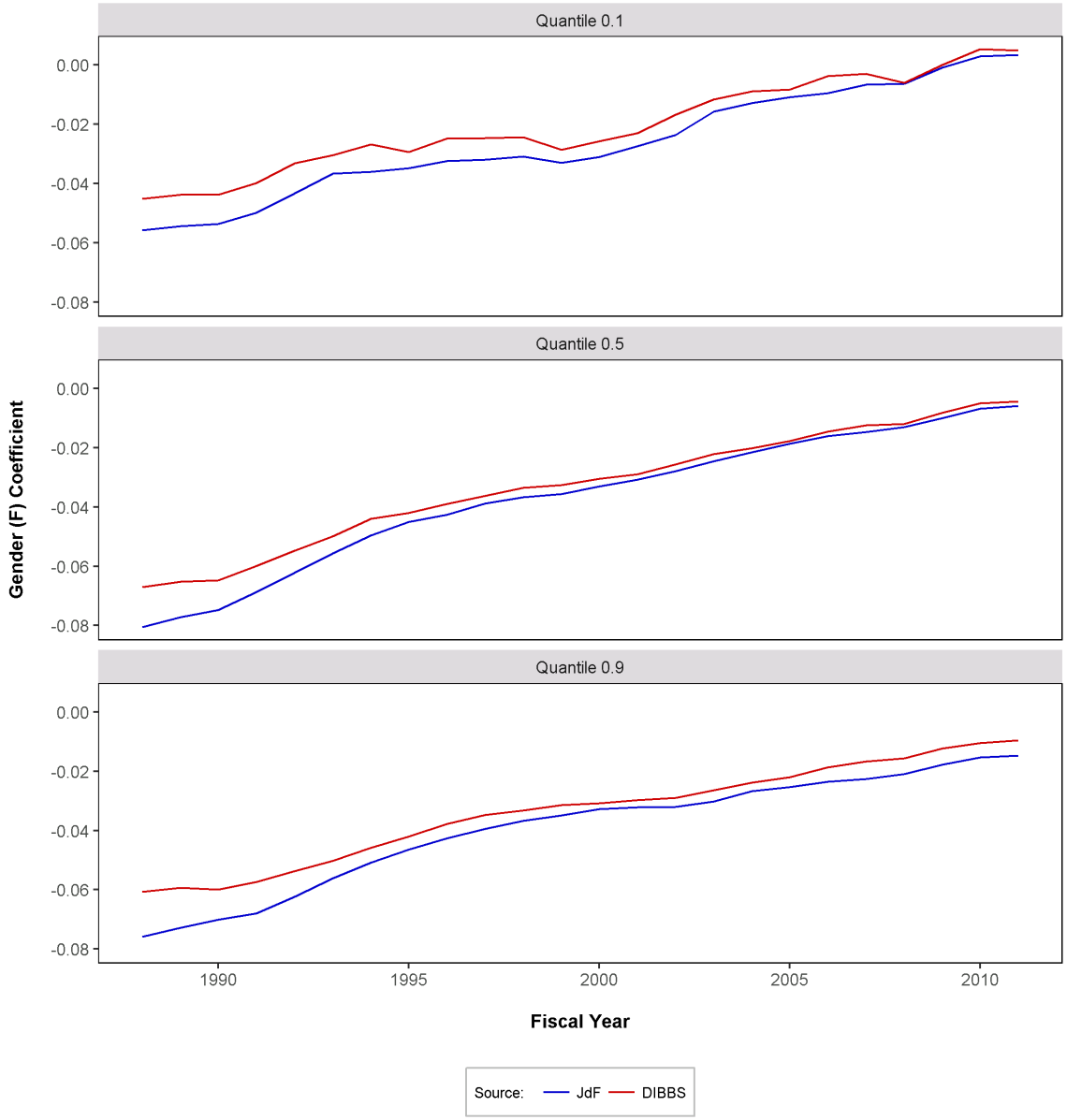


Figure 16: Quantile estimates from gender pay disparity fixed effects model. Upper line from synthetic data, lower line from authentic.

GENDER PROPORTION BY RACE, EDUCATION, AND YEAR

Proportion observations by gender is critical in models involving gender effects. Figure 17 plots proportion female employees by race, education, and year. Fitted lines are logistic regression models.

Observation: This four-way comparison (sex, race, education, and year) confirms good representation in synthetic data of gender proportion among important variable combinations in the authentic data. Fitted logistic regression models have nearly identical trends through fiscal years. Note the slight degradation in fit as observation count (n) decreases.

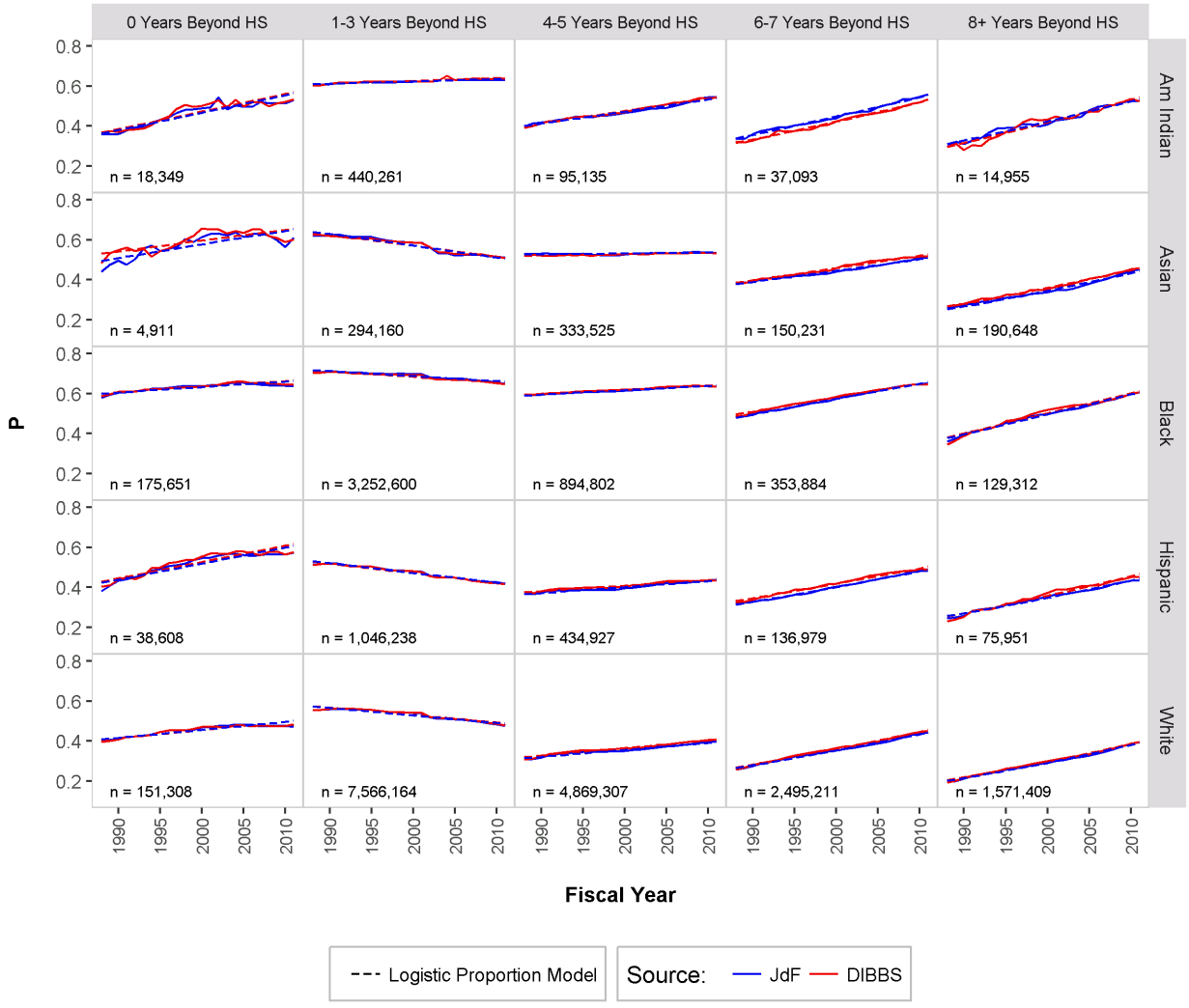


Figure 17: Proportion female observations by race, education, and year. Fitted lines are logistic regression estimates.

GENDER PROPORTION BY RACE, AGE, AND YEAR

Figures 18 through 22, show for each race, proportion female employees by age and year. Fitted lines are logistic regression models.

Observation: These four-way comparisons (sex, race, age, and year) confirm good representation in synthetic data of gender proportion among important variable combinations in the authentic data. Fitted logistic regression models have nearly identical trends through fiscal years.

Logistic models reveal agreement in trends between data sets.

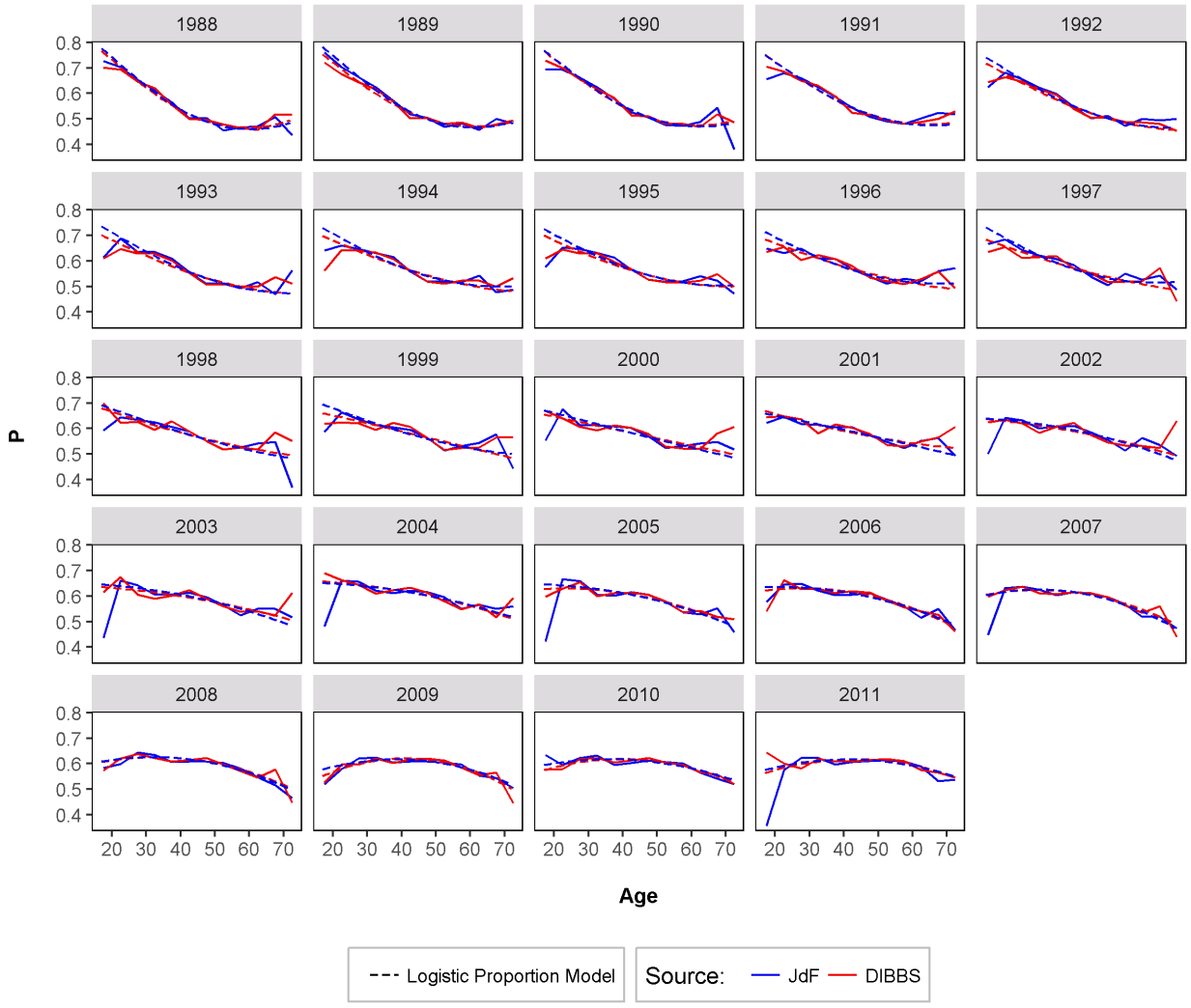


Figure 18: Proportion female observations by education and year. Race Native American. Fitted lines are logistic regression estimates.

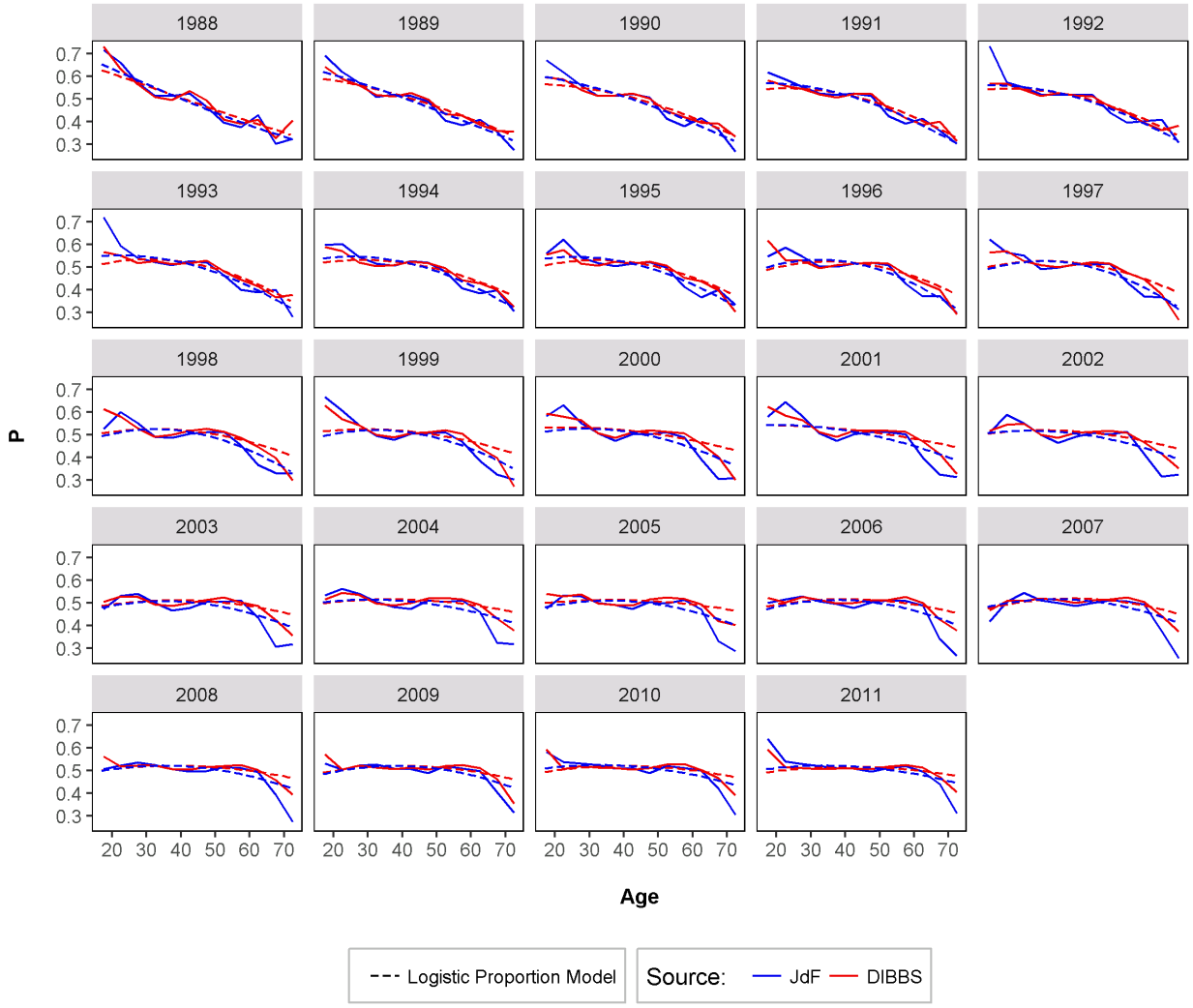


Figure 19: Proportion female observations by education and year. Race Asian. Fitted lines are logistic regression estimates.

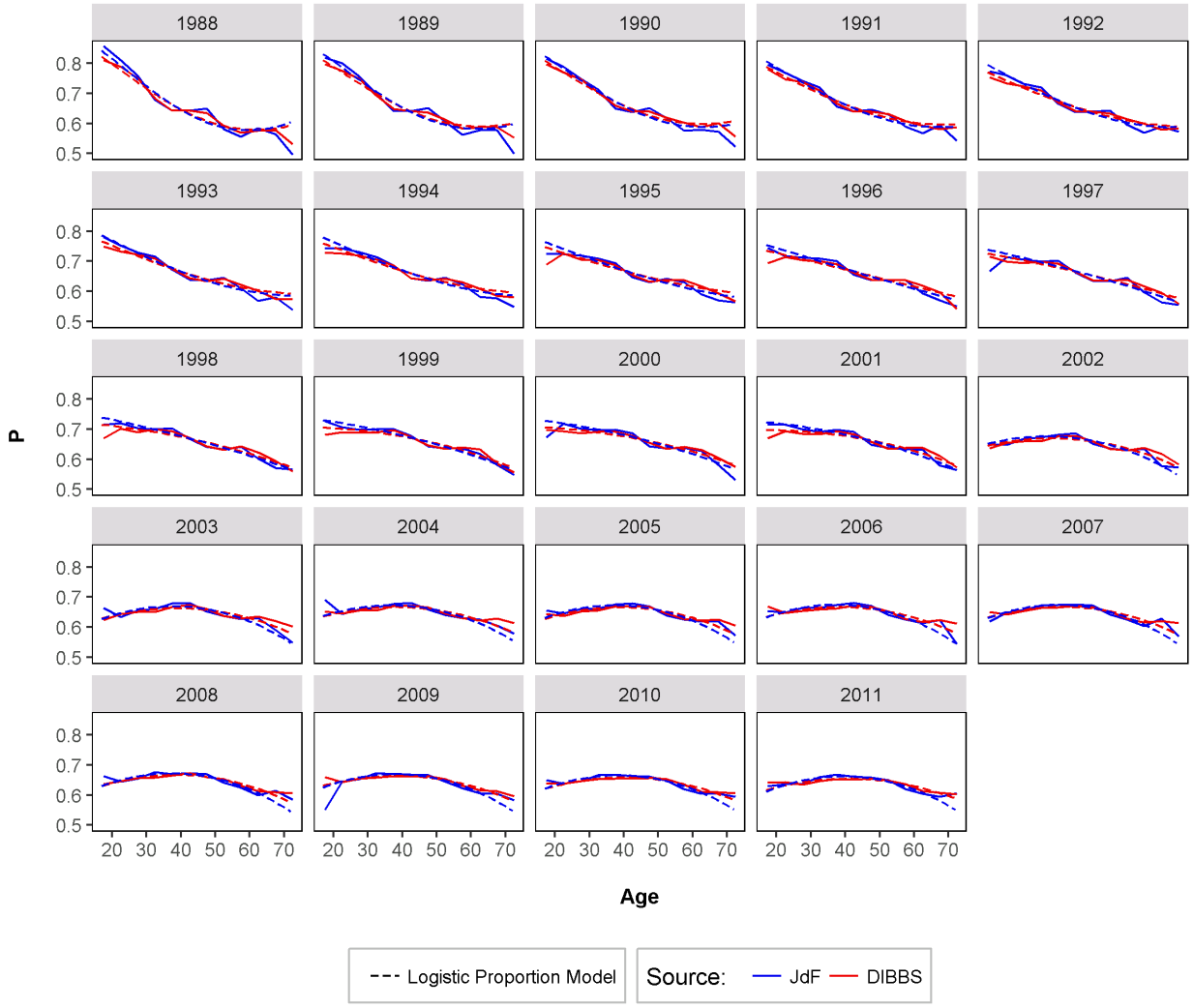


Figure 20: Proportion female observations by education and year. Race black. Fitted lines are logistic regression estimates.

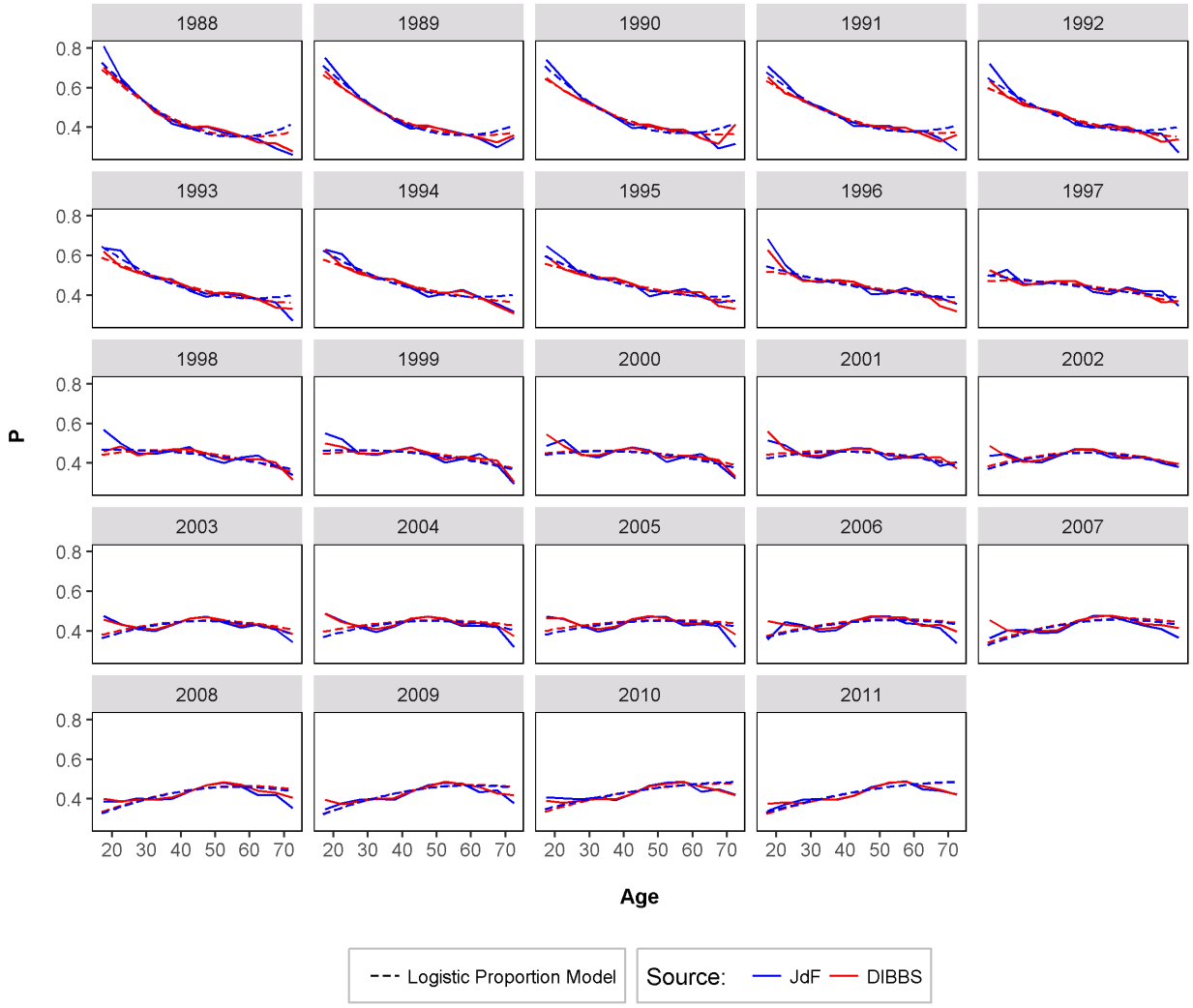


Figure 21: Proportion female observations by education and year. Race Hispanic. Fitted lines are logistic regression estimates.

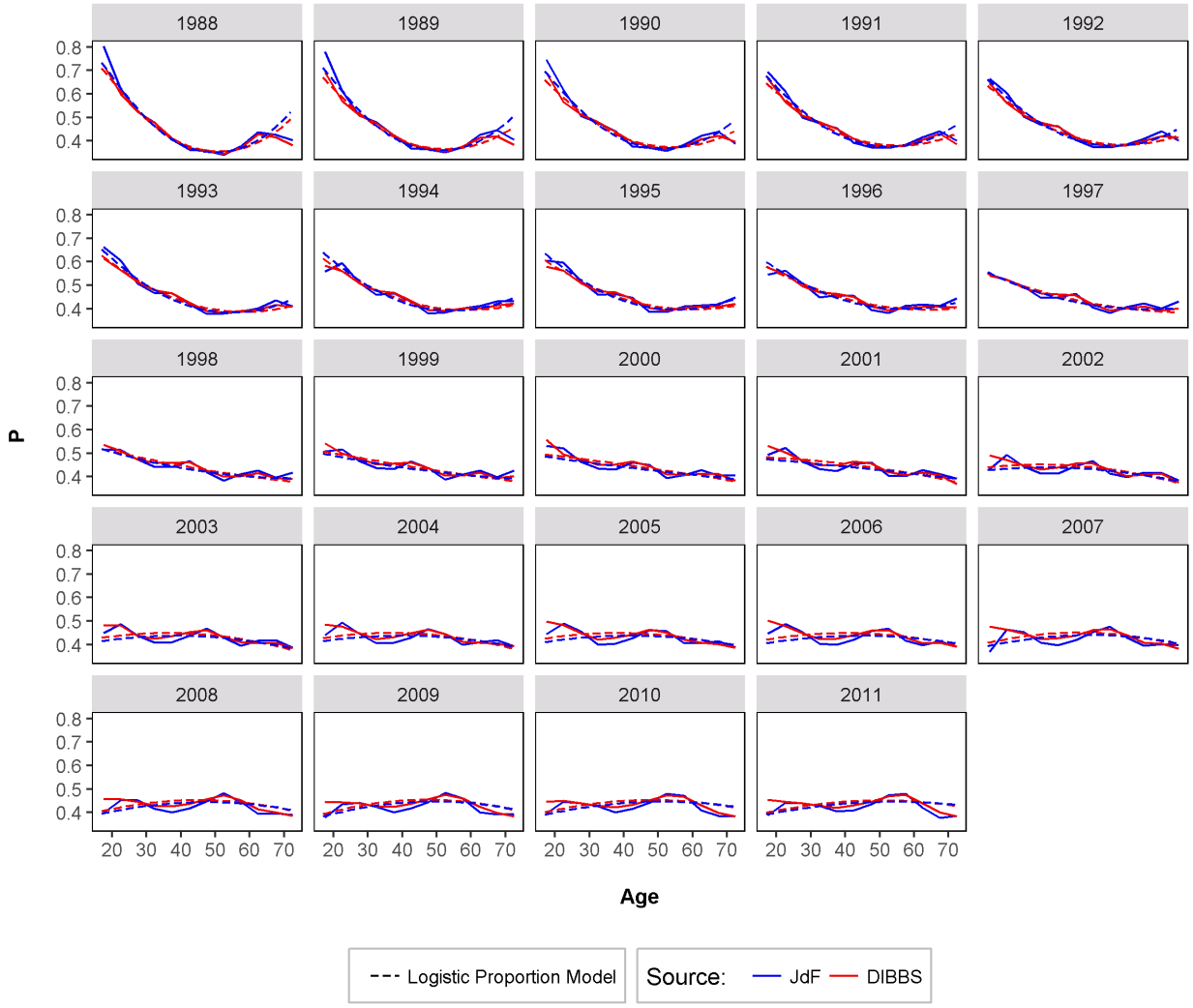


Figure 22: Proportion female observations by education and year. Race white. Fitted lines are logistic regression estimates.