GEMINI Guide

Martin R. Holmer
Policy Simulation Group

TEL 202-526-0406
NET martin.holmer@gmail.com
WEB www.polsim.com

February 2016

Preface

This is an introductory guide for policy analysts using GEMINI, a microsimulation model for distributional analysis of social security and individual accounts policy. The model was originally developed by the Policy Simulation Group during 2000 and 2001 without any financial support from other organizations. The guide has five main sections.

Section 1, *Surveying the Model*, provides an introduction to the model's overall structure and components.

Section 2, *Installing the Model*, describes the GEMINI installation procedure.

For *Running the Model*, which gives detailed instructions about how to operate the model, see Section 3 of the SSASIM Guide.

Section 3, *Validating Model Output*, discusses results of validation tests performed on GEMINI output results.

Section 4, *Analyzing Model Output*, discusses different methods of tabulating data contained in the GEMINI output results files.

Contents

1 Surveying the Model

GEMINI is a dynamic microsimulation model that enables policy analysts to conduct distributional analysis of policy changes in the Old Age and Survivors Insurance (OASI) and the Disability Insurance (DI) programs. In addition to current-law policy, the model can simulate the individual effects of a variety of incremental reforms to these defined-benefit programs, as well as the individual effects of a range of structural reforms that introduce a defined-contribution tier of self-managed individual accounts into social security.

A broad range of demographic and economic assumption parameters specify an environment. The environment can be represented as a *certain environment*, in which case it is simulated using one simulation *scenario*. Alternatively, an *uncertain environment* can be simulated by using Monte Carlo methods (Hammersley and Handscomb 1964) to sample many scenarios from dynamic probability distributions (i.e., stochastic processes) that represent the assumed expectations concerning the range of future values for the demographic and economic parameters.

GEMINI simulates the lifetime social security experience of a representative sample of people born in a specified year (i.e., a birth cohort), who live their lives in the context of the environment scenario(s) specified in the model run. The ability to simulate individual experiences under a representative sample of scenarios allows GEMINI to estimate the implications of an uncertain environment on people's lifetime social security experience.

A detailed set of specified policy parameters defines a *policy regime*, the implications of which are simulated in a *model run* that can assume either a certain or an uncertain environment. The consequences of a reform relative to current-law policy (or some other reform) are determined by comparing the output results from two model runs that assume the same environment: one run that specifies the reform policy regime and another run that specifies the current-law policy or some other reform. The model simulates the consequences of a policy regime for a specified cohort sample simulating all the years of the cohort individuals' lives. Comparing the consequences of two different policy regimes involves comparing the cohort sample distributions of several policy performance measures. This comparison will determine which sample members experienced an increase or decrease in a policy performance measure, and the magnitude of those reform-induced changes.

This first section of the guide describes the logical structure of GEMINI and provides information on the model's major features.

1.1 Modes of Operation

GEMINI can be operated as a free-standing model (as described below), or it can operate as a SSASIM add-on. When operating as an add-on, GEMINI is started automatically by SSASIM for one of two purposes. GEMINI can enable the SSASIM macro model to operate in the Over-Lapping Cohorts (OLC) mode or it can enable the SSASIM micro model to operate in the Representative Cohort Sample (RCS) mode. The SSASIM RCS mode is essentially the same as operating GEMINI as a free-standing model except that it provides options that maximize the use of available computer CPU chips and cores to speed the execution of several GEMINI runs. The SSASIM OLC mode requests GEMINI to produce samples for each cohort born after 1934 in order to build up aggregate payroll tax revenues and OASDI benefit expenditures for each calendar year, which are used by SSASIM to calculate standard trust-fund solvency statistics. In either mode, GEMINI operates with the same logic, but typically with smaller cohort sample sizes in OLC mode than in the RCS or stand-alone-model mode.

1.2 Model Architecture

The GEMINI model is designed to have a modular architecture. Modular architecture rejects the idea of the model as a single computer program that provides a large number of very different kinds of services: providing access to input data, calculating simulation results, and summarizing or visualizing output results. Instead, modular architecture calls for breaking up the model into a number of separate computer programs, each one of which specializes in providing one kind of service.

The main advantage of a model designed with a modular architecture is the increase in productivity that flows from the specialization allowed by the division of labor among component programs. Data-handling tasks can be performed by a program developed using a database management system that has superior data retrieval and ease-of-use features. Repetitive Monte Carlo calculations can be performed quickly by a compiled program that has been custom designed to do the necessary simulation computations. Statistical analysis and visualization of the simulated output can be done with statistical packages or custom-developed output analyzer programs.

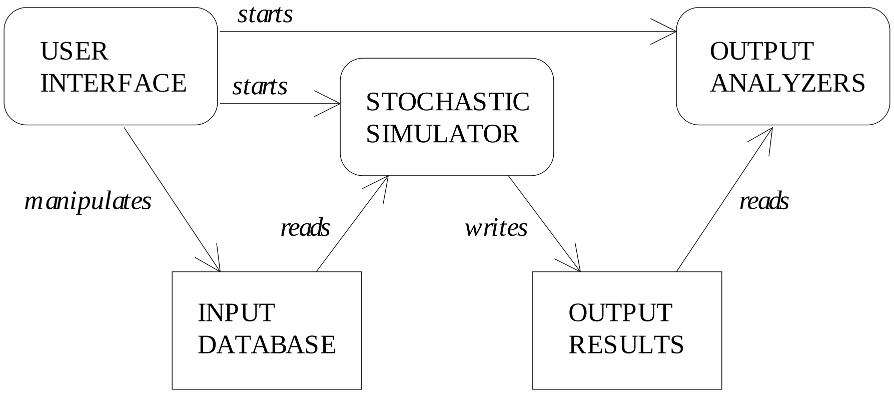

Figure 1: ***Modular Architecture of the GEMINI Model.*** *This architecture allows specialization among component programs and, if execution speed on a single computer chip is too slow, provides a development path to a multi-threaded architecture that permits distributed processing on multiple-chip computers.*

This modular approach goes well beyond the recommendations of a National Academy of Sciences panel that reviewed the design of policy simulation models (Citro and Hanushek 1991, Cotton and Sadowsky 1991).

The modular architecture of the GEMINI model is shown in Figure 1. The model consists of two sets of disk files and three separate kinds of computer programs, all of which are accessed from the model's user interface program. The two sets of disk files are: the input database tables, and several output results files for each model run, which consists of the simulation of one policy regime under one set of assumptions. The three computer programs are: a user interface, a stochastic simulator, and output analyzers.

1.2.1 User Interface Program

Read the *Getting Started with the RSF Toolkit* document, which is available at ⟨http://www.polsim.com/rsfhelp.pdf⟩, for a description of how to specify GEMINI (that is, SSASIM RCS mode) runs and execute the runs. To begin analyzing output results, start the EDA Toolkit from the RSF Toolkit Analyze menu and read *Getting Started with the EDA Toolkit* by pressing the Toolkit Help button.

1.2.2 Input Database Tables

The model's input database is organized as a relational database, a collection of linked database tables. Relational databases facilitate the organization of complex data on the attributes of and relationships between several different kinds of objects in a way that is intuitive, efficient, and flexible. The utility of relational principles for database design and implementation is widely appreciated. For an accessible introduction to relational databases by one of the original developers, see Date (1983).

The input database tables are stored as a single-file SQLite database, and the GEMINI installation contains a command-line tool (sqlite.exe) that can be used to manipulate the database in special cases where the standard methods are not adequate.

1.2.3 Stochastic Simulator Program

The GEMINI model is linked with two other models that produce information needed for a GEMINI simulation run. When GEMINI requests information from other models, they execute in a reduced-feature mode that generates only the requested information.

The GEMINI model's stochastic simulator program reads model input parameters from the GEMINI input database, reads cohort environment information from a SSASIM-generated environment file (envNNNNN, where NNNNN denotes the SSASIM run number specified in the PENSIM:RUN.ssasim_dir and PENSIM:RUN.ssasim_rid fields), reads cohort life history information from a PENSIM-generated cohort file (cohMMMMM, where MMMMM denotes the PENSIM run number specified in the GEMINI:SAMPLE.pensim_dir and GEMINI:SAMPLE.pensim_rid fields), then performs the requested run, and finally writes detailed information about the lifetime social security experience of each sampled individual to text files. The GEMINI output results files contain information about a sample of people born in the PENSIM:RUN.birth_year for the PENSIM run pointed to by the GEMINI:SAMPLE.pensim_dir and GEMINI:SAMPLE.pensim_rid fields. Figure 2 illustrates these model links.

It is important to realize PENSIM reads the envNNNNN file as input when it generates a cohMMMMM file for GEMINI. Note that if the required environment and cohort files do not exist, then GEMINI will request that

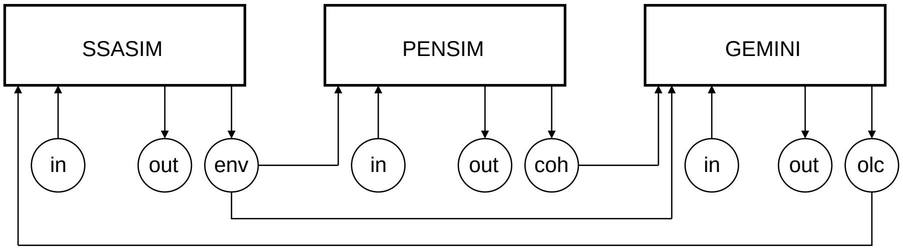

Figure 2: ***GEMINI Model Links.*** *Links between the GEMINI model and SSASIM and PENSIM.*

they be generated by SSASIM and/or PENSIM.

A highly significant implication of these model linkages is that if a user changes the specification of a SSASIM run that has already generated an envNNNNN file, it is essential for the user to delete that envNNNNN file. This is done automatically by pressing the Use button in the RSF Toolkit to respecify the run.

In an analogous manner, if a user changes the specification of a PENSIM run that has already generated a cohMMMMM file, it is essential for the user to delete that file or those files. This is done automatically by pressing the Use button in the RSF Toolkit to respecify the run.

The deletion of all envNNNNN and cohNNNNN.* files — note the same run numbers — plus all runNNNNN.* files can be accomplished quickly by using the `resclean` utility. For more information about how to control what is does, enter `resclean help` at the Windows command prompt. In particular, note the useful optional second parameter (saveOLCcoh) that permits saving the many OLC coh files (cohNNNNN.935, cohNNNNN.936, ...) while deleting the GEMINI coh file (cohNNNNN), the contents of which, unlike the OLC coh files, depends on the values of PENSIM:RUN.birth_year and PENSIM:RUN.sample_pct.

The GEMINI simulator processes one or more scheduled model runs automatically in a non-interactive manner. Ill-specified model runs generate detailed error messages in an error-log file and are aborted, but do not prevent the processing of other scheduled runs. A complete transcript of a long series of runs can be saved to a run-log file rather than being shown on the screen.

The stochastic simulator program is written in the C++ language and is organized in a modular fashion using C++ classes that correspond to real-world entities related to social security. The simulator is written using object-oriented and other modern programming styles (Coplien 1992, for example).

1.2.4 Output Results Files

Many of the GEMINI model's output results are written to text files that have been formatted to look like relational database tables. Some detailed cohort individual results are written to text files that have a compound relational structure. The output results files produced by the simulator during the course of each model run are labeled with the number of the run. This enables the storage of output results from many different runs and the comparison of results across runs long after the model runs were processed by the simulator. And because the specification of each run remains in the input database, the input parameters that generate the results for a run are completely documented.

The text output results files are quite large so the cannot be read by most text editors and cannot be imported into a spreadsheet program. See the below for a description of some output utility programs included in the GEMINI installation package, and see Section 4 for a more complete discussion of alternative methods of analyzing the GEMINI output results files.

1.2.5 Output Analyzer Programs

The model's output analyzer programs include any software program that is appropriate for the desired post-simulation analysis of the output results. Because the text output results are formatted in a manner that allows them to be easily imported into a wide variety of commercial software, there is enough flexibility in the selection of output analyzer programs to ensure that statistical analysis of the simulation output, output visualization, and preparation of presentation graphics are conducted with software that is both well suited to the task and familiar to the policy analyst using the model.

Although several output analyzer programs have been custom developed to facilitate routine analysis of output results, it should be noted that other options are available. Experience shows that one can never fully anticipate what post-simulation analysis must be done, and that unique, unexpected analysis needs arise routinely. The approach adopted here — writing detailed output and summary results from each model run to a set of text files, formatted in a way that allows them to be imported into a wide variety of commercial software — permits a flexible combination of anticipated analysis (using one of the existing output analyzer programs, for example) and unanticipated analysis (via custom-developed programs written in some statistical analysis language, for example).

1.3 Simulator's Internal Logic

The GEMINI stochastic simulator has OASDI-benefit-calculation capabilities identical to those of the SSASIM micro model except for one major difference. The SSASIM micro model simulates the social security experience of only a handful of exemplary individuals, while the GEMINI micro model simulates the social security experience of a large sample of individuals born in a specified year (that is, a representative sample of a birth cohort).

GEMINI relies on SSASIM to simulate the environment of the birth cohort and on PENSIM to simulate social-security-relevant life histories for each cohort sample individual and spouses. This means that GEMINI reads all the SSASIM input database tables except for the SSASIM:INDS table (and its child tables) and the SSASIM:COUPLES table (and its child table). Of course, GEMINI does read the SSASIM COHORT table to determine the policy regime being simulated (by reading COHORT.policy_id and the specified row(s) in the POLICY table and its child tables. However, the COHORT.birth_year value is ignored by GEMINI because the birth year of the simulated cohort is specified by PENSIM:RUN.birth_year for the PENSIM run pointed to by the GEMINI:SAMPLE.pensim_rid and GEMINI:SAMPLE.pensim_rid fields. And second, the input parameter setting in the SSASIM and PENSIM runs that support a GEMINI run must be carefully coordinated.

1.4 Simulator's Environment Input Files and SSASIM

GEMINI requires information on the demographic and economic environment of the specified birth cohort in order to estimate their lifetime social security experience. Such cohort input files are generated using SSASIM (Holmer SSASIM Guide), which permits specification of either certain (deterministic or one-scenario) environments or uncertain (stochastic or many-scenario) environments. GEMINI asks SSASIM to generate an environment input file on demand, but the request is not made if the appropriate file already exists.

It is important to realize that environment input files contain variables whose values depend on SSASIM policy parameters. These variables include the ERA, the NRA, and earnings-test policy parameters. This means that if a change in these policies are being simulated, a new environment input file needs to be generated by SSASIM. The environment input file also contains maximum taxable earnings, but this policy parameter is used only in employer-sponsored pension calculations.

1.5 Simulator's Cohort Input Files and PENSIM

GEMINI requires life histories for a large sample of birth cohort members in order to estimate their lifetime social security experience. Such cohort input files are generated using a restricted-use, no-pension version of PENSIM, which is a dynamic microsimulation model being developed for the U.S. Department of Labor to support policy analysis in the employer-sponsored pension area (Holmer et al. PENSIM Overview). PENSIM is distributed as part of GEMINI and it is used to generate the life histories stored in the GEMINI cohort input files. GEMINI asks PENSIM to generate a cohort input file on demand, but the request is not made if the appropriate file already exists.

PENSIM generates life histories for a representative sample individuals from a specified birth cohort and all their spouses, using an environment input file generated by SSASIM. These model linkages provide the capability of specifying a GEMINI run that uses either certain (one-scenario or deterministic) or uncertain (many-scenario or stochastic) cohort environments, with any reasonable cohort sample size. Total cohort samples (that is, the aggregation of the scenario samples) of 100,000 individuals plus their spouses are usually large enough to meet most policy analysis needs. But the model execution times are fast enough to make a total sample of 1,000,000 (an average sample size of 1,000 for each of 1,000 stochastic scenarios) a practical option.

The simulated life history information for each sample member is summarized in a binary record that is an element of a GEMINI cohort input file. Each individual record in a GEMINI cohort input file contains not only the individual's annual earnings history, but also information on gender, education, spouses (including their earnings histories and marriage termination events), children, disability, and age of death.

To generate these cohort samples with PENSIM, calibration targets are chosen for a range of statistics that describe life expectancy, educational attainment, employment pattern, earnings level, and marital status. In some cases these calibration targets may be known historical statistics (e.g., period life expectancy at birth for a cohort that has already been born). In other cases the calibration targets are chosen to be equal to forecasted values of the statistics (e.g., cohort life expectancy at age 65 as shown in a recent social security Trustees' Report).

After specifying the calibration targets for a particular birth cohort, PENSIM input parameters are adjusted so that when the resulting simulated cohort sample is tabulated, the value of the sample statistics match the calibration targets. In this manner, PENSIM can be used to generate cohort samples that reflect the demographic and economic differences between different birth cohorts.

The current version of PENSIM has the capability of generating cohort input files with any environment that can be simulated in SSASIM for any cohorts born in 1935 and subsequent years.

1.6 Model's Structure in Detail

The final part of this first section of the guide identifies all the input database tables and all the output results files associated with GEMINI. Brief descriptions of each table and complete descriptions of the output files are given.

1.6.1 Input Database Tables

Read the *Getting Started with the RSF Toolkit* document, which is available at ⟨http://www.polsim.com/rsfhelp.pdf⟩, for a description of the modern user interface program that uses run specification files and eliminates the need to work directly with the tables in the GEMINI input database.

The model's stochastic simulator program reads the input database tables in order to specify a model run. Each table contains an id number field — called a key in relational database terminology — that identifies table rows (or records, as the table rows are sometimes called). The tables are structured in a hierarchy in which the relevant row or group of rows in a child table is identified by the value of a table-specific row id number — called a foreign key — in the parent table. For more information about the logical structure of a relational database, see Date (1983).

Details on each table field can be found using the hypertext documentation available from the Help menu of the RSF Toolkit.

The names of the input database tables are presented below in the order that they are displayed in the documentation with the degree of indention representing their hierarchical level:

QUEUE Queue table rows contain the id numbers of the GEMINI runs that are scheduled to be processed by the stochastic simulator.

- **RUN** Run table rows point to the SSASIM run being used to specify the details of this GEMINI run and point to a row in the sample child table.
  - **SAMPLE** Sample table rows contain information about the birth cohort sample being used in this GEMINI run and also point to rows in child tables that provided more detailed specification of the run.
    - **TRACE** Trace table rows contain information about whether or not detailed social security calculations are show, and if so, for which scenario and sample individual.
    - **STATS** Statistics table rows contain information about which GEMINI output results files are to be written during the run.
    - **AAAF_RI** Annual actuarial adjustment factor table rows contain the assumed annual values of the OASI benefit adjustment factor.

1.6.2 Simulator Logic Modules

The model's stochastic simulator is composed of numerous modules or C++ classes. The source code for the each of the modules is spread across two files: a .h file, which contains the public interface to the module and defines the class's private data structures, and a .cpp file, which contains the module's private logic and algorithms.

The GEMINI logic modules are similar to those of SSASIM, which are described elsewhere (Holmer SSASIM Guide).

1.6.3 Output Results Files

The model's stochastic simulator program writes several large output files that describe different aspects of each sampled individual's lifetime social security experience.

The output results files written by the simulator for a model run all have the same file name, but have unique (three character) file-name extensions. The file name itself indicates the model run id number which is a positive integer with no more than five digits. This means that the results file name for run 306 is run00306, while run 2 would generate a results file name of run00002. If a particular results file had an extension of zzz, then the complete name of that results file would be run99999.zzz for the run with id 99999.

Documentation for each GEMINI output results file is included below.

Documentation of .adq Output File

```
DOCUMENTATION FOR TEXT OUTPUT RESULTS FILE .ADQ CREATED BY GEMINI

--- .ADQ file contains nominal pretax social security and pension
         benefits at each age for each cohort sample individual.

The GEMINI runNNNNN.adq file has no heading lines and no summary
lines.  The file contains individual-specific information about
each sampled individual in the cohort input file.  In addition,
there is age-specific information beginning as many as ten years
before the age of first OASDI benefit or private pension receipt
(or, the lesser of the age of death and 65, if that is before
the first benefit age) and continuing through the age of death.
Individual records, therefore, may contain different numbers of
lines.  The statistics on each line are separated by the tab
character.

The logical structure of an individual record is as follows:

IND LINE:   [there is an ind line for each sampled individual]
(1) "I" (without the quotes)
(2) scenar: scenario number
(3) ind_id: individual number (starts at 1 for each scenario)
(4) uncove: integer percent of nominal lifetime earnings that are uncovered
(5) sp_a_d: spouse's age difference in year indiv receives first benefit
            [sp_a_d = spouse's age - indiv's age (range is -9 to +9)]
(6) f__age: indiv's age on first age line (see age line info below)
(7) l__age: indiv's age on last age line (see age line info below)

AGE LINE:   [there is an age line for each age indicated on ind line]
            [age lines for indiv are arranged by ascending age]
            [dollar amounts are nominal thousands of dollars per year]
            [dollar amounts set to zero when less than or equal to 0.0005]
(1) indiv's current marital status expressed as one-digit code:
       0 ==> never married (marriage defined as lasting at least 0.5 year)
       1 ==> divorced (with no marriage lasting ten+ years so far)
       2 ==> divorced (with one+ marriage lasting ten+ years so far)
       3 ==> married
       4 ==> widowed (from most recent marriage)
(2) indiv's nominal pretax OASDI benefit ($K/year)
(3) indiv's largest-benefit-rule benefit type as one-digit code:
       [this differs from SSA primary-benefit-rule classification]
       0 ==> none
       1 ==> retired worker (may include kids benefits)
       2 ==> spouse of retired worker dually entitled
       3 ==> spouse of retired worker
       4 ==> widowed parent (may include kids benefits)
       5 ==> disabled widow (may include kids benefits)
       6 ==> aged widow
       7 ==> disabled worker (may include kids benefits)
       8 ==> spouse of disabled worker
       9 ==> only kids benefits
(4) couple's nominal pretax OASDI benefit ($K/year)
(5) indiv's nominal earnings ($K/year)
(6) couple's nominal earnings ($K/year)
(7) indiv's nominal pretax total employer-sponsored pension benefit ($K/year)
(8) couple's nominal pretax total employer-sponsored pension benefit ($K/year)
(9) family equivalence scale (adult equivalents in family at this age)
       NOTE: determined by family composition and specified values of
       the STATS.equivs_p, STATS.max_k_age, and STATS.equivs_f parameters.
(10) couple's nominal pretax DC pension/socsec benefit ($K/year)
       NOTE: (10) is part of (8) when STATS.adq_dc_pen=T or
             (10) is part of (4) when STATS.adq_dc_pen=F.
(11) indiv's nominal pretax DC pension/socsec benefit ($K/year) [part of (10)]
(12) current spouse's number of years married before marrying sample individual
       NOTE: zero if individual not married at this age; see (1) on this line
```

Documentation of .arc Output File

```
DOCUMENTATION FOR TEXT OUTPUT RESULTS FILE .ARC CREATED BY GEMINI

--- .ARC file contains revenue and cost statistics for social-
         security-account immediate-annuity provider.

The GEMINI runNNNNN.arc file has no header or footer lines.
There is one line for each cohort age for each scenario as well as one
summary line for each scenario.
The statistics on a line are separated by the tab character.

The arc statistics on each scenario AGE LINE are as follows:
(1) scenar: scenario number
(2)    age: age of individuals in simulated birth cohort
(3) ap_rev: annuity provider revenue from selling annuities at this age
(4) ap_cst: annuity provider cost from making annuity payments at this age
(5) ap_d_r: annuity provider discount rate at this age (in percent)

The arc statistics on each scenario SUMMARY LINE are as follows:
(1) scenar: scenario number
(2)         pv@65
(3) pv_rev: present value of annuity provider revenue from selling annuities
```

```
(4) pv_cst: present value of annuity provider cost from making annuity payments
(5) rc_rat: ratio of pv_rev to pv_cst for this scenario (pv_rev/pv_cst)

On all lines, rev and cst statistics, which have been inflated up to
population totals using the PENSIM:RUN.sample_pct, are expressed in
billions of SSASIM:COHORT.cpi_year dollars.

The present value statistics are computed using nominal cashflows and the
nominal discount rate (ap_d_r, which equals the nominal yield on Treasury
bonds), and then the nominal pv amount at age 65 is converted to real terms
so that it is also expressed in billions SSASIM:COHORT.cpi_year dollars.

NOTE: in a stochastic run that has more than one scenario, be sure
      that the all-scenario pv@65 revenue-to-cost ratio is at least
      one.  To compute the all-scenario ratio at the Windows command
      prompt in the directory containing the runNNNNN.arc file use
      the following two commands:
      > gawk "$2~/pv/" runNNNNN.arc | mean - 3
      > gawk "$2~/pv/" runNNNNN.arc | mean - 4
      The first command calculates the mean value across all scenarios
      of the revenue statistic (3), while the second command calculates
      the mean value across all scenarios of the cost statistic (4).
      The all-scenario ratio is simply the first mean divided by the
      second mean.
```

Documentation of .bAA Output File

```
DOCUMENTATION FOR TEXT OUTPUT RESULTS FILE .BAA CREATED BY OUT2BAGE UTILITY

--- .BAA file contains pretax benefits at age AA for each
         individual who is a living resident at age AA, where the
         statistics are created from the .sum and .adq output
         files by the out2bage utility, which is operated from
         the "GEMINI Data" menu of the EDA Toolkit or from the
         command line or in a tab_script.

The runNNNNN.bAA file has no heading lines and no summary lines.
There is one line for each individual who is a living resident at
the specified age AA.

NOTE: if AA < adq.f__age, then marital status, family equivalence
scale, and earnings statistics are unknown and have a value of u.
There will be u values only when the specified age AA is less than 55.

The statistics on a line are separated by the tab character.
The value of deflator is 1/cpi, where cpi is one when pretax benefits
are expressed in nominal terms and cpi is read from the runNNNNN.cpi
file for age AA when pretax benefits are expressed in real terms (i.e.,
in SSASIM:COHORT.cpi_year dollars).  See GEMINI documentation in the
out_sum.h, out_adq.h, and out_cpi.h files for more details.

NOTE: statistics that are exactly the same as in the .sum file are
      denoted as sum[n] where n is the .sum file statistic number;
      statistics from the age-line of the .adq file are denoted as
      adqa[n] where n is the .adq age-line statistic number.  Be sure
      to consult the documentation for the .sum and .adq statistics.
```

The statistics (numbered in parentheses) on each line are as follows:

```
// . . write runNNNNN.bAA file row with extracted statistics
long initben = 0; // this is NOT the initial OASDI benefit amount
if ( first_oasdi_benefit_age == age )
  initben = 1; // this age's benefit is the first OASDI benefit
float ssb_ind = convert( adqa[2] ) * deflator;
float ssb_cpl = convert( adqa[4] ) * deflator;
float ern_ind = convert( adqa[5] ) * deflator;
float ern_cpl = convert( adqa[6] ) * deflator;
float pnb_ind = convert( adqa[7] ) * deflator;
float pnb_cpl = convert( adqa[8] ) * deflator;
float feqivs  = convert( adqa[9] );
float pnc_cpl = convert( adqa[10] ) * deflator;
float pnc_ind = convert( adqa[11] ) * deflator;
long p_m_yrs = strtol( adqa[12], NULL, 10 );
fprintf( outfile,
         "%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s",
         sum[1],   // ( 1) scenar (see .sum doc for details)
         sum[2],   // ( 2) ind_id (see .sum doc for details)
         sum[3],   // ( 3) gender (see .sum doc for details)
         sum[4],   // ( 4) educat (see .sum doc for details)
         sum[5],   // ( 5) pvearn (see .sum doc for details)
         sum[24],  // ( 6) im_age (see .sum doc for details)
         sum[25],  // ( 7) em_age (see .sum doc for details)
         sum[16],  // ( 8) do_age (see .sum doc for details)
         sum[17],  // ( 9) dr_age (see .sum doc for details)
         sum[27] );// (10) fpvear (see .sum doc for details)
fprintf( outfile,
         "\t%s\t%s\t%ld\t%.3f\t%.3f\t%.3f\t%ld\t%.3f\t%.3f\t%.3f",
         adqa[1],  // (11) marital_status (u==>unknown) (see .adq doc)
         adqa[3],  // (12) ss_benefit_type (see .adq doc for details)
         initben,  // (13) is_first_ss_ben (0==>no;1==>yes)
         ssb_ind,  // (14) ss_ben_ind (above and .adq doc)
         ssb_cpl,  // (15) ss_ben_cpl (=ss_ben_ind+ss_ben_spouse) (")
         feqivs,   // (16) number of adult equivalents at bage (above)
         bage,     // (17) age in this cross-sectional sample
         pnb_ind,  // (18) total_pension_benefit_ind (above and .adq)
         pnb_cpl,  // (19) total_pension_benefit_cpl (=ind+spouse) (")
         ern_ind );// (20) earnings_ind (u==>unknown) (above and .adq)
fprintf( outfile,
         "\t%.3f\t%s\t%s\t%s\t%s\t%s\t%.3f\t%.3f\t%ld\t%.3f\t%.3f",
         ern_cpl,  // (21) earnings_cpl (u==>unknown)(=ind+spouse)
         sum[40],  // (22) siearn (see .sum doc for details)
         sum[41],  // (23) fsiern (see .sum doc for details)
         sum[43],  // (24) docage (see .sum doc for details)
         sum[44],  // (25) numqoc (see .sum doc for details)
         sum[12],  // (26) dieage (see .sum doc for details)
         dcb_cpl,  // (27) DC_pen/ss_benefit_cpl (=ind+spouse)(above)
         dcb_ind,  // (28) DC_pen/ss_benefit_ind [part of (27)]
         p_m_yrs );// (29) current spouse's prior marriage years
fprintf( outfile,
         "\t%s\t%s\n",
         sum[46],  // (30) siearx (see .sum doc for details)
         sum[47] );// (31) fsierx (see .sum doc for details)
```

Documentation of .bta Output File

```
DOCUMENTATION FOR TEXT OUTPUT RESULTS FILE .BTA CREATED BY GEMINI

--- .BTA file contains age-indexed pretax social security benefit
         and payroll tax aggregates for the cohort sample.

The GEMINI runNNNNN.bta file has no heading lines and no summary lines.
There is one line for each age for the only scenario in the run.
The statistics on a line are separated by the tab character.

The statistics on each line are as follows:
(1) birth year of cohort sample individuals
(2) age of cohort sample individuals [0,125]
(3) aggregate OASDI pretax benefits received by native-born cohort inds at age
(4) aggregate OASDI payroll taxes paid by native-born cohort individuals at age

Benefits and taxes of foreign-born individuals in the cohort sample are
not included in the aggregates.  Benefits and taxes are expressed in
thousands of nominal dollars per annum.  There is no inflation adjustment
to the benefit and tax aggregates.  Individual benefits and taxes (rather
than couple benefits and taxes that are adjusted by the adult-equivalent
size of the family) are used to tally the benefit and tax aggregates.
The aggregates are unweighted by the sampling weight implied by sample_pct.

IMPORTANT NOTE: be sure that each run that writes .BTA results is
simulating a solvent OASDI program over the life of the birth
cohort, otherwise the money's worth statistics computed from the
.BTA output results file using the Willingness-To-Pay CALCulator
(wtpcalc) will be meaningless.
```

Documentation of .bti Output File

```
DOCUMENTATION FOR TEXT OUTPUT RESULTS FILE .BTI CREATED BY GEMINI

--- .BTI file contains present value of lifetime earnings and
         present value of lifetime net social security benefits
         for each cohort sample individual included in .bta file.

The GEMINI runNNNNN.bti file has no heading lines and no summary lines.
There is one line for each native-born individual for the only scenario
in the run.  The statistics on a line are separated by the tab character.

The statistics on each line are as follows:
(1) ind_id: individual number [same as statistic (2) in .sum file]
(2) present value of adult-equivalent-adjusted couple earnings
(3) present value of OASDI pretax benefits minus OASDI payroll taxes

Individuals included in this .bti output results file are the same as
used to compile aggregate statistics in the .bta output results file.
Present values are for the birth year, are calculated using nominal
dollars and the constant annual discount rate specified by the
GEMINI:STATS.bti_drate parameter, and are expressed in thousands
of dollars per annum.  Individual benefits and taxes (rather than
couple benefits and taxes that are adjusted by the adult-equivalent
size of the family) are used to tally benefit-minus-tax present value.

IMPORTANT NOTE: be sure that each run that writes .BTI results is
```

```
simulating a solvent OASDI program over the life of the birth cohort.
Also, the .BTA output results file must be written whenever writing
the .BTI output results file.
```

Documentation of .cpi Output File

```
DOCUMENTATION FOR TEXT OUTPUT RESULTS FILE .CPI CREATED BY GEMINI

--- .CPI file contains consumer price index values by cohort age for
         each scenario.

The GEMINI runNNNNN.cpi file has no heading lines and no summary lines.
There is one line for each relevant cohort age for scenario in the run.
The statistics on a line are separated by the tab character.

The statistics on each line are as follows:
(1) scenario number
(2) cohort age (from 10 through 110)
(3) CPI value (with a value of about 1.0 in SSASIM:COHORT.cpi_year)

The CPI value will be exactly 1.0 for the age cohort individuals are in
the SSASIM:COHORT.cpi_year only if the simulated inflation rate over the
years between SSASIM:RUN.year_zero to SSASIM:COHORT.cpi_year is exactly
the same as the historical inflation rate documented in the SSASIM:HISTORY
table. With a positive inflation rate, the CPI value will rise above 1.0
as age increases beyond the cohort's age in SSASIM:COHORT.cpi_year.

To convert a nominal amount to a real amount (expressed in cpi_year
dollars), simply divide the nominal amount at age A by the CPI value
for age A.
```

Documentation of .scn Output File

```
DOCUMENTATION FOR TEXT OUTPUT RESULTS FILE .SCN CREATED BY GEMINI

--- .SCN file contains a variety of aggregate cohort sample statistics
         for each scenario.

The GEMINI runNNNNN.scn file has no heading lines and no summary lines.
There is one line for each scenario in the run.
The statistics on a line are separated by the tab character.

Definitions:
Cohort sample individuals who are foreign born or emigrate or have
  no retirement years (as defined below) are excluded from the
  group of individuals used to calculate .scn statistics.
Retirement begins at disability or retirement, whichever is first, and
  these events occur at the start of the year.
Retirement ends with death, which occurs at the end of the year.
So, for example, disability at age 57 and death at age 58 implies this
  individual has two retirement years.
Real pretax retirement income (RRI) in a retirement year is the total of
  individual's pretax OASDI benefit (adq:2 or share of adq:4) and
  ind's pretax employer-sponsored pension benefit (adq:7 or share of adq:8),
  where the total is converted from nominal to real terms by expressing it
  in thousands of SSASIM:COHORT.cpi_year dollars.
The value of STATS.scnsid_ind determines whether pretax retirement income
  is computed using the individual's benefits or the individual's adult-
  equivalent share of the couple benefits.  The nature of the sharing
  is controlled by the three STATS table parameters that are used in
  the National Academy of Sciences formula (see documentation of the
  GEMINI STATS table).
So, for example, if a scenario sample has 10,000 individuals who each have
  15 retirement years on average, then there will be 150,000 retirement
  years in the whole scenario sample, and there will be 150,000 RRI values
  for the whole scenario sample.
AWI denotes "average wage index", the mean annual earnings used in the wage
  indexing of various OASDI amounts, where AWI is converted from
  nominal to real terms by expressing it in thousands of
  SSASIM:COHORT.cpi_year dollars.  Note that the AWI in the retirement years
  of the cohort individuals refers to the average earnings of mostly younger,
  working-age people (not the cohort individuals whose retirement income is
  being measured).

The pretax scenario statistics on each line are as follows:
( 1) scenario number
( 2) aggregate number of retirement years in scenario sample
( 3) number of individuals in scenario sample with retirement years
( 4) mean real AWI (averaged over all retirement years for all scenario indivs)
( 5) mean RRI (averaged over all retirement years for all scenario individuals)
( 6) mean OASDI benefit (a component of mean RRI defined above)
( 7) mean pension benefit (sum of pension components of mean RRI defined above)
( 8) mean steady AWI-indexed earnings (sum:40 or sum:41 depending on scnsid_ind)
( 9) mean first retirement age (averaged over all scenario individuals)
(10) certainty-equivalent RRI over all scenario individuals (see method below)
(11) certainty-equivalent sben over all scenario individuals (see method below)
(12) certainty-equivalent pben over all scenario individuals (see method below)

So, (5)=(6)+(7) for each scenario, apart from rounding error.
But (10) is not necessarily equal to (11)+(12) because of method (see below).

Method for Calculating a Certainty-Equivalent Amount
As described in the documentation of the .sid output results file, a
certainty-equivalent amount is computed for each individual even if the
.sid output file is not being written.  Using the certainty-equivalent
amounts for each individual in a scenario, the following methods are used
to compute a certainty-equivalent amount for the scenario sample.
   Standard expected-utility-theory methods are used with the addition of
a special rule to handle the individuals with a zero certainty-equivalent
amount.  A constant relative risk aversion (CRRA) utility function is
assumed with the crra parameter value being set by the scnsid_ra field in
the STATS table of the GEMINI input database.  That utility function is used
to compute the certainty-equivalent amount for individuals with positive
certainty-equivalent amounts, and that scenario certainty-equivalent amount
is multiplied by the proportion of scenario individuals who have positive
certainty-equivalent amounts.
```

Documentation of .sid Output File

```
DOCUMENTATION FOR TEXT OUTPUT RESULTS FILE .SID CREATED BY GEMINI
```

```
--- .SID file contains individual statistics for each scenario statistic
         in .scn file.

Read the documentation of the GEMINI runNNNNN.scn file before reading below.

The GEMINI runNNNNN.sid file has no heading lines and no summary lines.
There is one line for each individual tabulated in the .scn statistics.
The individual statistics on a line are separated by the tab character.

Definitions:
The individual statistics included in the .sid file for a scenario should
  add up (using retirement years as the weights for dollar amounts) to the
  aggregate scenario statistics in the .scn file.
Consult documentation of .scn and .adq and .sum output files for more details.

The individual statistics on each line are as follows:
( 1) scenario number
( 2) gender: 0 ==> male, 1 ==> female
( 3) number of retirement years
( 4) mean real AWI (averaged over all retirement years for individual)
( 5) mean pretax RRI (averaged over all retirement years for individual)
( 6) mean pretax OASDI benefit (a component of mean RRI)
( 7) mean pretax pension benefit (a component of mean RRI)
( 8) steady indexed earnings (sum:40 or sum:41 depending on scnsid_ind)
( 9) first age in retirement-years range (as defined in the .scn output file)
(10) certainty-equivalent pretax RRI (see note below on method)
(11) certainty-equivalent pretax OASDI benefit (see note below on method)
(12) certainty-equivalent pretax pension benefit (see note below on method)
(13) pretax RRI at last retirement age (i.e., in final year of life)
(14) pretax OASDI benefit at last retirement age (i.e., in final year of life)
(15) pretax pension benefit at last retirement age (i.e., in final year of life)
(16) pension rollover account balance at death when have no surviving spouse
     Note that (16) equals (sum:20) when (sum:22) equals zero
                                         (i.e., there is no surviving spouse)
           and (16) equals zero     when (sum:22) equals one
                                         (i.e., there is a surviving spouse)
     where balance is expressed in thousands of SSASIM:COHORT.cpi_year dollars.
(17) number of retirement years in which RRI is zero

So, (5)=(6)+(7) for each individual, apart from rounding error.
Also, (13)=(14)+(15) for each individual, apart from rounding error.
But (10) is not necessarily equal to (11)+(12) because of method (see below).

Method for Calculating a Certainty-Equivalent Amount
Standard expected-utility-theory methods are used with the addition of a
special rule to handle the retirement years with zero amounts.  A constant
relative risk aversion (CRRA) utility function is assumed with the crra
parameter value being set by the scnsid_ra field in the STATS table of the
GEMINI input database.  The certainty-equivalent amount for years in which
there are positive amounts is computed as the certain amount whose utility
is equal to the expected utility of the annual positive amounts.  If there
are years with zero amounts, the positive-years certainty-equivalent amount
is multiplied by the ratio of the number of positive-amount years to the
total number of retirement years.
```

Documentation of .sum Output File

```
DOCUMENTATION FOR TEXT OUTPUT RESULTS FILE .SUM CREATED BY GEMINI

--- .SUM file contains lifetime statistics for each individual in
         cohort sample.

The GEMINI runNNNNN.sum file has no heading lines and no summary lines.
There is one line for each sampled individual in the cohort input file.
The statistics on a line are separated by the tab character.

NOTE: annual pretax benefits used to compute lifetime present values are not
      rounded down to zero when below 0.0005 thousand dollars per year.

The summary statistics on each line are as follows:
(***see SSASIM documentation, where same names are used, for more details***)
( 1) scenar: scenario number
( 2) ind_id: individual number (starts at 1 for each scenario)
( 3) gender: 0 ==> male, 1 ==> female
( 4) educat: 0 ==> no high school (died before entering high school)
             1 ==> high school dropout;    2 ==> high school graduate
             3 ==> attended some college;  4 ==> four-year college degree or
                                               graduate degree
Lifetime PV measures for OASI program [.=r] and DI program [.=d] and pensions:
    - all earnings and benefits are pretax (i.e., before federal income tax)
    - monetary measures expressed in thousands of dollars [PV=present-value]
    - present values discounted to year in which cohort is 65 years old
    - present values expressed in SSASIM:COHORT.cpi_year dollars
    - present values calculated using Treasury bond yields as discount rates
( 5) pvearn: present value of ind lifetime earnings (thousands of dollars)
( 6) pvtaxr: present value of ind lifetime OASI DB taxes and DC contributions
( 7) pvbenr: present value of ind lifetime OASI DB benefits and DC withdrawals
( 8) pvtaxd: present value of ind lifetime DI taxes
( 9) pvbend: present value of ind lifetime DI DB benefits and DC withdrawals
(10) irrind: real IRR of ind lifetime OASDI benefits/withd & taxes/contr (%)
(11) irri_2: 0 ==> one IRR value, 1==> multiple IRRs [(10) from HP-12C rule]
(12) dieage: death occurs at end of year just before next birthday
(13) aprice: OASDI account annuity price (current dollars to buy $1/year)
(14)  aror%: lifetime gross rate of return on OASDI account assets (%)
(15)  imed%: indiv-market lifetime equity rate of return difference (%)
(16) do_age: disability onset age (999 ==> no disability onset)
(17) dr_age: disability recovery age (999 ==> no disability recovery)
(18) ownpen: earned own pension & saved it for retire (0=no;1=DB;2=DC;3=DB+DC)
(19) pvpenb: PV of ind lifetime employer-sponsored pretax pension benefits
(20) pbalda: pension rollover account balance at death (same $K as pv stats)
(21) lngmar: at least one marriage of ten or more years (0 ==> no; 1 ==> yes)
(22) survbs: survived by spouse (0 ==> no; 1 ==> yes)
(23) n_kids: number of own children ("own" means biological, not step, kids)
(24) im_age: immigration (at start of year) age (0 ==> native-born indiv)
(25) em_age: emigration (at start of year) age (999 ==> no emigration)
(26) earn50: pretax earnings at age fifty (nominal thousands of dollars)
(27) fpvern: pv of family(equiv-scale-adj) lifetime earnings ($K)
(28) fpvtri: pv of family(equiv-scale-adj) lifetime OASI taxes/contribs ($K)
(29) fpvbri: pv of family(equiv-scale-adj) lifetime OASI benefits/withds ($K)
(30) fpvtdi: pv of family(equivalence-scale-adjusted) lifetime DI taxes ($K)
(31) fpvbdi: pv of family(equivalence-scale-adjusted) lifetime DI benefits ($K)
(32) fpvpen: pv of family(equiv-scale-adj) lifetime e-s pension benefits ($K)
(33) irrfrc: real IRR of family lifetime OASDI benefits/withd & taxes/contr (%)
(34) irrf_2: 0 ==> one IRR value, 1==> multiple IRRs [(33) from HP-12C rule]
```

```
(35) ib_amt: real initial OASDI benefit amount ($K in COHORT.cpi_year dollars)
(36) ib_age: age of receipt of initial OASDI benefit (999 if ib_amt is zero)
(37) pvb1nr: DB (tier-1) portion of pvbenr; DC portion: pvb2nr=pvbenr-pvb1nr
(38) pvb1nd: DB (tier-1) portion of pvbend; DC portion: pvb2nd=pvbend-pvb1nd
(39) lobnar: low-benefit+earnings avoidance rate [like SSASIM .c?f lobnar] (%)
NOTE: benefits in lobnar include family OASDI benefits AND pension benefits;
      family wage and salary earnings are added to these retirement benefits
(40) siearn: steady own earnings at age 65 for individual
(41) fsiern: steady equiv-scale-adjusted earnings at age 65 for family
NOTE: the "steady indexed earnings" statistics [(40) and (41)] are the
      constant AWI-indexed earnings from age 21 thru 65 that produce
      the same present value of earnings as the individual's actual
      present value of AWI-indexed earnings, where the present values
      are computed using the fixed real discount rate specified by the
      GEMINI:STATS.siearn_dr input parameter.  The statistics are
      "steady indexed earnings" at age 65 and are expressed in
      thousands of real (SSASIM:COHORT.cpi_year) dollars.  These
      statistics are suitable as the denominator in a replacement rate
      if the benefits in the numerator are also expressed in thousands
      of real dollars where the real dollars are for the
      SSASIM:COHORT.cpi_year.
MORE: additional information about replacement rate calculations is in the
      EDA Toolkit Operations section of the "Getting Started with the EDA
      Toolkit" document.
(42) icpvep: ind-to-cpl pvearn percentage ratio (%) (-1 ==> cpl pvearn is zero)
NOTE: the icpvep numerator is pvearn (5) and the denominator is similar to
      fpvern (27) except no family-size (or equiv-scale) adjustment is made.
(43) docage: age at which gains USA immigration documentation
NOTE: the value of docage is 0 for native-born individuals.
(44) numQoC: quarters of coverage at last insured-status/PIA calculation
(45) undocpaytax: whether undocumented immigrant pays OASDI payroll taxes
NOTE: 0==>no, 1==>yes (meaningless at ages when not an undocumented immigrant).
(46) siearx: same as statistic (40) except uses steady CPI-indexed earnings.
(47) fsierx: same as statistic (41) except uses steady CPI-indexed earnings.
NOTE: the "steady indexed earnings" statistics [(46) and (47)] are the
      constant CPI-indexed earnings from age 21 thru 65 that produce
      the same present value of earnings as the individual's actual
      present value of CPI-indexed earnings, where the present values
      are computed using the fixed real discount rate specified by the
      GEMINI:STATS.siearn_dr input parameter.  The statistics are
      "steady indexed earnings" at age 65 and are expressed in
      thousands of real (SSASIM:COHORT.cpi_year) dollars.  These
      statistics are suitable as the denominator in a replacement rate
      if the benefits in the numerator are also expressed in thousands
      of real dollars where the real dollars are for the
      SSASIM:COHORT.cpi_year.
MORE: additional information about replacement rate calculations is in the
      EDA Toolkit Operations section of the "Getting Started with the EDA
      Toolkit" document.
(48) eh_yrs: number of years with earnings above GEMINI STATS.eh_awi_pct
NOTE: read the STATS.eh_awi_pct input documentation for details.
(49) afsern: average final substantial real earnings (CBO-style rr denominator)
NOTE: read STATS.rr_aie_pct & STATS.rr_n_years input documentation for details;
      the value of afsern is zero in either of two cases:
      (a) AIME_at_age_62 was not computed (OAI-ineligible or other benefit),
      (b) number of substantial earnings years is less than STATS.rr_n_years;
      positive afsern values are expressed in thousands of age-62 dollars.
```

```
(50) aime62: AIME*12 at age 62 expressed in thousands of age-62 dollars
NOTE: aime62 will be zero if ind has no eligibility event at age 62 or is not
      eligible for retired-worker benefits at age 62.
```

Documentation of .xsa Output File

```
DOCUMENTATION FOR TEXT OUTPUT RESULTS FILE .XSA

--- .XSA file contains cross-sectional survey information for cohort sample
         members who are alive and are native-born or have immigrated by
         survey time, where the survey information is gathered in
         STATS.xsa_year, using supplementary questions in STATS.xsa_suppl,
         only if STATS.xsa_olc is true and if operating in OLC mode.
    IMPORTANT NOTE: If STATS.xsa_olc is true, be sure to combine the
         many xsaNNNNN.YYY files into a single runNNNNN.xsa file using
         the "GEMINI Data" capabilities of the EDA Toolkit (or use the
         xsall utility to produce the xsaNNNNN.all file with exactly
         the same contents as the runNNNNN.xsa file).

The GEMINI runNNNNN.xsa file has no heading lines and no summary
lines.  Individual survey results are written to a tab-delimited
file that has one row per person interview, making the survey
results file suitable for tabulating with an AWK program.
NOTE: the runNNNNN.xsa file must be constructed from the
component cohort files named xsaNNNNN.YYY using the "GEMINI Data"
capabilities in the EDA Toolkit for GEMINI or the xsall utility.

>>> IMPORTANT NOTE: the runNNNNN.xsa file never contains
                    individuals born before 1935.

FORMAT OF INDIVIDUAL SINGLE-ROW RECORD:
------------------------------------------------------------------------------
**** CORE SURVEY QUESTIONS
------------------------------------------------------------------------------
**** Core statistics are statistics for the xsa_year, many of which are the
**** the same as statistics in the .sum output file or the relevant age line
**** in the .adq output file.  The square bracket information indicates
**** whether the statistic is the same as a statistic in one of those files.
( 1) scenario: (positive integer) [sum:1]
( 2) indiv id: (positive integer) [sum:2]
( 3) indiv gender: 1=female, 0=male [sum:3]
( 4) indiv highest lifetime schooling: [sum:4]
            0 = never in high school,  1 = high school attended,
            2 = high school completed, 3 = college attended,
            4 = four-year college degree or graduate degree
( 5) indiv annual total earnings (thousands of nominal dollars) [adq:5]
( 6) im_age: indiv immigration age (0 ==> native-born indiv) [sum:24]
( 7) em_age: indiv emigration age [sum:25]
( 8) docage: indiv age at which gains USA immigration documentation [sum:43]
( 9) indiv xsa age: indiv age when xsa cross-section survey conducted
(10) indiv marital status [adq:1]
            0 = never married (marriage defined as lasting at least 0.5 year)
            1 = divorced (with no marriage lasting ten+ years so far)
            2 = divorced (with one+ marriage lasting ten+ years so far)
            3 = married
            4 = widowed (from most recent marriage)
```

```
(11) indiv family equivalence scale (adult equivalents in family) [adq:9]
            NOTE: determined by family composition and specified values of
            the STATS.equivs_p, STATS.max_k_age, and STATS.equivs_f parameters.
(12) couple annual total earnings (thousands of nominal dollars) [adq:6]
(13) indiv annual pretax OASDI benefit (thousands of nominal dollars) [adq:2]
NOTE that statistics 14-16 classify people using the SSA primary-benefit rule.
(14) number of OAI beneficiaries receiving the indiv OASDI benefit in (13)
(15) number of  SI beneficiaries receiving the indiv OASDI benefit in (13)
(16) number of  DI beneficiaries receiving the indiv OASDI benefit in (13)
NOTE that statistics 14-16 classify people using the SSA primary-benefit rule.
(17) couple annual pretax OASDI benefit (thousands of nominal dollars) [adq:4]
(18) indiv annual pretax employer-sponsored pension benefit (ditto) [adq:7]
(19) couple annual pretax employer-sponsored pension benefit (ditto) [adq:8]
(20) indiv annual pretax OASDI-covered earnings (thousands of nominal dollars)
(21) couple annual pretax OASDI-covered earnings (thousands of nominal dollars)
(22) indiv annual OASDI-taxable earnings (thousands of nominal dollars)
(23) couple annual OASDI-taxable earnings (thousands of nominal dollars)
(24) indiv adq-file benefit type (value 0 through 9) [adq:3]
------------------------------------------------------------------------------
END OF CORE SURVEY QUESTIONS
------------------------------------------------------------------------------
++++ SUPPLEMENT 1 QUESTIONS for OAI recipients:
---------------------------     ^^^^^^^^^^^^^^
++++ the OAI-recipient supplement:
++++ questions posed only to indiv with (24) equal to 1, 2, or 3.
(25) number of people getting OAI benefits (regardless of primary-benefit type)
(26) number of kids getting OAI benefits
(27) kids OAI benefits included in core statistic 13 (same units as 13)
------------------------------------------------------------------------------
++++ SUPPLEMENT 2 QUESTIONS for SI recipients:
---------------------------     ^^^^^^^^^^^^^
++++ the SI-recipient supplement:
++++ questions posed only to indiv with (24) equal to 4, 5, 6, or 9.
(25) number of people getting SI benefits (regardless of primary-benefit type)
(26) number of kids getting SI benefits
(27) kids SI benefits included in core statistic 13 (same units as 13)
------------------------------------------------------------------------------
++++ SUPPLEMENT 3 QUESTIONS for DI recipients:
---------------------------     ^^^^^^^^^^^^^
++++ the DI-recipient supplement:
++++ questions posed only to indiv with (24) equal to 7 or 8.
(25) do_age: disability onset age (999 ==> no onset) [sum:16]
(26) dr_age: disability recovery age (999 ==> no recovery) [sum:17]
(27) first age at which received DI benefits (should never be negative)
(28) kids DI benefits included in core statistic 13 (same units as 13)
------------------------------------------------------------------------------
++++ SUPPLEMENT 4 QUESTIONS for ALL individuals:
---------------------------     ^^^^^^^^^^^^^^^
++++ the federal income tax supplement (see PENSIM Overview, Chapter 7):
++++ questions posed to all individuals included in output file.
(25) income tax filing status:
       -1 = not a tax unit because is a dependent on another tax return,
        0 = single (apply unit weight in income tax tabulations),
        1 = married filing jointly (apply half weight in tabulations),
        2 = head of household (apply unit weight in tabulations).
          The parenthetical statements above are instructions about
          how to tabulate aggregate tax-return statistics from a GEMINI
          cross-sectional sample of individuals; failing to apply a
          one-half (0.5) weight to individuals who file jointly will result
          in the double counting of joint tax units.
(26) tax unit is nonfiler (0=no, 1=yes).
     NOTE: can be nonfiler because of -1 filing status or low income.
     IMPORTANT NOTE: ignore any non-zero tax statistics for a nonfiler.
(27) elderly tax unit with at least one person age 65+ (0=no, 1=yes).
(28) nominal annual AGI of individual's tax filing unit ($K).
(29) nominal annual OASDI benefit included in AGI ($K).
(30) nominal annual federal income tax liability ($K).
(31) nominal annual EITC included in tax liability ($K).
(32) number of EITC-qualified kids in individual's tax filing unit.
(33) nominal annual savers credit paid to individual's tax filing unit ($K).
(34) filing unit has positive OASDI income this year (0=no, 1=yes).
(35) filing unit has positive DI income this year (0=no, 1=yes).
(36) nominal annual imputed divinc this year ($K); details in PENSIM Overview.
(37) nominal annual imputed capinc this year ($K); details in PENSIM Overview.
(38) nominal annual imputed munint this year ($K); details in PENSIM Overview.
(39) nominal annual imputed businc this year ($K); details in PENSIM Overview.
(40) nominal annual imputed uicomp this year ($K); details in PENSIM Overview.
(41) nominal annual OASDI payroll tax credit included in tax liability ($K).
(42) nominal annual family employee DC pension contribution ($K).
(43) nominal annual family employer DC pension contribution ($K).
------------------------------------------------------------------------------
```

Documentation of .xsb Output File

```
DOCUMENTATION FOR TEXT OUTPUT RESULTS FILE .XSB

--- .XSB file contains cross-sectional survey information for cohort sample
         members who are alive and are native-born or have immigrated by
         survey time, where the survey information is gathered in
         STATS.xsb_year, using supplementary questions in STATS.xsb_suppl,
         only if STATS.xsb_olc is true and if operating in OLC mode.
    IMPORTANT NOTE: If STATS.xsb_olc is true, be sure to combine the
         many xsbNNNNN.YYY files into a single runNNNNN.xsb file using
         the "GEMINI Data" capabilities of the EDA Toolkit (or use the
         xsall utility to produce the xsbNNNNN.all file with exactly
         the same contents as the runNNNNN.xsb file).

The GEMINI runNNNNN.xsb file has no heading lines and no summary
lines.  Individual survey results are written to a tab-delimited
file that has one row per person interview, making the survey
results file suitable for tabulating with an AWK program.
NOTE: the runNNNNN.xsb file must be constructed from the
component cohort files named xsbNNNNN.YYY using the "GEMINI Data"
capabilities in the EDA Toolkit for GEMINI or the xsall utility.

>>> IMPORTANT NOTE: the runNNNNN.xsb file never contains
                   individuals born before 1935.

FORMAT OF INDIVIDUAL SINGLE-ROW RECORD:
------------------------------------------------------------------------------
**** CORE SURVEY QUESTIONS
------------------------------------------------------------------------------
**** Core statistics are statistics for the xsb_year, many of which are the
**** the same as statistics in the .sum output file or the relevant age line
```

```
**** in the .adq output file.  The square bracket information indicates
**** whether the statistic is the same as a statistic in one of those files.
( 1) scenario: (positive integer) [sum:1]
( 2) indiv id: (positive integer) [sum:2]
( 3) indiv gender: 1=female, 0=male [sum:3]
( 4) indiv highest lifetime schooling: [sum:4]
            0 = never in high school,  1 = high school attended,
            2 = high school completed, 3 = college attended,
            4 = four-year college degree or graduate degree
( 5) indiv annual total earnings (thousands of nominal dollars) [adq:5]
( 6) im_age: indiv immigration age (0 ==> native-born indiv) [sum:24]
( 7) em_age: indiv emigration age [sum:25]
( 8) docage: indiv age at which gains USA immigration documentation [sum:43]
( 9) indiv xsb age: indiv age when xsb cross-section survey conducted
(10) indiv marital status [adq:1]
            0 = never married (marriage defined as lasting at least 0.5 year)
            1 = divorced (with no marriage lasting ten+ years so far)
            2 = divorced (with one+ marriage lasting ten+ years so far)
            3 = married
            4 = widowed (from most recent marriage)
(11) indiv family equivalence scale (adult equivalents in family) [adq:9]
            NOTE: determined by family composition and specified values of
            the STATS.equivs_p, STATS.max_k_age, and STATS.equivs_f parameters.
(12) couple annual total earnings (thousands of nominal dollars) [adq:6]
(13) indiv annual pretax OASDI benefit (thousands of nominal dollars) [adq:2]
NOTE that statistics 14-16 classify people using the SSA primary-benefit rule.
(14) number of OAI beneficiaries receiving the indiv OASDI benefit in (13)
(15) number of  SI beneficiaries receiving the indiv OASDI benefit in (13)
(16) number of  DI beneficiaries receiving the indiv OASDI benefit in (13)
NOTE that statistics 14-16 classify people using the SSA primary-benefit rule.
(17) couple annual pretax OASDI benefit (thousands of nominal dollars) [adq:4]
(18) indiv annual pretax employer-sponsored pension benefit (ditto) [adq:7]
(19) couple annual pretax employer-sponsored pension benefit (ditto) [adq:8]
(20) indiv annual pretax OASDI-covered earnings (thousands of nominal dollars)
(21) couple annual pretax OASDI-covered earnings (thousands of nominal dollars)
(22) indiv annual OASDI-taxable earnings (thousands of nominal dollars)
(23) couple annual OASDI-taxable earnings (thousands of nominal dollars)
(24) indiv adq-file benefit type (value 0 through 9) [adq:3]
------------------------------------------------------------------------------
END OF CORE SURVEY QUESTIONS
------------------------------------------------------------------------------
++++ SUPPLEMENT 1 QUESTIONS for OAI recipients:
---------------------------     ^^^^^^^^^^^^^^
++++ the OAI-recipient supplement:
++++ questions posed only to indiv with (24) equal to 1, 2, or 3.
(25) number of people getting OAI benefits (regardless of primary-benefit type)
(26) number of kids getting OAI benefits
(27) kids OAI benefits included in core statistic 13 (same units as 13)
------------------------------------------------------------------------------
++++ SUPPLEMENT 2 QUESTIONS for SI recipients:
---------------------------     ^^^^^^^^^^^^^
++++ the SI-recipient supplement:
++++ questions posed only to indiv with (24) equal to 4, 5, 6, or 9.
(25) number of people getting SI benefits (regardless of primary-benefit type)
(26) number of kids getting SI benefits
(27) kids SI benefits included in core statistic 13 (same units as 13)
------------------------------------------------------------------------------
```

```
++++ SUPPLEMENT 3 QUESTIONS for DI recipients:
---------------------------     ^^^^^^^^^^^^^
++++ the DI-recipient supplement:
++++ questions posed only to indiv with (24) equal to 7 or 8.
(25) do_age: disability onset age (999 ==> no onset) [sum:16]
(26) dr_age: disability recovery age (999 ==> no recovery) [sum:17]
(27) first age at which received DI benefits (should never be negative)
(28) kids DI benefits included in core statistic 13 (same units as 13)
-------------------------------------------------------------------------------
++++ SUPPLEMENT 4 QUESTIONS for ALL individuals:
---------------------------     ^^^^^^^^^^^^^^^
++++ the federal income tax supplement (see PENSIM Overview, Chapter 7):
++++ questions posed to all individuals included in output file.
(25) income tax filing status:
       -1 = not a tax unit because is a dependent on another tax return,
        0 = single (apply unit weight in income tax tabulations),
        1 = married filing jointly (apply half weight in tabulations),
        2 = head of household (apply unit weight in tabulations).
          The parenthetical statements above are instructions about
          how to tabulate aggregate tax-return statistics from a GEMINI
          cross-sectional sample of individuals; failing to apply a
          one-half (0.5) weight to individuals who file jointly will result
          in the double counting of joint tax units.
(26) tax unit is nonfiler (0=no, 1=yes).
     NOTE: can be nonfiler because of -1 filing status or low income.
     IMPORTANT NOTE: ignore any non-zero tax statistics for a nonfiler.
(27) elderly tax unit with at least one person age 65+ (0=no, 1=yes).
(28) nominal annual AGI of individual's tax filing unit ($K).
(29) nominal annual OASDI benefit included in AGI ($K).
(30) nominal annual federal income tax liability ($K).
(31) nominal annual EITC included in tax liability ($K).
(32) number of EITC-qualified kids in individual's tax filing unit.
(33) nominal annual savers credit paid to individual's tax filing unit ($K).
(34) filing unit has positive OASDI income this year (0=no, 1=yes).
(35) filing unit has positive DI income this year (0=no, 1=yes).
(36) nominal annual imputed divinc this year ($K); details in PENSIM Overview.
(37) nominal annual imputed capinc this year ($K); details in PENSIM Overview.
(38) nominal annual imputed munint this year ($K); details in PENSIM Overview.
(39) nominal annual imputed businc this year ($K); details in PENSIM Overview.
(40) nominal annual imputed uicomp this year ($K); details in PENSIM Overview.
(41) nominal annual OASDI payroll tax credit included in tax liability ($K).
(42) nominal annual family employee DC pension contribution ($K).
(43) nominal annual family employer DC pension contribution ($K).
-------------------------------------------------------------------------------
```

Documentation of .xsc Output File

```
DOCUMENTATION FOR TEXT OUTPUT RESULTS FILE .XSC

--- .XSC file contains cross-sectional survey information for cohort sample
         members who are alive and are native-born or have immigrated by
         survey time, where the survey information is gathered in
         STATS.xsc_year, using supplementary questions in STATS.xsc_suppl,
         only if STATS.xsc_olc is true and if operating in OLC mode.
    IMPORTANT NOTE: If STATS.xsc_olc is true, be sure to combine the
         many xscNNNNN.YYY files into a single runNNNNN.xsc file using
         the "GEMINI Data" capabilities of the EDA Toolkit (or use the
         xsall utility to produce the xscNNNNN.all file with exactly
         the same contents as the runNNNNN.xsc file).

The GEMINI runNNNNN.xsc file has no heading lines and no summary
lines.  Individual survey results are written to a tab-delimited
file that has one row per person interview, making the survey
results file suitable for tabulating with an AWK program.
NOTE: the runNNNNN.xsc file must be constructed from the
component cohort files named xscNNNNN.YYY using the "GEMINI Data"
capabilities in the EDA Toolkit for GEMINI or the xsall utility.

>>> IMPORTANT NOTE: the runNNNNN.xsc file never contains
                    individuals born before 1935.

FORMAT OF INDIVIDUAL SINGLE-ROW RECORD:
------------------------------------------------------------------------------
**** CORE SURVEY QUESTIONS
------------------------------------------------------------------------------
**** Core statistics are statistics for the xsc_year, many of which are the
**** the same as statistics in the .sum output file or the relevant age line
**** in the .adq output file.  The square bracket information indicates
**** whether the statistic is the same as a statistic in one of those files.
( 1) scenario: (positive integer) [sum:1]
( 2) indiv id: (positive integer) [sum:2]
( 3) indiv gender: 1=female, 0=male [sum:3]
( 4) indiv highest lifetime schooling: [sum:4]
           0 = never in high school,  1 = high school attended,
           2 = high school completed, 3 = college attended,
           4 = four-year college degree or graduate degree
( 5) indiv annual total earnings (thousands of nominal dollars) [adq:5]
( 6) im_age: indiv immigration age (0 ==> native-born indiv) [sum:24]
( 7) em_age: indiv emigration age [sum:25]
( 8) docage: indiv age at which gains USA immigration documentation [sum:43]
( 9) indiv xsc age: indiv age when xsc cross-section survey conducted
(10) indiv marital status [adq:1]
           0 = never married (marriage defined as lasting at least 0.5 year)
           1 = divorced (with no marriage lasting ten+ years so far)
           2 = divorced (with one+ marriage lasting ten+ years so far)
           3 = married
           4 = widowed (from most recent marriage)
(11) indiv family equivalence scale (adult equivalents in family) [adq:9]
           NOTE: determined by family composition and specified values of
           the STATS.equivs_p, STATS.max_k_age, and STATS.equivs_f parameters.
(12) couple annual total earnings (thousands of nominal dollars) [adq:6]
(13) indiv annual pretax OASDI benefit (thousands of nominal dollars) [adq:2]
NOTE that statistics 14-16 classify people using the SSA primary-benefit rule.
(14) number of OAI beneficiaries receiving the indiv OASDI benefit in (13)
(15) number of  SI beneficiaries receiving the indiv OASDI benefit in (13)
(16) number of  DI beneficiaries receiving the indiv OASDI benefit in (13)
NOTE that statistics 14-16 classify people using the SSA primary-benefit rule.
(17) couple annual pretax OASDI benefit (thousands of nominal dollars) [adq:4]
(18) indiv annual pretax employer-sponsored pension benefit (ditto) [adq:7]
(19) couple annual pretax employer-sponsored pension benefit (ditto) [adq:8]
(20) indiv annual pretax OASDI-covered earnings (thousands of nominal dollars)
(21) couple annual pretax OASDI-covered earnings (thousands of nominal dollars)
(22) indiv annual OASDI-taxable earnings (thousands of nominal dollars)
(23) couple annual OASDI-taxable earnings (thousands of nominal dollars)
```

```
(24) indiv adq-file benefit type (value 0 through 9) [adq:3]
------------------------------------------------------------------------------
END OF CORE SURVEY QUESTIONS
------------------------------------------------------------------------------
++++ SUPPLEMENT 1 QUESTIONS for OAI recipients:
---------------------------     ^^^^^^^^^^^^^^
++++ the OAI-recipient supplement:
++++ questions posed only to indiv with (24) equal to 1, 2, or 3.
(25) number of people getting OAI benefits (regardless of primary-benefit type)
(26) number of kids getting OAI benefits
(27) kids OAI benefits included in core statistic 13 (same units as 13)
------------------------------------------------------------------------------
++++ SUPPLEMENT 2 QUESTIONS for SI recipients:
---------------------------     ^^^^^^^^^^^^^
++++ the SI-recipient supplement:
++++ questions posed only to indiv with (24) equal to 4, 5, 6, or 9.
(25) number of people getting SI benefits (regardless of primary-benefit type)
(26) number of kids getting SI benefits
(27) kids SI benefits included in core statistic 13 (same units as 13)
------------------------------------------------------------------------------
++++ SUPPLEMENT 3 QUESTIONS for DI recipients:
---------------------------     ^^^^^^^^^^^^^
++++ the DI-recipient supplement:
++++ questions posed only to indiv with (24) equal to 7 or 8.
(25) do_age: disability onset age (999 ==> no onset) [sum:16]
(26) dr_age: disability recovery age (999 ==> no recovery) [sum:17]
(27) first age at which received DI benefits (should never be negative)
(28) kids DI benefits included in core statistic 13 (same units as 13)
------------------------------------------------------------------------------
++++ SUPPLEMENT 4 QUESTIONS for ALL individuals:
---------------------------     ^^^^^^^^^^^^^^^
++++ the federal income tax supplement (see PENSIM Overview, Chapter 7):
++++ questions posed to all individuals included in output file.
(25) income tax filing status:
       -1 = not a tax unit because is a dependent on another tax return,
        0 = single (apply unit weight in income tax tabulations),
        1 = married filing jointly (apply half weight in tabulations),
        2 = head of household (apply unit weight in tabulations).
          The parenthetical statements above are instructions about
          how to tabulate aggregate tax-return statistics from a GEMINI
          cross-sectional sample of individuals; failing to apply a
          one-half (0.5) weight to individuals who file jointly will result
          in the double counting of joint tax units.
(26) tax unit is nonfiler (0=no, 1=yes).
     NOTE: can be nonfiler because of -1 filing status or low income.
     IMPORTANT NOTE: ignore any non-zero tax statistics for a nonfiler.
(27) elderly tax unit with at least one person age 65+ (0=no, 1=yes).
(28) nominal annual AGI of individual's tax filing unit ($K).
(29) nominal annual OASDI benefit included in AGI ($K).
(30) nominal annual federal income tax liability ($K).
(31) nominal annual EITC included in tax liability ($K).
(32) number of EITC-qualified kids in individual's tax filing unit.
(33) nominal annual savers credit paid to individual's tax filing unit ($K).
(34) filing unit has positive OASDI income this year (0=no, 1=yes).
(35) filing unit has positive DI income this year (0=no, 1=yes).
(36) nominal annual imputed divinc this year ($K); details in PENSIM Overview.
(37) nominal annual imputed capinc this year ($K); details in PENSIM Overview.
```

```
(38) nominal annual imputed munint this year ($K); details in PENSIM Overview.
(39) nominal annual imputed businc this year ($K); details in PENSIM Overview.
(40) nominal annual imputed uicomp this year ($K); details in PENSIM Overview.
(41) nominal annual OASDI payroll tax credit included in tax liability ($K).
(42) nominal annual family employee DC pension contribution ($K).
(43) nominal annual family employer DC pension contribution ($K).
------------------------------------------------------------------------------
```

2 Installing the Model

In earlier versions of the GEMINI Guide, this section explained how to install the GEMINI and PENSIM model. Installation instructions for the GEMINI, PENSIM, and SSASIM models can now be found in the "Installation of PSG Models" section of *Getting Started with the PSG Models* ⟨http://www.polsim.com/psghelp.pdf⟩.

3 Validating Model Output

This section of the guide describes the validation tests that have been performed on PENSIM life histories output and on GEMINI social security output.

3.1 PENSIM Validation

PENSIM output has been subjected to a broad range of validation tests that show it produces cohort samples that are realistic in both the demographic and economic aspects of people's life histories. For a complete discussion of these validation tests and PENSIM's performance on each test, see Chapter 7 of Holmer et al. (PENSIM Overview).

3.2 GEMINI Validation

Even though PENSIM life histories (including the annual earnings histories) appear to be realistic, it would be desirable to compare initial retired-worker social security benefits, simulated by GEMINI using a cohort input file generated by PENSIM, with benefits actually received by that birth cohort. The rest of this section describes such a validity test that was conducted for the 1935 birth cohort.

The basic idea of this validity test is to compare for men and for women average initial retired-worker benefits at age 62 for those born in 1935, as reported in the Social Security Administration's Annual Statistical Supplement, with simulated gender-specific benefits produced by GEMINI using a 1935 cohort input file generated by PENSIM.

The calibration target statistics chosen for the cohort born in 1935, the PENSIM parameters that were adjusted to match these calibration targets, and the closeness of the match between sample statistics and the calibration targets, are all described in Chapter 6 of Holmer et al. (PENSIM Overview).

Our validation test of GEMINI-simulated social security benefits focuses on comparing actual data on initial retired-worker awards made to those 62 years old in 1997 with simulated data on initial retired-worker awards in a GEMINI run using the 1935 birth cohort and the assumption that everyone (who is not already receiving social security benefits) quits working and applies for benefits at age 62.

Even this narrowly-focused validity test may have a selectivity problem caused by the fact that, in the actual data, not everyone eligible to apply for retired worker benefits does so at age 62. If the propensity to retire early at age 62 varies by lifetime earnings level, then the fact that only about sixty percent actually apply at 62 will complicate the comparison with statistics from the GEMINI simulation that assumes everyone applies at age 62.

We have investigated the extent to which average initial benefit levels in the actual data may be affected by this selectivity problem using primary insurance amount (PIA) data for the 1935 birth cohort, for which there is initial award data through age 65 at this writing. We analyze initial award data from the Social Security Bulletin's Annual Statistical Supplement (1998–2001) for initial awards in the 1935 cohort at ages 62–65, which accounts for about 93 percent of the total initial retired-worker awards made at age 62 and beyond according to Toder et al. (1999, Table 5-2, column F, p. 136). After inflation adjusting the average PIA of each group that retired after age 62 to be comparable to the average PIA of those who retired at age 62 in 1996, we find that the average PIA for men retiring at age 62 is 3.0 percent below the average PIA for men retiring at ages 62–65. The average PIA for women retiring at age 62 is 7.6 percent below the average PIA for women retiring at ages 62–65. Attributing all these differences to the selectivity problem would suggest that the average initial benefits at age 62 in the actual data should be adjusted upward by a factor equal to 1.031 for men and 1.083 (1/0.924) for women.

But we think it is likely that some of this difference is caused, not by the selectivity problem discussed above, but by the fact that some people delay their retirement in order to increase their PIA by replacing one to three positive-earnings years with zero-earnings years in their average indexed monthly earnings (AIME) calculation. This is quite plausible given the relatively low life-time employment rates for women born in the 1930s. Given this consideration, we think it is reasonable to think that the average initial benefits at age 62 in the actual data should be adjusted upward by a factor that is somewhat less than the factors calculated above.

The results of a validity test on average initial benefit levels are presented in Table 1, where we show both the unadjusted average from the actual data and the range of plausible adjusted averages given the above discussion of the selectivity problem and AIME effect.

The simulated gender composition of initial retired-worker awards at age 62 is similar to that indicated by the actual data for ages 62–65. The sim-

Table 1: ***Simulated and Actual Statistics on Average Initial Retired-Worker Awards in 1997 for People Aged 62***. *Simulated statistics from GEMINI (10/3/03 version) run 35100, which assumes everyone born in 1935 retires in 1997 at age 62, for retired-worker beneficiaries and retired-worker spouse beneficiaries who are dually entitled. Actual data on initial retired-worker awards for those born in 1935 retiring at age 62 in 1997 from Tables 6.B1 and 6.B2 in Social Security Bulletin's 1998 Annual Statistical Supplement (SSB-ASS). The ranges of the adjusted actual average are based on the discussion in the text. The gender composition data combines actual initial award data for ages 62–65.*

1935 Birth Cohort Statistic	GEMINI Value	SSB-ASS Value
Gender composition of awardees in percent:		
Men	61	57
Women	39	43
Average initial monthly benefit of awardees in dollars:		
Men	819	795–820
Women	517	509–551

ulated average monthly benefit level for initial retired-worker benefits for men is at the higher end of the range of actual data. The simulated average monthly benefit level for women is toward the bottom of the actual data range.

Comparing the distribution of initial retired-worker benefit awards is even more difficult, given available statistics. Table 6.B3 in Social Security Bulletin's 1998 Annual Statistical Supplement contains the distribution of 1977 initial benefits for all those whose benefit has been reduced for early retirement (that is, those whose initial award was in 1997 at age 62, age 63, or age 64). The presence in this table of awards that involved smaller actuarial reductions in benefits than applied to those retiring at age 62, not only increases the average benefit, but increases the absolute measures of benefit dispersion, in the actual data. In addition, the distribution of these data might be affected by extra positive-earnings years AIME effect discussed above.

Table 6.B3 data suggest that the inter-quartile range of monthly benefits — the difference between the 25th and 75th percentile values — is roughly

$350 for men and roughly $240 for women. The GEMINI-simulated interquartile range in initial benefits for only those whose initial benefit award was at age 62 in 1997 is $375 for men and $251 for women. The simulated dispersion in awards is relatively close to that suggested by data on actual benefits.

4 Analyzing Model Output

This section describes OLD methods of output analysis. These old methods have been made much less important by the introduction of the GEMINI Exploratory Data Analysis Toolkit. This EDA Toolkit provides easy data extraction and data visualization capabilities for quickly analyzing GEMINI output files. It is difficult to think of a situation in which the first step in analyzing GEMINI output would not involve using the EDA Toolkit. For a short introduction, start the EDA Toolkit from the Analyze menu of the RSF Toolkit, click on the Toolkit Help button, and then on the toolkit button, to read the *Getting Started with the EDA Toolkit* document.

The remainder of this section discusses pre-Toolkit methods of analyzing GEMINI output.

This final section of the guide discusses various methods for analyzing the GEMINI output results files, which are large in size and sometimes complex in structure (in order to conserve disk space). Many policy analysts would immediately think the best tool to use for this purpose would be a statistical analysis package like SAS or SPSS.

While these packages have many advantages, they have some major disadvantages: namely they are slow at some operations like sorting (this is particularly true of SAS). Given this situation, the Policy Simulation Group has developed alternative capabilities for GEMINI output analysis that may be used as a compliment to the statistical packages to work around their weaknesses or may be used as a substitute, thus eliminating the need to acquire additional statistical analysis software. This alternative capability is built around the AWK data processing language and the standard SSASIM output analyzer programs.

The remainder of this section discusses the GEMINI output utilities, the use of the AWK language for processing GEMINI output, the use of SSASIM output analyzer programs to analyze the processed output, the use of AWK programs to analyze GEMINI output, and the use of scripts (.bat files) to automate analysis of GEMINI output.

4.1 Using GEMINI Output Utility Programs

The GEMINI installation package contains a number of output utility programs. They are grouped below by function and described briefly with examples.

4.1.1 Showing Individual Output Records

Looking at an individual record in one of the GEMINI output results files is a very simple, but sometimes indispensable, way to analyze a GEMINI output results file. There is one utility program for showing an individual record for each kind of GEMINI text output file.

Documentation on how to use the "show" utilities can be obtained by selecting the appropriate Output analyzer program documentation item on the RSF Toolkit Analyze menu. When working at the Windows command prompt, simply enter the name of the show utility to get help with the command-line syntax and an example of how to use the utility.

The trace capability (see the TRACE table entries in the input database documentation) is a powerful feature of GEMINI, and its power is enhanced when using the show utilities. A typical situation is that when analyzing GEMINI output a particular individual is found to have puzzling results, perhaps because some results are statistical outliers. The show utilities can be used to quickly look at that individual's output records. And if this does not resolve the puzzle, then this run can be re-executed tracing this individual to see the details of that person's lifetime social security calculations.

Examples. To show the .sum output file record of individual 34592 in scenario 1 of a GEMINI run numbered 23456, enter the following at the Windows command prompt:

```
sumshow run23456 1 34592
```

To show the .adq output file record of individual 987 in scenario 9 of a GEMINI run numbered 77, enter the following:

```
adqshow run00077 9 987
```

4.1.2 Converting Longitudinal to Cross-Sectional Output

The .sum output file contains cross-section statistics (with one row for each sample individual), while the .adq output file contains longitudinal statistics on benefits and earnings (with many age rows for each sample individual. Often it is useful to merge information from the .sum file with information from the .adq file for a specific age. This can be done using the out2bage utility program.

Documentation on how to use this "out2bage" utility program can be obtained via the RSF Toolkit Analyze menu. When working at the Windows command prompt, simply enter `out2bage help` to get help with the command-line syntax.

Documentation on the file created by the out2bage utility program can be obtained via the RSF Toolkit Analyze menu. When working at the Windows command prompt, simply enter `out2bage odoc` to get the same documentation.

Example. To combine information from .sum output file and the .adq output file for age 65, which are generated by GEMINI run 34567, into a file suitable for cross-sectional statistical analysis, enter the following at the Windows command prompt:

```
out2bage 34567 65
```

The combined information is contained in the run34567.b65 file.

4.2 Using SSASIM Output Analyzer Programs

The text output analyzer programs that come with SSASIM provide a wide range of tabulation capabilities. These capabilities are described in the written documentation (Holmer SSASIM Guide) and on the RSF Toolkit Analyze menu under Output analyzer program documentation (the analyzers for text output files all start with “res”).

The GEMINI .sum and .bAA output files have a simple rectangular structure with one row per sampled individual. This means that all the SSASIM output analyzer programs can be used to analyze these files. The SSASIM output analyzer programs can be run interactively from the Windows command prompt or they can be called in scripts (.bat or .tcl files) that automate analysis of the output from many GEMINI runs.

4.3 Using AWK Programs to Tabulate Output

Sometimes GEMINI output needs to be processed in some way that is not possible using either the GEMINI or SSASIM output analyzer programs. If this is the case, then a custom statistical analysis program needs to be written. If available and familiar, the custom program can be written in the language understood by SAS or SPSS. In addition, there is the option of writing the custom program in the AWK language because an AWK language interpreter (Aho, Kernighan, and Weinberger 1988) is included in the SSASIM distribution package.

4.4 Using Scripts to Automate Processing

Often the GEMINI and SSASIM output analyzer programs, or the AWK programs, are invoked from the Windows command prompt. This approach is suitable for analyzing output from one or two GEMINI runs, but is not ideal when faced with the task of processing output from many runs. What is needed is a method of automating or scripting output processing and analysis tasks, and then applying that script repetitively to output from many runs.

The Windows operating system provides just such a scripting capability through the use of a *batch file* (or batch program). Basically, any command entered at the command prompt can be executed as part of a script by including that command on a line in a batch file (i.e., a file that ends in .bat). For more information, select the Help item on the Windows Start menu, click on the Index tab, enter “batch files” (without the quotes), and display the overview topic or any other topic of interest.

There is no limit to the number of output analysis tasks that can be automated using batch files.

References

Aho, Alfred V.; Kernighan, Brian W.; and Weinberger, Peter J. *The AWK Programming Language*. Reading, MA: Addison-Wesley, 1988.

Citro, Constance F. and Hanushek, Eric A., eds. *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling*, *Volume I: Review and Recommendations*. Washington, DC: National Academy Press, 1991.

Coplien, James O. *Advanced C++ Programming Styles and Idioms*. Reading, MA: Addison-Wesley, 1992.

Cotton, Paul and Sadowsky, George. "Future Computing Environments for Microsimulation Modeling," in Constance F. Citro and Eric A. Hanushek, eds., *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling*, *Volume II: Technical Papers*. Washington, DC: National Academy Press, 1991.

Date, C. J. *Database: A Primer*. Reading, MA: Addison-Wesley, 1983.

Hammersley, J.M. and Handscomb, D.C. *Monte Carlo Methods*. London, UK: Chapman and Hall, 1964.

Holmer, Martin R. *SSASIM Guide*. Washington, DC: Policy Simulation Group, various dates. The most recent version is available at ⟨http://www.polsim.com/doc/guide.pdf⟩

Holmer, Martin; Janney, Asa; and Cohen, Bob. *PENSIM Overview*. Washington, DC: Policy Simulation Group, various dates. The most recent version is available at ⟨http://www.polsim.com/doc/overview.pdf⟩

Toder, Eric; Uccello, Cori; O'Hare, John; Favreault, Melissa; Ratcliffe, Caroline; Smith, Karen; Burtless, Gary; and Bosworth, Barry. *Final Report: Modeling Income in the Near Term — Projections of Retirement Income Through 2020 for the 1931–60 Birth Cohorts*. Contract report to the Social Security Administration on MINT model development. Washington, DC: The Urban Institute, September, 1999.