JOURNAL OF SMALL BUSINESS STRATEGY

APPROACHING THE PSED: "SOME ASSEMBLY REQUIRED"

Kelly G. Shaver
College of Charleston
sharverk@cofc.edu

Amy E. Davis
College of Charleston
davisae@cofc.edu

Mark S. Kindy
Medical University of South Carolina
Ralph H. Johnson VA Medical Center
kindyms@musc.edu

Carrie A. Blair
College of Charleston
messalc@cofc.edu

ABSTRACT

The Panel Studies of Entrepreneurial Dynamics (PSED I and PSED II) are nationally representative longitudinal surveys of individuals in the United States who are in the process of starting businesses. These nascent entrepreneurs have been followed for three to four years (PSED I, N = 1,261, over 6,000 variables), or for six years (PSED II, N = 1,214, over 8,000 variables). As of this writing there are over 150 publications based on the PSED, but there could be even more if some of the critical data cleaning and data combining instructions were widely available. This article presents code (both SPSS and STATA) that can be used to check on the inclusion criteria, to renormalize weights for subgroup analysis, and to combine the data for PSED I with those for PSED II.

Keywords: PSED, longitudinal research, nascent entrepreneurship, syntax codes

Editor's Note (G. Hills): As noted in the conclusion, the PSED data set is the only representative national sample reflecting the firm creation process. Commenting as a member of the 'start up' PSED I team, there was an entrepreneurial spirit at the inception. There were 20 universities (ultimately growing to 34). Each pledged $20,000, which provided the initial funding. Paul Reynolds, Kelly Shaver, and many others deserve great credit for advancing scholarship and knowledge in the entrepreneurship field.

Preparation of this article was supported by the National Science Foundation Partnerships for Innovation (PFI) Program under Grant # IIP-0917987, Kelly G. Shaver, PI. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

INTRODUCTION

The two Panel Studies of Entrepreneurial Dynamics, PSED I and PSED II, provide entrepreneurship researchers with an extremely valuable resource for examining the process of creating a new business venture. Each of these datasets is a nationally representative sample of people who are in the process of creating new businesses (PSED I also includes data from a comparison group of individuals who were not starting businesses). Each dataset is both wide (over 6,000 variables for PSED I and over 8,000 variables for PSED II) and long, with a sizable array of questions asked of respondents who were then followed for four years (PSED I) or six years (PSED II). The longitudinal design allows researchers to identify the characteristics of start-up efforts that have (and have not) succeeded (Reynolds & Curtin, 2011).

As of this writing 13 books, over 50 book chapters, and more than 90 peer-reviewed publications based on the PSED studies have appeared in the literature (Frid, Gordon, & Davidsson, 2011). PSED I has been described in detail in a book edited by Gartner, Shaver, Carter, and Reynolds (2004) with chapters written by members of the research teams who contributed the variables included. A comprehensive description of the outcomes of PSED I was written by Reynolds (2007). PSED II has been described in a book edited by Reynolds and Curtin (2009) with chapters written by researchers interested in particular topics included in the data. A comprehensive description of the outcomes of PSED II has been written by Reynolds and Curtin (2008). Codebooks and interview schedules for both datasets are publicly available from the Institute of Social Research (ISR) at the University of Michigan, http://www.psed.isr.umich .edu/psed/home.

Details of the construction and use of each dataset (and of the combined or "harmonized" dataset created by Reynolds & Curtin, 2011) have been published (e.g., Appendices A-C in the Gartner, et al. book; the Reynolds & Curtin, 2011 paper). Together these resources are excellent references for researchers who have some prior experience with the data. But for many entrepreneurship researchers who have not yet dipped a toe into the PSED ocean, the technical details can appear overwhelming. There is a good reason that for every parent, the three most dreaded words at holiday time are "some assembly required." Where the PSED is concerned, there is a great deal of assembly required. In fact, there is enough so that with funding from the Ewing Marion Kauffman Foundation the first two authors of this paper have for four years taught a three-day course for doctoral students and faculty called "PSED 101." The present article presents some of the principles developed in that

course and includes parallel SPSS and STATA syntax required to accomplish the purposes described.

SCREENING TO IDENTIFY POTENTIAL RESPONDENTS

Collection of each dataset began with a "screener" that was embedded in a larger market research process. For PSED I, 64,622 individuals were reached by Market Facts (now Synovate) from July 1998 to January 2000 through random digit dialing and asked two screening questions:

1. Are you, alone or with others, currently trying to start a new business, including any form of self-employment?

2. Are you, alone or with others, now trying to start a new business for your employer? An effort that is part of your job assignment?

Respondents who answered affirmatively to either question (or to both) were then asked if they expected to be owners of the new firm and whether they had been active in the past 12 months in trying to establish the firm. Those who expected to be owners (in whole or in part) and who had been active were then asked if they could be contacted by a university-based survey research laboratory that was conducting research on the creation of new businesses in the United States. Early in the project the interviews were done by the University of Wisconsin Survey Research Laboratory. A third screening criterion – whether the business organizing effort is still in the start-up phase – was asked early in the University-based interviews. When the UWSRL closed, the PSED I interviews were completed by the University of Michigan's Survey Research Center (SRC).

PSED I consisted of two phases, an initial telephone interview (1,261 people) followed by a mail survey returned by 871 of the 1,261. Of the 1,261, 830 were nascent entrepreneurs, 431 were in a comparison group composed of people who had initially said they were *not* organizing a business. It is important for the data analysis to note that because of a gap in funding, roughly half of those screened for PSED I were reached a year later than the beginning of the first screening.

For PSED II, 31,845 individuals were reached by ORC International from October 2005 to January 2006 and asked three screening questions, rather than two, on the basis of what had been learned about inclusion criteria from the Global Entrepreneurship Monitor (GEM). The three PSED II screening questions were:

1. Are you, alone or with others, currently trying to start a new business, including any self-employment or selling goods and services to others?

2. Are you, alone or with others, currently trying to start a new business or new venture for your employer, an effort that is part of your normal work?

3. Are you, alone or with others, currently the owner of a business you help manage, including self-employment or selling any goods or services to others?

Respondents who answered one or more of these questions affirmatively were then asked additional questions to determine whether they had been active in the past 12 months, whether they personally would

own all, part, or none of the new business, and whether the business had received revenues sufficient to cover expenses including salaries or wages of the owners (this last point was actually three separate questions). Respondents who had engaged some start-up activity in the preceding 12 months, expected to own all or the major part of the new firm, and had not achieved revenues sufficient to be classified as an ongoing new firm were asked if they would consent to a telephone interview by the University of Michigan's Survey Research Center. Interviews were completed with 1,214 respondents (there was no mail survey, and no comparison group).

CHECKING THE INCLUSION CRITERIA

Researchers interested in firm-level issues typically include all 2,475 individuals, but researchers focusing on person-level variables often elect to remove the 53 people (45 in PSED I and 7 in PSED II) who did not meet a strict definition of the inclusion criteria. In PSED I, the SPSS code to accomplish these reductions was written by Paul Reynolds and modified by Kelly Shaver. The corresponding STATA code was written by Amy Davis. In the Tables that follow, the SPSS code precedes the corresponding STATA code (descriptions of what is being accomplished obviously apply in both cases).

In PSED I, 6 people had achieved positive cash flow for more than 90 days, so were by definition no longer in the organizing phase. An additional 7 individuals expected that some institution (technically, ownership by an entity that was not a person) would own more than 50% of the business. It was later discovered that 32 members of what was supposed to be the comparison group had actually been organizing a business at the time of the interview. The syntax to accomplish this data cleaning is contained in Table 1. *NOTE: The syntax to be used appears in Courier type.* When respondents who do not meet the strict definitions of the inclusion criteria are eliminated, there are 1,216 people left in PSED I (817 of whom are nascent entrepreneurs, 399 of whom are in the comparison group).

WEIGHTS AND RENORMALIZING

One of the major advantages of using PSED studies is that the data can be made to be nationally representative. In any survey research there is the possibility that biases will be introduced when contacting potential respondents. For the PSED studies the primary biases are differential selection probabilities and differential rates of non-response. For example, in PSED I five subsamples of data were collected. These were the initial sample (known as the "mixed gender sample," identified by the variable RTYPE with a score of 10), an oversample of women collected with funding from the National Science Foundation (RTYPE = 11), an oversample of minorities also collected with funding from the National Science Foundation (RTYPE = 12), a comparison group (RTYPE = 20) collected contemporaneously with the mixed gender sample, and a second comparison group (RTYPE = 21) collected contemporaneously with the minority oversample. The clearest conceptual indication of the need for weighting the data is provided by, for example, the oversample of women:

Table 1: Checking the Inclusion Criteria

STEP 1.

Eliminate 6 infant businesses that should have been screened out because they had positive cash flow including owner salary for more than 3 months prior to the date of interview. The cash flow variable is CFPHLAG (for Cash Flow PHone LAG). Eliminating these 6 infant businesses reduces the sample to 1255. (Individual respondents can be identified by sorting the data in descending order on the variable of interest. This puts the problematic respondents at the top of the data list.) RESPIDs for these 6 cases are 328100601, 37800137, 328100395, 328100124, 328100541, and 328100145.

SPSS CODE

```
FILTER OFF.
USE ALL.
SELECT IF (sysmis(cfphlag=1) or (cfphlag < 90)).
EXECUTE.

FREQ cfphlag.
```

STATA CODE

```
gen cfphlag1=cfphlag
recode cfphlag1 .=0
keep if cfphlag1<90
```

STEP 2.

Eliminate cases in which institutional ownership will exceed 50%. NPOWNPC was created by Paul Reynolds on the basis of Q217 (who will own?) answered as "not a person" and percentage of ownership (Q207C). This variable identifies 18 people (out of the total of 830 nascents) who expect that non-persons will own some percentage of the business. Of the 18, 7 show an expected non-person ownership greater than 50% (one at 66%, 1 at 82%, 1 at 85%, four at 100%). Delete these cases eliminates RESPIDs 328100020, 328100183, 328100255, 328100267, 328100443, 328100572, and 337800154, reducing the sample to 1248.

SPSS CODE

```
FILTER OFF.
USE ALL.
SELECT IF (sysmis(npownpc=1) or (npownpc LE 50)).
EXECUTE.

FREQ npownpc.
```

STATA CODE

```
drop if autonsu==5
```

Table 1 Continued

STEP 3.

Minority oversample Comparison Group participants, who are RTYPE 21, were asked the screening questions about start-up activities. Any who answered affirmatively to the question about start-up involvement, SUINVOL should be deleted from the Comparison Group. This represents a total of 28 people, 14 females and 14 males, leaving an overall total of 1220. (The total number of respondents can be seen by viewing the end of the listing in the "Data View" of the data file.)

SPSS CODE

```
DO IF (RTYPE = 21).
SELECT IF (SUINVOL = 1).
END IF.
EXECUTE.

FREQ SUINVOL.
```

STATA CODE

```
gen cgbiz=rtype
recode cgbiz 21=1 else=0
replace cgbiz=0 if suinvol==1
drop if cgbiz==1
```

STEP 4.

Respondents targeted for the ERC Mixed Gender and NSF Women (all of whom are RTYPE 20) comparison group were not asked about their start-up involvement. In the one-year follow-up, these respondents were asked about their start-up activities. (They should have had none.) The variable representing involvement is CGSUACT. This variable identifies four individuals who should be removed. RESPID numbers are 328200046, 328200059, 328200084, and 328200115, reducing the sample to 1,216.

SPSS CODE

```
DO IF (RTYPE = 20).
SELECT IF (CGSUACT NE 1).
END IF.
EXECUTE.

FREQ RTYPE.
```

STATA CODE

```
gen cgbiz1=rtype
recode cgbiz1 20=1 else=0
replace cgbiz1=0 if cgsuact==0
drop if cgbiz==1
```

Table 1 Continued:

STEP 5.

Finally, there is an error in one value assignment for AUTONSU. A frequency count of AUTONSU will show a total of 716 individuals, one of whose ventures is said to be 1-50% owned by a nonperson. Any such ownership, of course, means that the person is not fully autonomous. The problem is identified by a crosstab between AUTONSU and AUTONSU4 (AUTONSU4 will at this point have only three categories, as the fourth – institutional ownership > 50% has been eliminated).

SPSS CODE

```
CROSSTABS
      /TABLES= AUTONSU BY AUTONSU4
      /FORMAT= AVALUE TABLES
      /CELLS= COUNT.
```

STATA CODE

```
ta autonsu autonsu4
```

STEP 6.

In this crosstab the column totals, which represent AUTONSU4, are correct (715, 102, 817). The RESPID in error is 337800099. The correct value, determined by comparison to the frequencies in Q190 is a score of 3. Correct the value for this person then check to ensure that the column and row totals agree.

SPSS CODE

```
IF (RESPID = 337800099) AUTONSU = 3.
EXECUTE.

CROSSTABS
      /TABLES= AUTONSU BY AUTONSU4
      /FORMAT= AVALUE TABLES
      /CELLS= COUNT.
```

STATA CODE

```
replace autonsu=3 if respid==337800099
ta autonsu autonsu4
```

If a male (potential) respondent answered the telephone during the collection of this oversample, that male's probability of inclusion was zero.

Whether the initial screening was done by Market Facts (PSED I) or ORC (PSED II) the screening organization conducted interviews in replicated waves of 1000 people per wave. As part of their services to clients, these organizations provided a separate weight for every sample of 1000. Once all of the screening had been accomplished, the staff at SRC reconfigured the weights based on the total sample, a change that substantially reduced the variance in the weights (e.g., in PSED I from a range of nearly 10 points to a range of 1.7 points) according to Curtin (2004). A similar procedure was followed for PSED II, with comparable results. In each case the weight created is for the entire sample screened (64,622 or 31,845).

The general procedure for creating weights is to compare the percentage of respondents in a particular demographic group (e.g., white women aged 18-29 with incomes of $40,000 to $60,000) to the proportion of that same group in the total population of the United States, according to data from the Current Population Survey (CPS) of the U.S. Department of the Census. People in the specified demographic group would then be weighted to bring the weighted proportion into line with the proportion shown by the CPS. For example, if white women aged 18-29 with incomes of $40,000 to $60,000 were 15% of the CPS population but only 7.5% of the PSED sample, each respondent would be given a weight of 2. The demographic characteristics actually used to cross-classify the cells to be compared to the CPS were different from PSED I to PSED II because there were too many missing values in the income data. For PSED I the cells were the cross-classification of Gender X Ethnic Background X Age X Educational Attainment. For PSED II the cells were the cross-classification of Gender X Ethnic Background X Age X Income. Across all individuals in the screener, the weights were then centered to equal the total number of individuals screened. In both PSED I and PSED II the resulting weight is the variable WT_SCRN.

Although the screener samples are quite useful for estimating such things as the proportion of business creation activity among individuals with different gender, ethnic characteristics, and educational attainment, both screeners were limited to a very few questions. For detailed consideration of the factors involved in start-up, one needs the interview datasets. This means, of course, that the sums of weights need to be 1,261 and 1,214, not a number in the thousands. The process of creating weights for the detailed datasets began with using the weighted screener results to generate the demographic characteristics of nascent entrepreneurs. Following the logic outlined above for producing the screener weights, post-stratification weights were then created for the two detailed datasets. In PSED I the screener weights were adjusted by the SRC to produce one normalized weight for members of the comparison group (WTCG) and one for the nascent entrepreneurs in Wave 1 (WTW1). In PSED II, where there was no comparison group, the initial weight for the detailed dataset was WT_WAVEA. In each dataset there are weights for subsequent waves, but for present purposes we will restrict the discussion to the Wave 1 weights in both datasets.

If one is interested only in all nascents, or all nascents compared to members of the

comparison group, the weights given (WTW1, WTCG; WT_WAVEA) are sufficient. On the other hand, many investigators are interested in gender differences, differences between nascents who are fully autonomous (no financial support from any "nonperson"), or variables that appear only in the mail survey. In any or all of these cases, the overall weights will need to be renormalized so that the sum of the weights equals the number of individuals in the particular subsample of interest. An example of the problem is shown in Table 2.

Table 2: Example of Need for Recentering of Weights.

Gender (NCGENDER)	Number of People	Sum of WTW1	Mean of WTW1
Females	403	305.25	.76
Males	427	524.75	1.23
Total	830	830	1.00

Table 2 is based on the original dataset for PSED I (ERCW14Q, N = 1,261) downloaded from the ISR website. As noted above, in that original dataset there are 830 nascent entrepreneurs and 431 members of the comparison group. WTW1 was computed so that the total of this weight would equal the number of nascent entrepreneurs (830), and the bottom row in Table 2 shows that this is the case. The problem arises in the other two rows. When the sample of entrepreneurs is split into females and males (using NCGENDER, the only gender variable recommended for general use) the sum of weights for females is too small, whereas the sum of weights for the males is too large. This imbalance can be corrected by multiplying the value for WTW1 by a fraction consisting of (the number of individuals)/(the total weight for that class of individuals). Specifically, the new weights are:

For females, WTW1 * (403/305.25), sum of which is 403;
For males, WTW1 * (427/524.75), sum of which is 427.

The renormalizing of weights becomes a bit more complicated when the sample is cut two or more times. For this reason, Table 3 contains the syntax necessary to renormalize weights when the data of interest have been split on two dimensions. This procedure can simply be generalized to as many different splits as are needed for the particular research question. One note of caution: the SPSS command "MEANS TABLES," is not available through the menu system in some older versions of SPSS. All versions, however, recognize the command when it is written out into a syntax file.

COMBINING PSED I AND PSED II

As valuable as PSED I and PSED II are separately, they allow researchers to answer even more questions if they are combined into a single dataset. For sophisticated users of SPSS or STATA, accomplishing this task is a relatively simple matter. On the other hand, for those of us who are accustomed to working with one dataset at a time, putting the two together can be a challenge in at least four ways.

First, depending on the active memory of the computer you use, combining a 1261 person x 6000+ variable dataset with one that is 1214 person x 8000+ variables, it is prudent to be prepared for a crash. Minimize the number of other applications that are open, and save your work early and often. Some university email systems will not accept a file as large as the resulting combined dataset, so if you are working with colleagues it may be necessary to compress or zip the file.

Second, there is the need to have in the combined file some variable that indicates the source of the data (PSED I or PSED II). There are several ways to accomplish this, one of which is to add a variable called PSED (or SOURCE, or whatever variable name makes the most sense to you) to each dataset before the two are combined.

Table 3. Syntax for Renormalizing Case Weights, Example for PSED I Mail Questionnaire

STEP 1.

When the overall sample is reduced by eliminating people who did not return the mail questionnaire, the weights will need to be renormalized. The 871 respondents who completed the mail questionnaire will have a valid (not missing) value for return year (MAILQYR). Then retain only those respondents with a valid MAILQYR.

The counts should be as follows:
Full autonomy (245 females, 235 males, total of 480).
Partial autonomy (41 females, 32 males, total of 73).
Comparison group (173 females, 145 males, total of 318).

SPSS CODE

```
FREQ mailqyr.

FILTER OFF.
USE ALL.
SELECT IF(SYSMIS(mailqyr) NE  1).
EXECUTE.

CROSSTABS
  /TABLES=autonsu4 BY ncgender
  /FORMAT= AVALUE TABLES
  /CELLS= COUNT.
```

STATA CODE

```
drop if mailqyr ==.
ta autonsu4 ncgender
```

Table 3 Continued:

STEP 2.

Next, compute a weight that for nascent entrepreneur respondents will be WTW1 but for comparison group respondents will be WTCG.

SPSS CODE

```
COMPUTE weight = 0.
EXECUTE.
IF (rtype = 10) weight = wtw1.
IF (rtype = 11) weight = wtw1.
IF (rtype = 12) weight = wtw1.
IF (rtype = 20) weight = wtcg.
IF (rtype = 21) weight = wtcg.
EXECUTE.
```

STATA CODE

```
gen weight=0
replace weight=wtw1 if rtype==10
replace weight=wtw1 if rtype==11
replace weight=wtw1 if rtype==12
replace weight=wtcg if rtype==20
replace weight=wtcg if rtype==21
```

STEP 3.

Next, check the weights for a Gender x Autonomy split (which will have six cells). The result will show the numbers to use as divisors in the fractions to renormalize.

SPSS CODE

```
MEANS TABLES= weight BY autonsu4 BY ncgender
  /CELLS SUM MEAN COUNT STDDEV.
```

STATA CODE

```
sort autonsu4
by autonsu4: su weight if ncgender==1
by autonsu4: su weight if ncgender==2
```

Table 3 Continued:

STEP 4.

Finally, renormalize the weights for each of these six cells. At this point the sum of the weights for a cell should agree with the number of individual respondents in that cell.

SPSS CODE

```
COMPUTE RENORMWT = 99.
EXECUTE.
IF ((ncgender = 2) and (autonsu4 = 100)) RENORMWT =
weight*(245/184.87).
IF ((ncgender = 2) and (autonsu4 = 200)) RENORMWT =
weight*(41/33.95).
IF ((ncgender = 2) and (autonsu4 = 400)) RENORMWT =
weight*(173/175.06).
IF ((ncgender = 1) and (autonsu4 = 100)) RENORMWT =
weight*(235/291.68).
IF ((ncgender = 1) and (autonsu4 = 200)) RENORMWT =
weight*(32/40.12).
IF ((ncgender = 1) and (autonsu4 = 400)) RENORMWT =
weight*(145/160.14).
EXECUTE.

MEANS TABLES= renormwt BY autonsu4 BY ncgender
  /CELLS SUM MEAN COUNT STDDEV.
```

STATA CODE

```
gen renormwt=99
replace renormwt=weight*(245/184.87) if ncgender==2 &
autonsu4==100
replace renormwt=weight*(41/33.95) if ncgender==2 &
autonsu4==200
replace renormwt=weight*(173/175.06) if ncgender==2 &
autonsu4==400
replace renormwt=weight*(235/291.68) if ncgender==1 &
autonsu4==100
replace renormwt=weight*(32/40.12) if ncgender==1 &
autonsu4==200
replace renormwt=weight*(145/160.15) if ncgender ==1 &
autonsu4==400

sort autonsu4
by autonsu4: su renormwt if ncgender==1
by autonsu4: su renormwt if ncgender==2
```

This variable would be given a value of 1 if the source dataset was PSED I, and a value of 2 if the source dataset was PSED II. This method is accomplished by the SPSS syntax in Table 4. Another way is to combine the datasets and *then* create a variable representing the source dataset. This method is accomplished in the STATA syntax in Table 4. NOTE: to make the combining as widely useful as possible, we show how to combine the two original datasets (with no elimination of respondents from either dataset).

Table 4. Syntax for Combining PSED I and PSED II

STEP 1.

This syntax contains pathnames where datasets are to be found. *Those pathnames will differ from user to user depending on where both files are stored.* First, download ERCW14Q.sav from the ISR website (this will be data file 2b under the PSED I heading). Save the file onto your desktop. Next, download psedii_scrn_ABCDEF.sav from the ISR website. Also save this file to your desktop.

STEP 2.

Next, retrieve the ercw14q.sav dataset from your desktop and create a variable called PSED to identify the source dataset. Make all values of PSED = 1. Then save the file back to your desktop with a new name (the example uses "psed1.sav").

SPSS CODE

```
GET   FILE='/Users/kellyshaver/Desktop/ERCW14Q.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.

COMPUTE psed = 1.
EXECUTE.
VARIABLE label psed 'source dataset'.
VALUE labels psed
      1     'from psed1'
      2     'from psed2'.
EXECUTE.
FREQ psed.

SAVE OUTFILE='/Users/kellyshaver/Desktop/psed1.sav'
  /COMPRESSED.
```

STEP 3.

Next, retrieve the psedii_scrn_ABCDEF dataset from your desktop, create a variable called PSED and make all values of PSED = 2. Then save the file back to your desktop with a new name (the example uses "psed2.sav").

SPSS CODE

```
GET FILE='/Users/kellyshaver/Desktop/psedii_scrn_ABCDEF.sav'.
DATASET NAME DataSet2 WINDOW=FRONT.

COMPUTE psed = 2.
EXECUTE.
VARIABLE label psed 'source dataset'.
VALUE labels psed
      1     'from psed1'
      2     'from psed2'.
EXECUTE.
FREQ psed.

SAVE OUTFILE='/Users/kellyshaver/Desktop/psed2.sav'
  /COMPRESSED.
```

NOW CLOSE psed2.

Table 4 Continued:

STEP 4.

Finally, with psed1.sav open, add psed2 to it. When the two files are combined, there should be a total of 2,475 people, which can be confirmed by checking the frequency of PSED. Of course, if you have previously eliminated respondents who did not return the PSED I mail survey, the total number will be (871+1214) = 2,085.

You will probably want to save this combined file so that you do not have to do the combining every time you care to do an analysis.

SPSS CODE

```
ADD files
  FILE='/Users/kellyshaver/Desktop/psed2.sav'
  FILE=*.
EXECUTE.

FREQ psed.
```

STATA CODE FOR THE OTHER METHOD OF COMBINING

```
format respid %20.0f
gen sampid=respid
sort sampid
format sampid %20.0f

append using "C:\Documents and Settings\davisae\My
Documents\research\PSEDIIhandbook\psedii_scrn_ABCDEF.dta"
gen psed=sampid
recode psed 328100000/537800160=1 50001/60000=2
```

Third, once the datasets have been combined, it is essential to check all variables of interest using the data and the relevant codebook (codebooks for both datasets are also available as PDF files from the ISR website). Not all of the items included in PSED I are present in PSED II, and the latter contains variables not present in the former. Even when the variables are identical across datasets, their names will not be. Variables in PSED I have their waves identified by a leading capital letter (Q for wave 1, R for wave 2, S for wave 3, and T for wave 4). In PSED II, by contrast, waves 1-6 are identified by the leading capital letters A-F. In the mail questionnaire for PSED I, different conceptual variables appear together, based on the nature of their response scales (e.g.,most variables with 5-point scales were grouped together, whether or not they were conceptually related). In PSED II the variables are grouped in "modules," but the placement of items into modules would not be done the same way by each of a dozen researchers interested in the topics. So if a variable of interest to you does not appear in the module where you expect it, don't stop looking. It could simply be somewhere else.

Fourth, check both the codebook and the variable listing to make sure that a particular variable of interest (a) had the same stem and response scale from PSED I to PSED II, and (b) that the numbers assigned to response alternatives were

identical from one dataset to the other (this is not always true). For example, in PSED I the conceptual variable of entrepreneurial intensity was assessed with four items in the mail questionnaire. These are ql1d (q-ell-one-d) to ql1g:

d. I would rather have my own business than pursue another promising career.

e. There is no limit to how long I would give maximum effort to establish my business.

f. My personal philosophy is to "do whatever it takes" to establish my own business.

g. Owning my own business is more important than spending time with my family.

For each item there was a response scale with five alternatives: completely untrue (1), mostly untrue (2), it depends (3), mostly true (4), completely true (5) such that higher numbers represent greater levels of intensity. In PSED II, however, only two of the items were repeated (e and f) appearing as AY9 and AY10. Here the response scale has six alternatives: strongly agree (1), agree (2), neither (3), disagree (4), strongly disagree (5) and not relevant (6). Ignoring the last alternative, it is clear that higher numbers represent lesser levels of intensity. Thus, in the combined dataset, a researcher would have only two items available and would have to reverse score those two.

STARTUP TEAMS

A distinctive feature of the PSED is its inclusion of secondary founders in its surveys. Most large-scale surveys of entrepreneurs and business owners only seek information from the primary owner of each business. Thus, many entrepreneurs remain "hidden" from scholarly inquiry. By contrast, in the PSED, the use of household telephone numbers as the sampling frame means that the originator of the entrepreneurial concept is just as likely to be interviewed as the fourth team member that he or she recruited to the startup. Indeed, in PSED II, more than 200 respondents listed their primary role as being something other than "general management" or "everything" and more than 100 respondents reported that someone else on the team was in charge of daily operations in the business (Davis, Longest, Kim, & Aldrich 2009). Therefore, researchers must be mindful that although the PSED is richer for its inclusion of secondary entrepreneurs, those studying individual differences or personality should control for team characteristics because the attitudes and behaviors of an individual who initiated the startup process may be considerably different from an individual who was recruited into an ongoing nascent venture.

PSED I and II contain information about team members' demographic characteristics, human capital characteristics, contributions to the startup, and relationships among team members. All of this information is reported from the point of view and recollection of the respondent. In PSED I, respondents were asked about their occupation, industry experience, entrepreneurial experience, and amounts of money and time invested in the business in different places depending on whether they were starting their business by themselves or as members of teams. For example, if someone new to the PSED ran an analysis of q197 (the amount of

money a respondent has invested in the startup), he or she would find only 376 responses out of the 830 potential answers. Therefore, for anyone interested in studying teams or human capital and startup investments in the PSED I, the most important variables are q210b_1 through q210b_5. These variables indicate whether the person about whom other questions are asked is the respondent or not. Note that these variables do not capture human capital and startup investments across team members but simply restore missing values for respondents on teams.

CONCLUDING REMARKS

The two PSED datasets are important resources for researchers who seek to examine the early stages of new business formation. Indeed, Reynolds and Curtin (2008) identified 26 separate datasets related to business creation, including those from the Bureau of Labor Statistics, the Census Bureau, Dun & Bradstreet, the Internal Revenue Service, the Kauffman Foundation, the National Opinion Research Center, the National Science Foundation, the Small Business Administration, and the University of Michigan. In their words, "Only one extant research program, the Panel Study of Entrepreneurial Dynamics, provides detailed information on a representative national sample reflecting the firm creation process" (p. 162). The purpose of this article is to make the PSED data more approachable by newcomers. In short, we hope we have provided an abbreviated diagram to help reduce the frustration of using the PSED given that there is "some assembly required."

REFERENCES

Davis, A. E., & Shaver, K. G. (2009). Social motives in the PSED II. In P. D. Reynolds & R. T. Curtin (Eds.), *New firm creation in the United States: Initial explorations with the PSED II data set* (pp.19-34). Dordrecht, Germany: Springer.

Davis, A. E., Longest, K. C., Kim, P. H., & Aldrich, H. E. (2009). Owner contributions and equity. In P. D. Reynolds & R. T. Curtin (Eds.), *New firm creation in the United States: Initial explorations with the PSED II data set* (pp.71-94). Dordrecht, Germany: Springer.

Frid, C., Gordon, S., & Davidsson, P. (2011). *Publications based on the Panel Study of Entrepreneurial Dynamics.* Downloaded from http://www.psed.isr.umich.edu/psed/documentation, May 24, 2012.

Gartner, W. B., Shaver, K. G., Carter, N. M., & Reynolds, P. D. (Eds.) (2004). *The handbook of entrepreneurial dynamics: The process of business creation.* Thousand Oaks, CA: Sage Publications.

Reynolds, P. D. (2007). New firm creation in the U.S.: A PSED I overview. . *Foundations and Trends® in Entrepreneurship, 3* (1), 1-149.

Reynolds, P. D., & Curtin, R. T. (2008). Business creation in the United States: Panel Study of Entrepreneurial Dynamics II Initial Assessment. *Foundations and Trends® in*

Entrepreneurship, 4 (3), 155-307. DOI 10.1561/0300000022.

Reynolds, P. D., & Curtin, R. T. (2011). *PSED I, II Harmonized transitions, outcomes data set.* Downloaded from http://www.psed.isr.umich.edu/psed/documentation, May 24, 2012.

Reynolds, P. D., & Curtin, R. T. (Eds.), New firm creation in the United States: Initial explorations with the PSED II data set. Dordrecht, Germany: Springer.

Kelly G. Shaver is Professor of Entrepreneurial Studies in the School of Business at the College of Charleston. His prior appointments include the College of William & Mary, the National Science Foundation, and the Entrepreneurship and Small Business Research Institute (ESBRI) in Stockholm, Sweden. Shaver served as a member of the PSED I Executive Committee and the PSED II Advisory Committee. His highly cited research has been supported by the Ewing Marion Kauffman Foundation, the National Institute of Mental Health, and the National Science Foundation.

Amy E. Davis is Assistant Professor of Entrepreneurship at the College of Charleston. Her research interests include social networks, startup teams, biomedical entrepreneurship, and gender. She has published in *Entrepreneurship Theory and Practice*, *Frontiers of Entrepreneurship Research*, and *Work and Occupations*.

Mark S. Kindy is the Admiral Pihl Professor of Neurosciences in the College of Medicine at the Medical University of South Carolina and Career Research Scientist at the Ralph H. Johnson VA Medical Center. Dr. Kindy was on the faculty at the University of Kentucky School of Medicine, is a member of several societies and has served on numerous editorial boards and review committees. Dr. Kindy has been supported by the National Institutes of Health, National Science Foundation, Veterans Administration, Department of Defense, and the American Heart Association.

Carrie Blair Messal is an Assistant Professor of Management in the School of Business at the College of Charleston. Her passion is leader development. She has designed and executed components of the Executive Education Leader Development Program at the University of Tennessee-Knoxville. She is the Founder and Director of the Schottland Scholars Program, a leader development program for School of Business undergraduates at the College of Charleston.