Blaise Instrument Design for Automated Food Coding

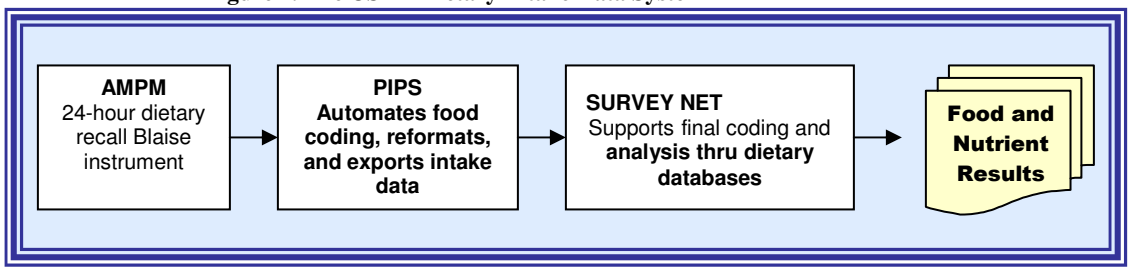
Ellen Anderson and Lois Steinfeldt, United States Department of Agriculture, Agricultural Research Service, Beltsville Human Nutrition Research Center, Food Surveys Research Group

1. Introduction

The Food Surveys Research Group of the United States Department of Agriculture has developed automated methods for collecting and processing food intake data. These methods are part of the Dietary Intake Data System (Figure 1), designed to efficiently collect and process high quality food intake data. The foundation of the system is the Automated Multiple Pass Method (AMPM) Blaise instrument, which is used to collect 24-hour dietary recalls. During the interview, individuals recall the foods and beverages that were consumed the day before the interview. Details about each food and beverage are collected as well as a description of the amount consumed. Information is also collected about the time of day the food was eaten, the name of the eating occasion, and where the food was obtained.

From the AMPM Blaise database, the intake data are extracted, reorganized, automatically coded, and reformatted by the Post Interview Processing System (PIPS). PIPS was developed with Visual Basic and the Blaise Application Programming Interface (API). These data are reformatted to be used by Survey Net, the computer-assisted food coding and dietary data processing system developed for use with USDA nationwide food consumption surveys. Trained food coders use Survey Net to search the Food and Nutrient Database for Dietary Studies (FNDDS) to code foods and amounts reported during the AMPM dietary recall interview. Government agencies and other researchers use the final data to examine nutrition and food safety issues affecting the U.S. population.

Figure 1. The USDA Dietary Intake Data System



Coding survey data via customary means (e.g. trained coding experts) often requires an inordinate amount of resources and may, at times, produce results that are subjective and erroneous (Speizer and Buckley, 1998; Biemer and Lyberg, 2003). Food coding can be especially challenging due to the variety of food items available in the U.S. marketplace and the diversity of the U.S. population. Automation of the food coding process may be a way to decrease data processing time and increase consistency and quality of food consumption data.

While the AMPM Blaise instrument was designed to facilitate the food intake interview, also considered was how the food detail data stored in the Blaise database would be extracted for manual and automated coding. With our method, automated coding of foods is accomplished as a data matching process between (a) the food question variable names and response values as reported in the AMPM interview, and (b) data in the Autocode table, which is part of an autocode pathways database. The series of questions and responses that describes a food establishes a food path for that food. If an exact match is found, then the food is automatically assigned a code. This method of automated coding was recently implemented as part of PIPS.

This paper will discuss how the design of the AMPM Blaise instrument facilitated the development and implementation of automated food coding in PIPS, as well as report on the successes and pitfalls of the automated coding process encountered thus far.

2. Overview of the AMPM Blaise instrument design

The AMPM Blaise instrument contains more than 2500 questions and more than 21,000 responses. Ninety-five percent of the questions are about specific food details, including the amount of food consumed. The types of questions asked for a food is determined by the food category assigned to that food. Foods were grouped into 132 categories represented by a unique 5 or 6 digit code. Each food category contains a series of questions designed to elicit specific details about that particular food. Certain questions are only asked of certain foods. In addition, certain questions may be skipped depending on the response to previous questions. Table 1 lists a variety of food categories and the number of food detail questions, not including questions about portion size, for each category. Collecting the proper level of detail for specific foods is an important first step in accurately coding food data. The capability of the AMPM Blaise instrument to ask only pertinent questions and collect all the necessary details for a particular food is a critical feature for implementing automated coding.

Table 1. Number of food detail questions (excluding portion size questions) associated with selected food categories

Food category	Number of food detail questions
Soda, pop, soft drinks	3
Milk	4
Yogurt	4
Cheese	8
Green salads	8
Infant formulas	8
Fruits, berries	10
Coffee	11
Mixed dishes	17
Meat sandwiches	45

Variable names for each question within a category were given a standard prefix. Initially, this was done to help the interviewer and food coder interpret the meaning of the variable names. For example, questions for the fruit category begin with the word "Fruit". Across food categories, the same question was given a standard suffix. For instance, the question "What kind was it?" in the fruit category is represented by the variable name "FruitKind". In the bread category, the variable name "BreadKind" represents the same question for bread. In addition to the variable name, each question in a food category is assigned an item number that indicates the order in which questions are asked. The item number consists of a three character mnemonic and a three digit number. Table 2 shows an example of these fields for food category 60060, yogurt. For the majority of food categories, item numbers were assigned systematically. Questions dealing with food details were assigned item numbers ending in 499 or less. The item number ending in 500 represents the standard question, "Did you add anything to the food?" Item numbers ending in 600 or greater represent questions about the amount of food consumed. The few exceptions to the item numbering order appear in the more complicated food categories such as sandwiches and salads. By standardizing the variable names and item numbers as much as possible, the process of identifying food detail questions that are important for automated coding was simplified.

Like the question variable names, the response values were written to be understandable to both the interviewer and the food coder. Open-ended responses were limited to "Other Specify" fields by the use of more than 90 lookup tables. The ready-to-eat cereal lookup table, for example, currently contains 289 possible responses. If the instrument required interviewers to enter responses into

open-ended fields, it is likely the same answer could be recorded in several different ways. For example, “Cap’n Crunch®” cereal could be entered as “Captain Crunch” or “Capn Crunch” or “Captain Krunch”, etc. With one selectable option for “Cap’n Crunch®”, the potential number of food paths is reduced. The standardization of questions and answers throughout the AMPM Blaise instrument proved to be important for implementing an autocoding system based on this exact text matching method. Limiting the variability of response entries helped to keep the food paths consistent and eliminated the need for complex parsing and decision making algorithms.

Table 2. Example of AMPM Blaise instrument field specifications for the yogurt food category

Item Number	Variable Name	Item Text
YOG005	YogurtFroz	Was it frozen yogurt?
YOG010	YogurtTypeMilk	Was it made from whole, lowfat, nonfat milk, or something else?
YOG015	YogurtLoCalSwtnr	Was it made with low calorie sweetener?
YOG020	YogurtFlav	What flavor was it? (Was it plain, vanilla, fruit, or something else?)
YOG500		GO TO STANDARD ADDITiON QUESTIONS, THEN CONTINUE WITH YOG600.
YOG600	YogurtAmt	How much yogurt did {YOU/SP NAME} actually eat?

3. The autocoding pathways database

The large number of questions and responses in the AMPM Blaise instrument produces an even larger number of skip patterns because the questions asked about a food depends on the responses to the previous questions. The number of possible food paths through the food detail questions is roughly estimated to be over 400,000. Identifying the most common food paths from the hundreds of thousands of possible pathways was the first step in setting up an automated coding system. The initial implementation of autocoding began in 2002 with the identification of commonly reported, simple foods and the creation of food paths linking to these foods. Using the 2002 AMPM Blaise instrument, 28 food paths were created and linked to 10 basic foods (whole, 2%, 1% and nonfat milk; diet and regular cola soft drinks; white sugar; catsup; apple; and banana). This initial implementation achieved an automated food coding rate of 11% of reported foods in the 2002 What We Eat in America, the dietary component of the National Health and Nutrition Examination Survey (NHANES). To pursue a more substantial rate of autocoding, a greater number of paths had to be identified and linked to food codes.

3.1 Development of the autocoding pathways database

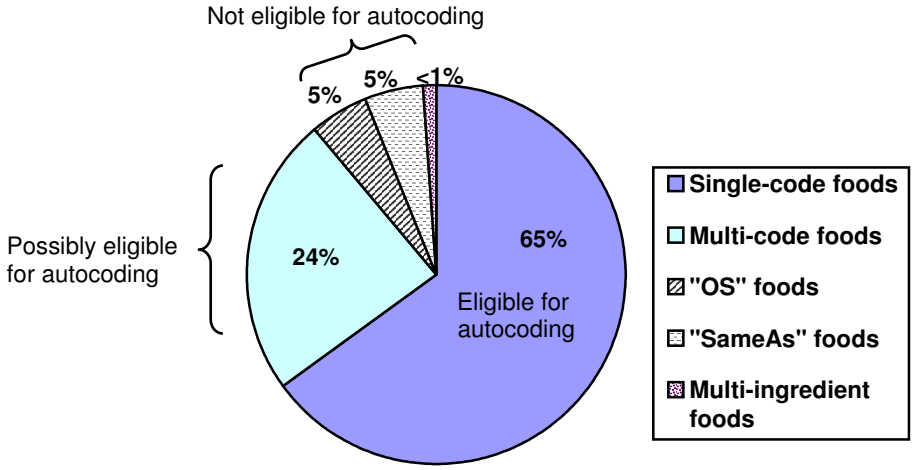
The development of the autocoding pathways database began with analysis of approximately 147,000 foods reported in the 2002 What We Eat in America survey and other food intake studies conducted by the Food Surveys Research Group. These data were collected using the AMPM Blaise instrument, coded by trained food coders, and thoroughly reviewed and monitored for quality control by nutritionists at USDA. A program was written in Visual Basic for Applications (Microsoft® Access 2000 (9.0.3821 SR-1)) that extracted food paths from the intake data and assigned a unique number to each food path within a food category. Extraneous information that does not contribute to the description of a food (e.g. the amount consumed) was not included in the food path identification process. Each combination of question and response that defines a food path was assigned a unique sequence number within that food path. Table 3 illustrates three food paths for the coffee category and the food codes that correspond to the food paths.

Table 3. Examples of food paths and food code links for the coffee food category

Food category		Sequence #	Question variable	Response value	Foodcode link
20010	1	1 2 3	CoffeeKind CoffeeCaffeine CoffeeForm	Coffee Regular Brewed	92101000 Coffee, made from ground, regular
20010	2	1 2 3	CoffeeKind CoffeeCaffeine CoffeeForm	Coffee Regular Instant	92103000 Coffee, made from powdered instant, regular
20010	3	1 2 3	CoffeeKind CoffeeCaffeine CoffeeForm	Coffee Regular Ready to drink	92100500 Coffee, regular, NS as to ground or instant

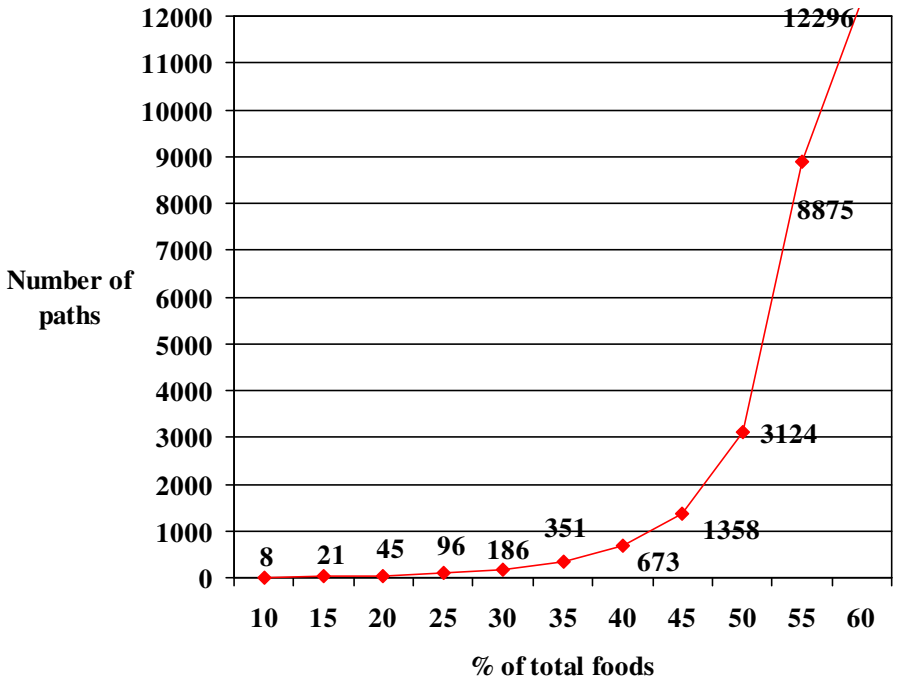
Certain foods were excluded from this initial construction of food paths. Complex foods that were coded with more than one food code were not considered for autocoding at this stage. These "multi-code" foods are items such as sandwiches or salads where individual ingredients (e.g. bread, meat, cheese, etc.) are coded as separate foods. Another category of foods that was not included during food path development was "Other Specify" or "OS" foods. During the AMPM interview, if a response is not listed as a choice in the Blaise instrument, then the interviewer may enter text in an open-ended "OS" field to record the exact response. "Same As" foods are foods that were previously reported by another member of a household. The AMPM Blaise instrument allows interviewers to record details such as "same as Mother's dinner, 7pm" in an open-ended "SameAsInstruction" field. A small percentage of foods have their main ingredients entered in an open-ended field during the AMPM interview. Homemade soups are an example of multi-ingredient foods which were deemed not eligible for autocoding at this initial stage. Accurate interpretation of these open-ended responses using a text matching method would be difficult. A breakdown of the approximately 147,000 foods analyzed during the autocoding developmental stage is shown in Figure 2.

Figure 2. Initial estimations of foods eligible for automated coding



From this initial analysis of 147,000 foods, we generated 12,304 unique food paths. Since the food intake data had already been coded, food codes that linked to the 12,304 food paths could be easily identified. We found that a relatively small number of food paths represented the majority of foods, as shown in Figure 3. Based on these data, we estimated that about 50% of total foods reported could be autocoded with approximately 3100 food paths.

Figure 3: Estimation of the number of food paths needed to achieve autocoding rate of total foods based on approximately 147,000 foods.



Before the food paths were incorporated into PIPS for autocoding, the data were reviewed by USDA nutritionists who verified that the food path was linked to the correct food code. Food paths were ranked by frequency to identify the paths that would capture the most autocoded foods. The food paths reported most often were reviewed first. Once a food path is approved as correct, it is added to the Autocode and FoodCodeLink tables used by PIPS for autocoding. The Autocode table contains fields for food category, path, sequence number, question variable and response value. The FoodCodeLink table links the food path to a food code. It contains fields for food category, path, and food code. These tables may be updated as necessary.

In addition to the review process, an adjudication test was performed before autocoding was implemented for the 2004 survey year. About 1500 dietary intakes were collected with the AMPM Blaise instrument and coded by trained food coders in 2002 and 2003 during a food intake study conducted by the Food Surveys Research Group. The Blaise intake data were re-processed using PIPS, which contained 1567 approved food path-food code links for autocoding. An autocoding rate of 50% of total foods reported was achieved for this dataset. The autocoded food items were then compared to the original food codes selected during manual coding. Out of the 12,224 autocoded foods, 98 food items (or less than 1% of all autocoded foods) did not match the original, manually-coded food code. Upon examination, some food path-food code links in the autocoding database were changed or removed. About half of the mismatched foods were due to additional interviewer comments that changed the outcome of the specificed food path. (The section titled “Limitations of

Autocoding", discusses these situations in greater detail.) This adjudication test verified that the autocoding process employed by PIPS produces accurate results in an efficient manner. The autocoding rate of 50% of total foods achieved with 1567 food path-food code links exceeded our estimates and expectations.

3.2 Implementation of the autocoding pathways database

Autocoding was put into operation for the 2004 What We Eat in America survey. In January 2004, we started with 1567 food paths in the autocoding database, which resulted in an autocoding rate of about 45% of reported foods. More food paths were added to the autocoding database during the course of the year. In June 2004, an autocoding rate of 59% was achieved. Currently, 2788 food paths are part of the autocoding database. In another food intake study conducted recently by the Food Surveys Research Group, we achieved an overall autocoding rate of 61% using the 2788 food paths in the current autocoding database. Again, these rates exceeded our estimations and expectations. Table 4 shows the 10 most frequently reported food paths since autocoding has been implemented.

Table 4: Top ten most frequently reported food paths being autocoded

Question Variable	Response Value
MilkKind	Whole milk
MilkKind	2% reduced fat milk
SugarKind	Granulated white sugar
CoffeeKind CoffeeCaffeine CoffeeForm	Coffee Regular Ground
FruitKind	Banana
SodaKind SodaCaffeineFree SodaType	Coke No (contains caffeine) Regular
FruitKind FruitType	Apple Fresh/raw
MilkKind	Skim/nonfat milk
ChipKind ChipType01	Potato chips (all flavors) Regular
MilkKind	1% lowfat milk

There is much agreement between the most frequently reported food paths and the most frequently reported food codes. However, there are some differences between their rankings. For example, lettuce and white bread are two of the most frequently reported food codes, but their food paths do not rank in the top ten. Because lettuce and bread are typically used as ingredients in sandwiches and salads, these foods are not autocoded very often with our current system. This is one of the limitations of our autocoding method that is discussed in the following section.

3.3 Limitations of the Autocoding Pathways Database

Although we have found that the AMPM Blaise instrument works successfully with the autocoding pathways database thus far, some limitations of this method exist. The autocoding process relies on text matching to link the food detail stored in the Blaise database to a food code. Changes to the AMPM Blaise instrument, such as added questions and responses and changes in the spelling of responses, may affect the autocoding food paths. Coordination between updating the AMPM Blaise instrument and maintaining the autocoding pathways database will be important. If a change occurs in the AMPM Blaise database, the autocoding pathways database will need to be reviewed to determine the effect, if any, on the food paths.

Open-ended responses are always a challenge to automated coding systems. The number of open-ended responses in the AMPM Blaise instrument was greatly reduced through the use of over 90 lookup tables. These lookup tables help to standardize the response options and facilitate the autocoding process. However, there are a few open-ended response variables ("Other Specify", "Same As" and multi-ingredient foods) that could not be handled with look up tables. These were mentioned previously in Section 3.1. The "Other Specify" or "OS" open-ended response is necessary when there is no response in the lookup table that corresponds to the answer given by the respondent. Text matching could be done on this type of open-ended response, but the variability of interviewers' entries would make pursuing this impractical with our method. We estimated that about 5% of total foods reported in a survey year contain an "OS" field (Figure 2). Frequently reported OS responses are added to the AMPM Blaise instrument. A similar situation exists for multi-ingredient foods where the ingredients in a mixed dish are listed in an open-ended response field. Only a small number of foods (approximately 0.4%) are handled in this way, so autocoding is not substantially affected. The "SameAsInstruction" open-ended field is an important feature of the AMPM Blaise instrument, allowing the interview to quickly record the same food consumed by more than one household member. Even though these foods cannot be autocoded with our method, this open-ended response saves the interviewer time and enhances the flow of the interview.

At any point in the interview, the interviewer may enter a text comment that gets stored in the Blaise database. Most often, these comments are entered to note time, eating occasion, or amount information volunteered by the respondent before the food details are collected. A small percentage of these comments contribute additional information to the food detail captured by the food path. Sometimes, these comments may contradict or change information collected in the food path. For example, a respondent reported consuming rye toast and the appropriate food path for rye toast was recorded. However, the interviewer entered a comment that said, "This was really wheat toast", which negates the food path recorded in the Blaise database. In this case, if there were a match in the autocoding pathways database for the rye toast, this food would have been inappropriately autocoded as rye toast. Fortunately, this type of situation does not happen very often. During the adjudication test described in Section 3.1, less than 1% of autocoded foods contained comments that would change the food path-food code link. The number of comments affecting food paths can be reduced further with additional interviewer training on when to enter comments versus when to change the response in the AMPM Blaise instrument.

Foods such as sandwiches and salads where individual components of the food are reported separately pose another challenge for the autocoding process. As it is currently designed, the Blaise database considers these multi-component foods as a single food. In order to capture the food paths that represent the individual ingredients, the multi-component food must be disaggregated into its component parts. For example, a tossed salad may disaggregate into iceberg lettuce, tomatoes, carrots, and croutons. Disaggregating these types of foods will increase the autocoding rate by an estimated 20-25%.

4. Conclusion

The creation of an autocoding pathways database and its implementation in PIPS for automated coding of food data has proven to be very successful. The success of this autocoding method is owed in large part to the design of the Blaise instrument and the ability to extract and reorganize data from the Blaise database. The amount of food data being captured with this automated method has exceeded our initial expectations. However, there is still room for improvement. In the future, we hope to explore other automated coding methods that would enhance the current system and contribute to our overall goal of high quality dietary data collection and processing.

Mention of commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture over others not mentioned.

5. References

Paul P. Biemer and Lars E. Lyberg, *Introduction to Survey Quality*. 2003, John Wiley & Sons, Inc.

Howard Speizer and Paul Buckley, "Automated Coding of Survey Data" in *Computer Assisted Survey Information Collection*, Edited by Mick P. Couper, Reginald P. Baker, Jelke Bethlehem, Cynthia Z.F. Clark, Jean Martin, William L. Nicholls II, and James M. O'Reilly. 1998, John Wiley & Sons, Inc.