Using SAS® to Facilitate Data Comparisons

M. Rita Thissen and Elizabeth Heath, RTI International, RTP, NC

ABSTRACT

As part of a longitudinal survey of children, we collected data about the primary subjects and their caregivers. Identifying information for many of the caregivers changed from one wave of interviewing to the next as the children moved from family settings to foster homes or other living situations. In retrospect, we needed to determine if the caregiver had changed or had remained the same. We used SAS to match caregivers by case ID and perform a logical comparison of selected identifier information, assigning each caregiver a value for a flag variable which denoted whether the caregiver was the same or different from the previous interview, or if the comparison was ambiguous.

In order to increase the usefulness of the flag variable, project staff manually reviewed data for each ambiguous case, to determine if the caregiver was the same, different, or if it was still ambiguous. Due to typographical errors in collected data, changes in caregiver marital status and family name, or information that was provided at one interview but not in the other, there were a number of interviews that appeared ambiguous to the programmed logic of variable comparisons, but which were quickly resolved as same or different caregivers by the human reviewers.

We used SAS code to pre-clean the data, assign rating values, select cases which remained ambiguous, export comparison and identity data to Excel® spreadsheets for review, and finally to incorporate the human ratings back into the original data set. The data management and file handling capabilities of SAS allowed the process to proceed in a reproducible and productive manner. We will review how we used SAS to handle data and prepare for manual review.

INTRODUCTION

The National Survey of Child and Adolescent Well-Being (NSCAW) is a long-term study of children who have been reported as being abused or neglected in the United States. By following these children's living situations, services received, health and accomplishments for several years, a wealth of information will be made available to analysts and policy makers, for use in determining desirable intervention options and social programs. The NSCAW study is sponsored by the Administration for Children and Families (ACF), in the U.S. Department of Health and Human Services, and has been conducted for the past several years by RTI International. Details of the study can be obtained at the ACF website (http://www.acf.hhs.gov/programs/core/ongoing_research/afc/wellbeing_intro.html). Refer to Figure 1 for definitions of survey terminology used in this paper.

Figure 1. Survey terminology

Term	Definition
Sample member	Person selected to participate in the survey.
Respondent	Person who completes an interview.
Longitudinal survey	Multiple administration of a survey to the sample members, over a time interval, in order to track how the data change over time
Wave	A single administration of the survey. The NSCAW study had four waves of interviews. Wave one is the first survey administration while wave four is the last.
Case set	Unique identifier used to associate the caregiver and other respondents with a specific child in the survey. There were approximately 6200 case sets in the NSCAW study.
Finalized non-complete interview	An interview was not completed by the sample member. A finalized non-complete wave 2 interview means that the sample member did not complete the survey at wave 2.
Finalized completed interview	An interview was completed by the sample member.
Kin-care	The caregiver is a relative, but not the biological parent, of the child.

Data were collected from approximately 6200 sets of children, caregivers, and other respondents. RTI attempted to complete four separate interviews with each caregiver. During this longitudinal survey, many children changed living situations and caregivers, as they moved in and out of parental care, kin-care, or foster care settings. Because respondent identifying information is not provided to NSCAW data set users, we needed to establish whether the caregiver had changed from one interview to another. In order to resolve whether the caregiver had changed for a pair of completed interviews, logical comparisons of caregiver identifier information were used to assign one of three values to a flag variable. The three values were: same caregiver, different caregiver, or it was not clear or was ambiguous. For discussion purposes, we will consider the comparisons made for completed wave 4 interviews and the most recently completed prior wave interview.

The logical comparisons assigned 820 case sets an ambiguous rating. Examination of a small set of this data revealed that the majority of the ambiguous cases actually involved the same caregiver. The cases had been assigned an ambiguous rating because of changes in caregiver last name (due to changes in caregiver marital status), identifying information that was provided by the caregiver at one interview but not another, and typographical errors in the collected data. Because it appeared that the number of cases with an ambiguous rating could be significantly reduced, two human reviewers were asked to independently examine identifying information for case sets that were assigned an ambiguous rating, and assign a rating of same caregiver, different caregiver, or ambiguous, in order to increase the usefulness of the flag variable.

Reviewer evaluations were checked for completeness and then reconciled. When the two reviewers agreed on a rating, that rating was used; when the two reviewer ratings differed, an ambiguous rating was assigned. The human reviewer ratings differed for only 19 case sets. After the human evaluation, 30 case sets had an ambiguous rating. Reconciled human ratings were merged with the original data set.

We will review examples of ambiguous data and how we used SAS to do: data pre-cleaning, automated logical comparisons and rating assignments, prepare Excel spreadsheets with caregiver identifying information for human review, import human reviewer ratings, and resolve the final ratings.

DATA PRE-CLEANING

The SAS compress function removes blanks and characters (specified by the user) from a variable. Compress function syntax is shown in Figure 2. In line (2a), the source variable (var1) contains blank spaces, underscores, a single quotation mark, a numeric character, and alphabetic characters. In line (2b), the characters to be removed are listed between double quotation marks because the list includes a single quotation mark. Note that the list also includes a blank space. As shown in line (2c), the compress function removes each occurrence of the specified characters and all blank spaces. When no characters are specified for removal in the compress function, as shown in line (2d), all blank spaces are removed, as shown in line (2e). In order to pre-clean the data for analysis, special characters were removed from values for caregiver names and relationship to child.

Figure 2. SAS compress function syntax

```
data _null;
var1 = " a b bbb c  d    3_ _'";                                    (2a)
x=compress(var1, "_ bc '");                                         (2b)
*     x=ad3;                                                        (2c)
y=compress(var1);                                                   (2d)
*     y=abbbbcd3__';                                                (2e)
run;
```

IDENTIFYING MOST RECENTLY COMPLETED PRIOR INTERVIEW

OVERVIEW

Four waves of interviews were conducted. Because about 27 percent of the caregivers interviewed at wave 4 did not complete the wave 3 interview, we had to identify the prior wave that was completed most recently. This meant that wave 2 or 1 might be the most recently completed prior interview. Data are sorted and processed by descending wave for each case set. If a wave 4 caregiver interview was not completed, the case set is deleted from the determination of the same caregiver flag. A wave 4 completion flag (w4comp) is initialized to a character missing value while a prior completed interview flag (gotprior) is initialized to N. Both flag values are retained. The general approach, to identifying the prior completed wave, can be followed with the example in Figure 3. In Figure 3, interviews were completed at waves 1, 2, and 4. For case set AAA1 in Figure 3, the wave 4 completion flag is set to Y for the completed wave 4 interview and the value is retained. The wave 3 interview is not completed and the wave 3 observation is deleted. The wave 2 interview is completed and the value of the prior completed interview flag is set

to Y and retained. The wave 1 interview is deleted because the prior completed interview flag was set for the wave 2 interview. If the wave 4 interview is the last interview in the case set, which means that wave 4 is the only interview completed by the caregiver, the wave 4 observation is deleted. This process is repeated for each case set that has a completed wave 4 interview.

Figure 3. Example identification of most recently completed prior interview

Case set	AAA1			
Wave	Interview completed at the wave	Wave 4 completed Interview flag	Completed prior wave interview flag	Delete observation
4	Yes	Y	N	No
3	No	Y	N	Yes
2	Yes	Y	Y	No
1	Yes	Y	Y (set for wave 2)	Yes

CODE REVIEW

Code for determining the most recently completed prior wave is shown in Figure 4. In line (4a), the data are processed by descending wave per the case set variable (casegrp). In lines (4b1) and (4b2), for the first observation in the case set, the wave 4 completion flag and prior completed interview flag are initialized and the values are retained.

The wave 4 completion flag is set for wave 4 cases in lines (4c1) through (4c3). If the wave 4 interview was completed, the wave 4 completion flag variable, w4comp, is set to a value of 'Y'. If the wave 4 interview was a finalized non-complete interview, the flag is set to a value of 'N' and the same caregiver flag is set to ambiguous (in the do loop in line (4c3)).

Lines (4d1) through (4d3) delete observations that contain: finalized non-complete wave 4 interview data in line (4d1), any earlier completed wave interview, when the most recently completed interview has already been identified with the prior completed interview flag, in line (4d2), and any finalized non-complete prior wave interview in line (4d3). There is no need to do data comparisons if the wave 4 interview was not completed or if wave 4 was the only completed interview. In line (4e) the prior completed interview flag is assigned a value of 'Y' for the most recently completed prior wave interview.

Figure 4. Determination of most recently completed interview

```
data sameCG;
   set sameCG;
   by casegrp descending wave ;                                          (4a)
   *flag w4comp - wave 4 was complete for the case set;
   *flag gotprior - a prior complete for the case set was found;
   length w4comp $1 gotprior $1;
   retain w4comp gotprior ;                                              (4b1)
   if first.casegrp then do;                                             (4b2)
      w4comp = '';
      gotprior = 'N';
   end;
   if wave = 4 then do;                                                  (4c1)
      if sumstat GE 490 then w4comp = 'Y';                               (4c2)
      else do;                                                           (4c3)
            w4comp = 'N';
            cg4same = -4;
      end; *else;
   end;  *if wave 4;
   else do;
      if w4comp = 'N' then delete;      *non-complete wave 4;            (4d1)
      else do;  *wave 4 complete;
        if gotprior = 'Y' then delete;  *have more recent prior wave;    (4d2)
        if sumstat LT 490 then delete;  *non-complete prior wave;        (4d3)
        else gotprior = 'Y';            *keep this one;                  (4e)
      end;
   end;            run;
```

LOGICAL COMPARISONS

OVERVIEW

During each interview, questions were repeated in a different context and in different sections, so the data could be checked for consistency. As a result, there were two source variables for the caregiver birth date and social security number.

In determining whether the caregiver is the same, logical comparisons are made for the caregiver first and last names, date of birth, and social security number, in the wave 4 and prior wave data. If this information matches, the caregiver is the same person.

Another method is used to identify where the caregiver has changed. Counts are totaled for differences in non-missing values for caregiver first name, last name, social security number, date of birth, relationship to the child, and gender. The totals are increased by one if the flag prSameR indicates that the caregiver has changed. If the total number of differences equals or exceeds 3, the caregiver is different. Case sets that have neither the same nor different caregiver are assigned an ambiguous rating.

CODE REVIEW

Data are sorted by case set and ascending wave, so that completed prior wave data are listed before completed wave 4 data for each case set. Part of the logical comparison code is shown in Figure 5. The data are processed by ascending wave per case set as shown in line (5a). Due to varying name lengths and missing values, variable lengths are set in line (5b). Prior wave values for caregiver first name, lastname, first and last name, date of birth, social security number, relation to child, and gender are retained in line (5c).

The code, in lines (5d) through (5f), checks data for the first observation in the case set. In line (5e), prior waves are assigned a value of -9 for the same caregiver flag, because the logical comparison rating will be assigned to the wave 4 observation. On the other hand, in line (5f), if the first observation is wave 4 data and the same caregiver flag has not been assigned, the flag is assigned a value of -3, for not applicable, and a message is written to the program log. If wave 4 is the first observation, it was the only completed interview.

In lines (5g) through (5h), new variables are created for the caregiver identifier information for the first observation in the case set. Short descriptions are provided beside the code for each new variable. Line (5i) ends processing of the first observation in the case set. Starting at line (5j), retained values of the new variables are compared against the corresponding wave 4 values, provided the wave 4 variable, sjboth, that is derived from the first and last names, has a non-missing value. In line (5k), the same caregiver flag is assigned a value of ambiguous if the names are missing in the prior wave data.

Logical comparisons for the same caregiver begin in line (5l). The caregiver is the same in the interviews completed at the prior wave and wave 4, if the wave 4 caregiver name matches the prior wave name in line (5l), either of the two wave 4 birth dates matches either of the two prior wave birth dates in line (5m), and either of the two wave 4 social security numbers matches either of the two prior wave social security numbers in line (5n). In line (5o), the same caregiver flag is assigned a value of 1, if the same caregiver criteria are met.

Logical comparisons for different caregiver start with line (5p). Line (5p) sets a cutoff of 3 for the minimum number of differences. If the same caregiver rating is missing in the wave 4 data in line (5q), the count of differences (score) is initialized to 0, in line (5r). In lines (5s1) to (5s12), the score is increased by 1 each time a difference is detected in: first name in line (5s1), last name in line (5s2), first birth date variables in line (5s3), wave 4 first birth date and prior wave second birth date in line (5s4), wave 4 second birth date and prior wave first birth date in line (5s5), second birth date variables in line (5s6), first social security number variables in line (5s7), wave 4 first social security number variable and prior wave second social security number variable in line (5s8), wave 4 second social security number variable and prior wave first social security number variable in line (5s9), second social security number variables in line (5s10), caregiver relationship to child in line (5s11), and gender in line (5s12). In line (5s13), if the prSameR flag indicates that the caregiver is not the same, the score is increased by one. If the score equals or exceeds 3, the flag is assigned a value of 2, for different caregiver, in line (5t). If a rating has not been assigned by this time and the caregiver names are not missing, the flag is assigned a value of –4, for ambiguous, in line (5u). If the caregiver names are missing in the prior wave data, an ambiguous rating is assigned in line (5v).

Figure 5. Logical comparisons of caregiver identifier information

```
*************************** main data step *********************;
data samecg;
set samecg;
by casegrp wave;  * ascending waves - prior wave comes first, then wave 4;      (5a)
length last $20 first $20 both $40 ssn1 $9 ssn2 $9 dob1 $8 dob2 $8 rel $20
   CG4Same 4 sjssn1 $9 sjssn2 $9 sjdob1 $8 sjdob2 $8 sjboth $40 ;               (5b)
retain last '' first '' both '' ssn1 '' ssn2 '' dob1 '' dob2 '' rel '' sex '';  (5c)
if w4comp = 'Y' then do;  * wave 4 was a completed interview;
*examples of data pre-cleaning are given in the data pre-cleaning discussion;
sjboth = compress(sjfirst||sjlast);
*these are renamed and retained from the prior wave for comparison;
if first.casegrp then do;                                                       (5d)
   if wave < 4 then cg4same = -9;                                               (5e)
   if wave = 4 and cg4same = . then do;                                         (5f)
      put '!!! lowest wave < 4, should have cg4same = -3 ' caseID= wave= cg4same= ;
      cg4same = -3;
   end;
   last = sjlast;         *last name;                                           (5g)
   first = sjfirst;       *first name;
   ssn1 = sjssn1;         *first social security number variable;
   ssn2 = sjssn2;         *second social security number variable;
   dob1 = sjdob1;         *first date of birth variable;
   dob2 = sjdob2;         *second date of birth variable;
   rel = sjrel;           *relationship to child;
   sex = sjsex;           *gender;
   both = sjboth;         *compressed first and last name;                       (5h)
end;                                                                            (5i)
else do;  *wave 4;
   if sjboth ne '' then do;  *have names this wave;                             (5j)
      *situation 1: no comparison due to missing names from prior wave;
      if both = '' then CG4Same = -4;                                           (5k)
      else do;  *situation 2: compare across waves;
   * if the first and last names, dob and ssn match, they are the
      same respondent;
   if sjboth = both                                                             (5l)
         and (
      (sjdob1 ne '' and (sjdob1 = dob1 or sjdob1 = dob2))
         or (sjdob2 ne '' and (sjdob2 = dob1 or sjdob2 = dob2))
               or (sjdob1 = '' and sjdob2 = '')     )                         (5m)
          and ( (sjssn1 ne '' and (sjssn1 = ssn1 or sjssn1 = ssn2))
         or (sjssn2 ne '' and (sjssn2 = ssn1 or sjssn2 = ssn2))
               or (sjssn1 = '' and sjssn2 = '')  )                            (5n)
   then CG4Same = 1;                                                            (5o)
   %let cutoff=3;                                                               (5p)
   *count the differences for situation 2;
   if CG4Same = . then do;                                                      (5q)
```

Figure 5. (continued)

```
      score = 0;                                                              (5r)
      if sjfirst ne first
       then score = score + 1;                                                (5s1)
      if sjlast ne last  then score = score + 1;                              (5s2)
      if sjdob1 ne '' and dob1 ne '' and sjdob1 ne dob1
       then score = score + 1;                                                (5s3)
      if sjdob1 ne '' and dob2 ne '' and sjdob1 ne dob2
       then score = score + 1;                                                (5s4)
      if sjdob2 ne '' and dob1 ne '' and sjdob2 ne dob1
       then score = score + 1;                                                (5s5)
      if sjdob2 ne '' and dob2 ne '' and sjdob2 ne dob2
       then score = score + 1;                                                (5s6)
      if sjssn1 ne '' and ssn1 ne '' and sjssn1 ne ssn1
       then score = score + 1;                                                (5s7)
      if sjssn1 ne '' and ssn2 ne '' and sjssn1 ne ssn2
       then score = score + 1;                                                (5s8)
      if sjssn2 ne '' and ssn1 ne '' and sjssn2 ne ssn1
       then score = score + 1;                                                (5s9)
      if sjssn2 ne '' and ssn2 ne '' and sjssn2 ne ssn2
       then score = score + 1;                                                (5s10)
      if sjrel ne '' and rel ne '' and sjrel ne rel
       then score = score + 1;                                                (5s11)
      if sjsex ne '' and sex ne '' and sjsex ne sex
       then score = score + 1;                                                (5s12)
      if prsamer = 2   then score = score + 1;                                (5s13)
      *criteria set at the top of the program;
      if score GE &cutoff then cg4same = 2;                                   (5t)
      *remaining cases are ambiguous;
      if cg4same = . then cg4same = -4;                                       (5u)
   end; *counting differences;
   end; *situation 2 - compare across waves;
   end;  *have names this wave;
   else do;  *missing names this wave;                                        (5v)
    cg4same = -4;
   end;  *missing names current wave;
end; *wave 4 or prior;
end; *else -- if w4comp = 'N';  * no action needed for wave 4 noncomplete;
run;
```

EXAMPLE OF SAME CAREGIVER DETERMINATION

Figure 6 shows sample data that would be assigned a rating of same caregiver, based on the logical comparison of wave 4 and wave 2 data. In line (6a), the names match. In line (6b), the first wave 4 birth date, 12251960, differs from the first wave 2 birth date, 12241960, by one day. Because the second wave 4 birth date has a missing value, no comparison is made using the second wave 4 birth date. However, the first wave 4 and second wave 2 birth dates match. In line (6c), the first wave 4 and second wave 2 social security numbers are missing. Because the second wave 4 social security number matches the first wave 2 social security number, the three comparisons are true and the case set is assigned a rating of same caregiver, or cg4same = 1.

Figure 6. Sample comparison - same caregiver

	Wave 2	**Wave 4**	**Comment**	**#**
First & last name	CANDYGA	CANDYGA	Names match	6a
First birth date	12241960	12251960	First wave 4 and second wave 2 dates match	6b
Second birth date	12251960	missing		
First social security #	777777777	missing	First wave 2 and second wave 4 numbers match	6c
Second social security #	missing	777777777		

EXAMPLE OF DIFFERENT CAREGIVER DETERMINATION

Figure 7 shows sample data that would be assigned a rating of different caregiver, based on the logical comparison of wave 4 and wave 3 data. From lines (7a) to (7c), the first names, last names, and first social security numbers differ. The score at line (7c) is three for these three differences. In line (7d), the comparison cannot be made with the missing wave 4 value for the second social security number. Different first birth dates increase the score to 4 in line (7e). Comparisons cannot be made in line (7f) due to a missing wave 3 value for the second birth date. Different relationships in line (7g) and genders in line (7h) increase the score to 6. In line (7i), the score is increased to 7 because the wave 4 interview flag (prSameR) indicates a different caregiver. With a total score of 7, the rating is assigned a value of 2, for different caregiver.

Figure 7. Sample comparison - different caregiver

	Wave 3	Wave4	Comment	Score	#
First name	ANDY	SANDY	No match	1	7a
Last name	NOONE	SOONE	No match	2	7b
First social security #	999999999	888888888	No match	3	7c
Second social security #	999999999	missing	Cannot compare	3	7d
First birth date	12251970	12251980	No match	4	7e
Second birth date	missing	12251980	Cannot compare	4	7f
Relationship to child	GUARDIAN	KINCARE	No match	5	7g
Gender	MALE	FEMALE	No match	6	7h
Interview flag	1	2	No match	7	7i

HUMAN REVIEW OF AMBIGUOUS DATA

For wave 4 and the most recently completed prior wave, the logical comparisons assigned an ambiguous rating to 820 case sets. Because a manual review of a small set of these ambiguous cases revealed that most of the ambiguous cases actually involved the same caregiver, and because the data analysts wanted to improve the usefulness of the same caregiver rating, two human reviewers were asked to independently examine identifying information for these case sets and assign a rating of same caregiver, different caregiver, or ambiguous.

Identifying information for the 820 case sets was written to a tab delimited file that was loaded into Excel. Figure 8 contains the code that prepared the tab delimited file. In line 8a, the data set generated by the logical comparisons, samecg, contains caregiver identifying information for case sets that were assigned the same, different, and ambiguous ratings. Because data are needed for only the cases that were assigned an ambiguous rating by the logical comparisons, a subset of the samecg data set is prepared that contains observations for completed wave 4 interviews that have an ambiguous rating. In line (8b), the subset contains only four variables that include: casegrp (the case set number), cg4same (the rating assigned by the logical comparisons), wave that is renamed wave4, and

w4comp (the wave 4 completion flag). Data set samecg is merged with its subset, by the case set number, in order to produce the desired data set. The data set is sorted in line (8c) in order to list the wave 4 data before the prior wave data, for each case set.

The data are written to a text file in line (8d), per case set in line (8e). With the first observation in line (8f), column names are written with a tab delimiter that is denoted by '09'x in line (8g), where '09'X is the ASCII value for a TAB character. Lines (8g) through (8i) label columns for the reviewer ID, review date, and rating to be assigned by the reviewer, while lines (8j) to (8k), label columns for the case and caregiver identifying information. From lines (8l) to (8m), the data are written with the tab delimiter per case set, with the wave 4 data listed first. In line (8n), a blank space is written after data have been written for the last data observation.

The text file was loaded into Excel. The two reviewers worked independently. Each reviewer recorded their ID, review date, and their rating, on the row with the wave 4 observation, in their copy of the spreadsheet.

Figure 8. Preparation of file with caregiver identifying data

```
* write a file of supporting data that can be loaded into Excel for review;
data samecg;
      merge samecg (in = in_s)                                              (8a)
            samecg4 (in=in_4 keep = casegrp cg4same wave w4comp
                  where = (w4comp = 'Y' and wave4 = '4' and cg4same = -4)
            rename= (wave = wave4));                                        (8b)
      by casegrp;  if in_s and in_4;
run;
proc sort data = samecg;
      by casegrp descending wave;                                           (8c)
run;
data _null_;
      file 'CG4Same.txt' lrecl=500 ;                                        (8d)
      set samecg ;
      by casegrp;                                                           (8e)
      if _n_ = 1 then put                                                   (8f)
            'Reviewer' '09'x                                                (8g)
            'Rating' '09'x                                                  (8h)
            'ReviewDate' '09'x                                              (8i)
            'Wave' '09'x                                                    (8j)
            'CaseID' '09'x      'Fldcode' '09'x
            'cidLastName' '09'x 'cidFirstName'  '09'x
            'Sex' '09'x         'SSN'  '09'x
            'Relation' '09'x    'FirstNameChg'  '09'x
            'FirstName' '09'x   'BirthMonth' '09'x
            'BirthDay' '09'x    'BirthYear' '09'x
            'FirstName'  '09'x  'MidName'  '09'x
            'LastName'  '09'x   'LegalFirstName'  '09'x
            'LegalMidName'  '09'x
            'LegalLastName' '09'x
            'MaidenName' '09'x  'Alias1'  '09'x
            'Alias2'  '09'x     ;                                           (8k)
put                                                                         (8l)
      ' ' '09'x
      Rating '09'x              ' ' '09'x
      wave '09'x                ZRID '09'x              SUMSTAT '09'x
      SJLast  '09'x             SJFirst  '09'x          SJSex '09'x
      PLF21A  '09'x             PHH6RELT '09'x          PHH4a  '09'x
      PHH4AA  '09'x             PHH7AM '09'x            PHH7AD '09'x
      PHH7AY '09'x              PNP5F  '09'x            PNP5M  '09'x
      PNP5L  '09'x              PLF1F  '09'x            PLF1M  '09'x
      PLF1L '09'x               PLF2a '09'x             PLF4a  '09'x
      PLF4b  '09'x       ;                                                (8m)
if last.casegrp then put ' ';                                               (8n)
run;
```

SAMPLE AMBIGUOUS IDENTIFIER DATA

Figure 9 shows examples of caregiver identifier data that were provided to human reviewers. In case set 9A, the last names differ and the prior wave name contains a middle initial. In case set 9B, the first names are variations of the same name, the last names differ, and the wave 4 relationship and birth date are missing. First names are variations of the same name, last names differ, and the prior wave birth date is missing in case set 9C, while the relationships and birth dates differ in case set 9D. Case set 9E has different first names, while case set 9F has variations of the same first name and a change in last name.

The human reviewers were given the following evaluation guidelines. Name variations and minor typographical errors should not be considered different information. In addition, names could change due to a change in marital status and the relation could change if the child was adopted or changed his/her living arrangement. However, caregiver gender and birth date should not change. When it is not clear, or there is too much missing data, the rating should be assigned as ambiguous. Based on the actual human reviewer ratings, case sets 9A and 9F would have been assigned a rating of same caregiver, while the other case sets would have been assigned a rating of ambiguous.

Figure 9. Simplified example ambiguous data

Wave 4	**Most recently completed prior wave**	**#**
MARY CAROLINA	MARY N. CARROLINA	9A
FEMALE	FEMALE	
PARENT	PARENT	
12/25/1960	12/25/1960	
BECKY KENTUCKY	REBECCA MARYLAND	9B
FEMALE	FEMALE	
MISSING RELATIONSHIP	PARENT	
MISSING BIRTH DATE	12/25/1960	
ANNE ALABAMA	ANN FLORIDA	9C
FEMALE	FEMALE	
PARENT	PARENT	
11/25/1960	MISSING BIRTH DATE	
SARAH GEORGIA	SARAH GEORGIA	9D
FEMALE	FEMALE	
GRANDPARENT	OTHER BLOOD RELATIVE	
12/23/1960	12/25/1957	
SUSIE VIRGINIA	TERESA VIRGINIA	9E
FEMALE	FEMALE	
PARENT	PARENT	
12/25/1960	12/25/1960	
BEV WEST	BEVERLEY W. VIRGINIA	9F
FEMALE	FEMALE	
PARENT	PARENT	
12/25/1960	12/25/1960	

RECONCILE REVIEWER RATINGS

Reviewer evaluations were imported and checked for completeness with the macro in Figure 10. In line (10a), the macro xls uses two macro variables, n and file. File is the Excel spreadsheet filename while n, the reviewer number, is 1 or 2. In line (10b), the macro imports data from the Excel spreadsheet (as the datafile) to a SAS data set that will be called rev1 for the first reviewer and rev2 for the second reviewer. Line (10c) keeps five variables that include: ID (the case set number), Reviewer (reviewer ID), Rating&N (rating assigned by the reviewer), caseID, and Wave. The reviewer rating variable is renamed in line (10d). A message is written to the program log if a rating is missing for a caseID, in line (10e). In line (10f), the program keeps wave 4 data that have non-missing caseIDs and ratings that are not missing and not -3 (for 'not applicable'). The data set is sorted in line (10g), in preparation for checking that the reviewers did not inadvertently delete a case set. In line 10(h), the data set with the reviewer's evaluations is

merged with the data set that contains all case sets that should have been reviewed. Line (10i) catches any cases that are missing in the reviewer's spreadsheet, while line (10j) outputs reviewer data that are expected. Data are prepared for a quick check by a human reviewer in lines (10k) through (10m). Data are listed for 25 observations for cases that should have been reviewed in line (10k) and for any cases that were not evaluated in line (10l). A frequency distribution of the reviewer ratings is generated in line (10m). The distribution allows a human reviewer to quickly check the range and total number of reviewer ratings. The macro code is closed in line (10n) and called with lines (10o) and (10p).

Figure 10. Check and reconcile the two reviewer ratings

```
********************** import and process review data ******************;
%macro xls (n, file);                                                    (10a)
   *spreadsheet from reviewer &N;
   proc import
      datafile = "c:\&file"
      out = rev&N ;                                                      (10b)
   run;
   data rev&N (keep = ID Reviewer Rating&N CaseID Wave);                 (10c)
      set rev&N (rename=(rating=rating&N ));                             (10d)
      if caseID NE '' and rating&N = '' then put
      '***** Case with no rating:  ' caseID= rating&N= wave= reviewer=;  (10e)
      if rating&N NE '' and caseID NE '' and rating&N NE -3
            and wave EQ 4;                                               (10f)
      ID = substr(left(caseID),2,6);
   run;

   proc sort data = rev&N ;
   by ID wave;      run;                                                (10g)

   * check IDs in the spreadsheet against those expected, ;
   data Rev&N._w4   NoRev&N.w4 ;
      merge rev&N (in=in_r) ambg4 (in=in_a);                             (10h)
      by ID wave;
      if in_a and not in_r then  output NoRev&N.w4;                      (10i)
      if in_a and in_r then output Rev&N._w4;                            (10j)
   run;

   title "Reviewer &N ratings for ambiguous CG4Same (top 25 obs)";
   proc print data = rev&N._w4 (obs = 25);                               (10k)
   run;

   title "Reviewer &N - Cases not found in spreadsheet for ambiguous CG4Same
      (top 25 obs)";
   proc print data = NoRev&N.w4 (obs = 25);                              (10l)
   run;

   title "Reviewer &N - rating distribution for CG4Same";
   proc freq data= rev&N._w4;
   tables Rating&N;                                                      (10m)
   run;
%mend;                                                                   (10n)
%xls (1, one_ambg_CCGs.xls);                                             (10o)
%xls (2, two_ambg_CCGs.xls);                                             (10p)
```

The project data set contained same caregiver flag values that were assigned by the logical comparisons. A data set was prepared that contained the rating and case identifying data for case sets that had been assigned the same rating by the two human reviewers. This data set was merged with the project data set, in order to reduce the number of case sets with ambiguous ratings. The code that created the data set that was merged with the project data set is shown in Figure 11.

In line (11a), two data sets are created for when the reviewer ratings agree (data set UpdCGSm4) and disagree (data set CG4_diff). The length of the flag variable cg4same is set in line (11b). Data sets for the two reviewers are merged in line (11c) by ID and wave in line (11d). If the ratings assigned by the two reviewers agree for the wave 4

case, in line (11e), the CG4Same flag is assigned the rating from the first reviewer, in line (11f), and output to the data set UpdCGSm4, in line (11g). In line (11h), data for case sets where the two reviewers disagree on the rating are output to the data set cg4_diff.

Figure 11. Reconcile reviewer ratings

```
******************** compare the two reviewers ratings *************;
data UpdCGSm4  (keep = ID NSCAWID CG2Same)
         CG4_diff (keep = ID NSCAWID Rating1 Rating2 wave);              (11a)
length GC4Same 4;                                                       (11b)

merge rev1_w4 (in=in1 keep=ID Rating1 wave)
       rev2_w4 (in=in2 keep=ID Rating2 wave);                           (11c)
by ID wave ;                                                            (11d)
if in1 and in2 and Rating1=Rating2 and wave =4 then do;                 (11e)
       CG4Same = Rating1;                                               (11f)
       output UpdCGSm4;                                                 (11g)
end;
else output cg4_diff;                                                   (11h)
run;
```

SUMMARY OF RECONCILED RATINGS

As shown in Figure 12, reconciliation of the human reviewer ratings reduced the number of cases that had an ambiguous rating from 820 to 30. The reviewers assigned the same rating for 801 case sets. Approximately 93 percent of the case sets were assigned a rating of same caregiver.

Figure 12. Comparison of Human Reviewer Ratings

Reviewers agree	Caregiver rating	Count	Percent of the 820 ambiguous case sets
Yes	Same	762	93.0
Yes	Different	28	3.4
Yes	Ambiguous	11	1.3
No	Not applicable	19	2.3

CONCLUSION

In our experience, determining if a respondent had changed, over the course of a longitudinal survey, could not be handled by logical comparisons alone. Due to last name changes, typographical errors, and inconsistencies, respondent identifying information needed additional comparison by human reviewers. Fortunately, SAS programming provided an automatic and reproducible way to conduct logical comparisons and process human evaluation of non-ideal data.

ACKNOWLDEGMENTS

We gratefully acknowledge guidance by R. Suresh, Jean Richardson, and members of the NSCAW project team.

The programming described here was funded by the U.S. Department of Health and Human Services Administration for Youth and Families, as part of the National Survey of Child and Adolescent Well-Being. (http://www.acf.hhs.gov/programs/core/ongoing_research/afc/wellbeing_intro.html)

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:
M. Rita Thissen
rthissen@rti.org
(919) 485-7728

Elizabeth Heath
eah@rti.org
(919) 485-2786

RTI International
P.O. Box 12194
Research Triangle Park, NC 27709-2194.