Overview

**PERSPECT** is a facility for extracting more conceptual information from an exhaustively categorical data encoding scheme. The facility was written using the SAS macro facility and those procedures available in base SAS software on Personal Computers.  

**PERSPECT** allows the user to decode data that has been encoded ineffectively. The result is the creation of a list of codes with a corresponding array of decode values that are more conveniently employed in statistical research.

Encoding for digital processing often means coding numbers. While this makes sense in the computer's world, it easily becomes esoteric outside that world.

In addition, researchers often find that computerized data has been encoded in great detail, but with a flat design. The detail is more an inventory than a description. This presents a dilemma: uni-dimensional data in a multi-dimensional world. The codeset may be defined and fully documented, but the task of analysis is greatly aggravated when observation from a population is unique in some way. The ambition of research is the creation of meaningful summaries, not comprehensive lists.

This ambition helps in the communication of knowledge. A graphic artist applies an aesthetic rule-of-thumb to the presentation of data in a figure: an audience can apprehend from three to six categories. A greater number requires greater effort— which loses the audience. Analysts and researchers must also consider their audience.

A condition of meaningful or valid statistical results is the creation of categories that do not beg the research questions. If blind and a priori methods are used in the formation of the categories, this is less of a concern. However, the user is free to assess any code series. The generic **PERSPECT** interface avails the facility both to those interested in decoding objective, empirical concepts and those interested in subjective, experiential knowledge.

Keeping an Analogy in Mind

If every piece of information were stored in the brain as a distinct item, the mind would hardly comprehend, and never learn. The mind places information in a multi-dimensional, highly connected space. By so doing, it can not only answer a question on a random topic, but knows whether the question is even appropriate to the topic. "How tall is a light?" makes no more sense than "What color is an inch?"

The mind imposes its network structure to facilitate tasks like sequencing, priorities, and sentence construction. The mind is, in a sense, a relentless research tool that sciences the world about it. It can invent new dimensions and readily place objects in it. With imagination, even the height of light can be conceptualized, the color of distance, and other such koans.

**SIC and UWASIS**

Two cases will help illustrate the phenomenon of super-encoding, and its enigma. The first example comes from the field of economics (specifically, industry, business, and other enterprises). The Federal Office of Management and Budget has published the Standard Industrial Classification (SIC) Manual. Much research is focused on the activities of the industrial sector. Identification of trends and habits can drive investment and marketing, management, and other business research applications. In particular, the Internal Revenue Service and state revenue departments use SIC codes to help identify users and the volume of businesses involved in particular activities. OMB revises SIC periodically.

The second example is from the social service domain. When agencies apply for funding through United Way, they must report their activities regularly. Each agency is assigned a United Way of America Service Identification System (UWASIS) code. The agency is asked to select the UWASIS code that best describes its activities. Descriptions of each activity are found in a two-inch volume dedicated to UWASIS classification. The codeset was developed by a special committee, and is updated periodically. Codes are assigned in such a manner as to allow extension of the system as new activities and areas of service evolve.

Defending the Giant

The wisdom of encoding an issue is not fully articulated by its proponents (i.e., those who formulate the codes). The task of collecting detailed information on a topic, through exhaustive consideration of all salient categories, is truly

---

1. **PERSPECT** did not take advantage of Screen Control Language, or other facilities coming in Release 6.06, for a most practical reason: these were not available at this time at this site. The speed and control may suffer, but the audience is serendipitously broader. Of course, the system can be freely modified to fit the needs of each site.

2. A Zen Buddhist riddle used to trick the mind into a truer sense of reality.

3. United Way offers UWASIS as a general scheme that other agencies may use freely. A major trend in social services is the matching of service needs to available services, and adoption of a standard code would greatly facilitate this program.
expansive. The result should justify the development time and other expenses. The benevolent goal, of providing a standardized format for recording data, is fully respectable. Execution of that goal, when not informed by a researcher's perspective, is implicitly menacing to the its own end, research.

The method used for some coding schemes may be well justified. Perhaps the plan is, in fact, to detail the world into unique entities, as do social security numbers. Of perhaps there are only a few categories of interest, and storage space is expensive, or an existing system only has enough padding in a record layout to add a new three-digit field. Perhaps separate reports are actually required for fresh dimensions of the issue. Applying PERSPECTIVE to an issue does not supplant the original coding effort.

Slaying the Giant
Unfortunately, most standard coding schemes are expressed in a format that impedes access to the information. In the two examples above, UWASIS uses digits of each code in a quasi-Dewey decimal system. The first digit is a main category, the second a subdivision, and the last digits are specific agencies or industries. This allows two or three levels of detail in viewing the data, which allows analysis of group, subgroup, or individual "species" behavior. At best, these levels represent two or three concepts for analysis. In the case of SIC, the concepts are Division, Major Group, and Industry Group, which are actually different grains of one concept: Industry.

Inherent in the Dewey decimal approach is quite a bit of redundancy. When reading UWASIS, you start under a major category like Mental-Health-Services (the need the service addresses) and find that each of the subgroups generally reflect the mode of service delivery (a finite list, e.g., In-the-Home, in-a-Center, or In-an-Institution) and that the final code is often based on who receives the service (e.g., the Elderly, Disabled, or Children). Other major groups follow this same scheme—some more and some less truly. Thus, to reflect each separate concept, a new code is assigned for each unique combination with existing concepts. 

While a consistent rationale for classification is comforting, access is still a problem. For example, it is not possible to extract all services for the elderly without referencing an irregular list of individual codes. The only concept that can be easily accessed across all codes is the value nine (9), which refers to a miscellaneous, other, or undifferentiated category. Unfortunately, this category is not often informative, except with a hypothesis that stray or infrequent types behave differently from mainstream types.

Adding PERSPECTIVES
While the main concept reflected by a coding scheme may prove useful as a predictor or category for some purposes, a wealth of more significant information may lie just beneath it. The volumes (which document the code sets) for SIC and UWASIS contain concise and attentive knowledge—but in a narrative format. As yet, semantic patterns cannot be machined into existence. But a careful human analyst can review most any narrative and extract a profound design. The experienced researcher can even flesh out valuable information by looking at just the title for each code of a set. To illustrate, a research specialist knows whether: particular SIC codes are private, public, non-profit, or mixed; particular social services are certified by Medicaid.

Working Around the Codes
The research endeavor may continue, using only the narrow band of information most easily extracted from a coding scheme (e.g., super-groupings). In some cases, the code may be the only data which provides demographic information about its populations. Response or dependent data composes the remainder of a record. In other cases, sufficient data is collected outside the code to make an analysis meaningful without the code. But, odds are, much of this information overlaps with data implicitly available in the code.

The origin of coding problems appears at the design of data collection protocols. The researcher should see this phase as an opportunity to interject a more meaningful essay.

The Genesis of a Code:
- Management needs information from the Researcher to make a decision.
- The Researcher needs source data not in the client database, and grumbles to Management about having to collect additional data.
- Management requests that the survey information be added to the client database.
- A Systems Analyst does a system analysis, and recommends upgrading the client database.
- The Management Committee defines a total information package for the Systems Analyst.
- A Forms Committee designs a new client data collection instrument, which includes a checklist of client characteristics.
- The Systems Analyst gives specs to a Programmer.
- The Programmer optimizes the new client database fields by encoding client characteristics as sequential numbers, which are added to the data collection instrument.
- Only codes are entered into the system at data entry.
- The Researcher needs to update the original analysis, which requires manually extracting the original client characteristics from the system's codes.

If demographics are explicitly included in a data collection instrument, along with a standard, comprehensive, and difficult-to-use coding scheme, the researcher may risk the respect of the respondents. From a respondent's perspective, if ten minutes are spent reading a two page list of code descriptions to ensure that the correct code is recorded, and the next few questions ask for data made obvious by the code just given, the task becomes insulting or tedious. A simple example is to ask for the respondent name, the respondent gender, the spouse name, and the

*With n concepts each having \(I_i\) levels, there are \(P(L_i)\) possible codes, where \(P\) means Product, and \(i = 1 \rightarrow n\). The alternative is an array of \(n\) columns each having \(I_i\) possible values. The total possible values is \(S(L_i)\), where \(S\) mean Sum, and \(i = 1 \rightarrow n\). So 3 concepts with 4, 3, and 7 levels respectively would either take 84 uni-dimensional codes or a three-dimensional combination having 14 possible values (total, not for each dimension).
spouse gender—it is fairly likely that the spouse gender will be the opposite of the respondent.

Another risk in supplementing a code response with explicit questions is a contradiction in the responses. While this may be a good data check, it can be confounding. Some respondents may have a tendency to overthink the explicit if they infect it with what they believe are implicit issues from the context. "What did you do at work today?" will merit a different response if prefaced by "Your boss called a few minutes ago." An alternative is to tell the respondent that you are expressing their activities in both ways for your convenience. The purpose of explicit data on a survey is seldom for data checking when it accompanies a coding scheme, as the coding scheme is seldom decoded (as presented in this paper) in such a way as to enable checking.

Working with the Codes
Alas, how would you proceed when the system or survey is completed, after a year or more of planning and programming, and not easily revised to incorporate more explicit data? Say a standard coding scheme was selected by a committee, assuming the research staff could extract all the demographics needed from the codeset. The researcher stores in the face of a monolithic task: "Explain the survey results." The researcher opens the code documentation and begins recording any meaningful groupings of the codes that can be construed (decoding).

The decoding process is best approached in a systematic manner. The features of each code should be considered to extract relevant descriptors (called "decodes," in the system, since they reverse super-'encoding'). Then each new code concept should be applied consistently to all the codes.

The PERSPECT facility mechanizes this process. Once a system2 of decoding is begun in PERSPECT, the integrity of the system must be assured. PERSPECT attempts to maintain such integrity through an intuitive interface and a highly supervised decoding process.6 PERSPECT then provides a mechanism for applying the new decode PERSPECTives to the originally encoded values.7

A System PERSPECTive
The first step in using PERSPECT is to establish the environment in the SAS Display Manager session. This includes modifying the AutoExec.SAS file in the ISASRoot directory to initialize the facility, creating a PERSPECT directory, and copying AutoSys.SAS (called by AutoExec.SAS) to the PERSPECT root. When the SAS system is invoked from the PERSPECT directory, required system macros are loaded to memory.8 The PERSPECT statement will then invoke the facility as its functions are required.

AutoExec.SAS:
/*
- Append these lines to Current AutoExec.SAS
in ISASRoot directory.
Set DOS Path to find SAS.EXE in ISASRoot.
Change to PERSPECT root directory and invoke SAS.
*/
LibName AutoSys 'AutoSys.SAS';
%Include AutoSys;

The PERSPECT menu displays actions that are appropriate for each point in the session. The menu forces the user to comply with a strict interface, allowing escape by no other means than those shown on the screen. This prevents the user from straying and getting lost outside the PERSPECT menu.

Defining a Codeset
At first, a new codeset system must be named and initialized. A separate directory is used to contain and manage each codeset. A codeset contains one variable named after the system for the list of codes in the set. Code values can be added, edited, or removed. In addition, two or more codes can be merged or one code can be split into multiple codes.

Whenever any code action occurs, all decode variables that are currently defined appear for processing (see decode options below). The user is forced to fill out the system so that the integrity of the codeset is protected.9 Otherwise, holes may go unfilled as the system grows. The user is thereby forced to think about a code or decode in context, fostering a more comprehensive system.

Adding a new code will trigger the routine for filling out values for existing decode variables; combining multiple codes will display all prior decode values as new values are supplied for the emergent code.

Lastly, a type option allows the user to change the definition of the code, which includes a description. The description is used throughout PERSPECT to assist in identifying codes on the screen.

Defining a Decode
The list of options for decodes is consistent with those under the code menu. The decode fields and values can also be added, edited, or removed. Each can also be merged or split. In addition, a type option for changing the definition of a decode is provided. The decode definition includes a description or label, but also includes the variable type and length. Decodes fields can be character or numeric.

A researcher should attempt to create as small a set of decode variables as possible while discriminating the implicit differences between each code of the codeset. In extremes, the number of decodes may equal the number of codes (a set of dummy variables) or one decode may

5PERSPECT allows the user to name a subdirectory that will contain a decoding system. A system may correspond to one codeset, or alternative systems may be created and maintained separately using the same codeset.
6Using SAS Display Manager windows and the SAS macro facility.
7Using the SAS Format procedure and the SAS macro facility.
8The amount of expanded memory available to the SAS system may affect your ability to load the facility.
9A special response, not "missing," but rather "not available," is designated to temporarily skip cells.
contain as many levels or unique values as the original codeset (a simple re-expression). Neither result is desirable for supporting the research endeavor.

**Outputting Decode Formats**

The next step, after a codeset is fully decoded, is to put the decode array into the format most useful to the SAS system. The format used happens to be SAS Formats. A Format Library is maintained for each SAS dataset which contains a new Format for each decode.

**Applying Decodes to Data**

The last step is to apply the decodes to a SAS dataset that utilizes a specific codeset. This menu option simply adds the decodes as new variables and matches the decode values to each observation. The descriptions of the decodes are used as labels in the dataset to assist in documentation. One, several, or all decodes can be incorporated into the specified dataset.

**Emergent Properties**

**Classification Formalized**

By translating a codeset into meaningful decodes, the classification rationale is formalized, or made explicit. The success of this process depends upon the decodes chosen. If the values of all the decodes place a specific code in a unique space—each decode is a dimension of a hypothetical space—the decode set is complete and data can be mapped from code to decode and successfully back again. If, however, the same set of decode values correspond to more than one code, two possibilities arise:

1. The decodes are not complete, and another decode should be added to differentiate the codes.
2. The codes contain duplication or the distinction between overlapping codes is not a valuable concept.

The inverse of duplication is a gap. A gap in the original codes may become apparent when certain logical combinations of decode values do not occur in the codeset. In this way, decoding can actually aid in identification of weaknesses in the codeset, which may provoke a change in the data collection instrument.

**Anticipating Expert Systems**

The ability to map back from decodes to codes in a complete system has still greater implications on data collection methods. Currently, a survey respondent must choose a code from a long flat list. If they are lucky, they are allowed to give a free-form response to the question, "Who are you?" But then a rater or data entry operator must encode from the free-form survey response. Neither position is better, since the respondent does not fully understand the coding scheme, and even an expert rater does not know enough about the respondent to apply the scheme reliably.

If a set of questions corresponding to the decodes were included on the survey, the responses could be translated to the codeset (the value of which is, though the survey instrument would change, the data entry and other established computer programs could retain the codeset without modification). Depending on the complexity of the mapping process, a rater could use an assignment matrix, or a separate computer program could be used. The program may be possible in a traditional programming language, but expert systems techniques make the task much easier to implement and maintain. The Texas Comptroller's Office\(^{1}\) has already developed a classification system for SIC codes using an expert system development shell.

**Familiarity with the Issue**

The most welcome effect of using PERSPECT to decode is the transformation upon the researcher. To open a code book the size of SIC or UWASIS requires a great deal of bravery—and a still greater ego to think that one can form a comprehensive understanding of a population by reading straight through once. Hopefully PERSPECT makes management of the decoding process less threatening, encourages the researcher to open the book, and involves the researcher in the issues of the domain.

Thus, a truly new research perspective is formed through sheer exposure to the population.

**The Boon to Analysis**

Expansion of a codeset to decode variables will be followed by an analysis using traditional statistical techniques. In particular, factor, cluster, and regression analysis will have rich concepts to model, facilitating interpretation.

A current Tennessee Department of Revenue application illustrates the value of decoding SIC codes. A Stepwise Regression was used to model tax audit potential. A Principia Industrial Activity (PIA) code, derived from SIC, was available for the tax base. To deal with the diversity of the population, a separate model was developed for each major PIA group. Most of the independent variables available were amounts from the tax returns. Access to concepts like activity, type of goods, public contact, and richer demographics were not available for modeling. The problem was not intractable, but resulting models had high variability and predictive power suffered.

Using PERSPECT, the PIA code is being broken out into meaningful dimensions by a specialist familiar with the paying population. The dimensions will offer cleaner explanatory power, and resulting models will reflect cogent theories about the population's behaviors. It will then be easier to establish typologies, make inferences about the qualities that affect tax payment, and better address taxpayers' needs.

**Conclusion**

The problems of encoding and decoding motivated the development of the PERSPECT facility. However, the resulting system addresses a variety of applications:

1. **PERSPECT** offers the interactive management of formats and an interface to facilitate their application to data.

---

\(^{1}\)Contact Katha Richards, Texas State Comptroller of Public Accounts, Information Resource Center, 111 East Seventeenth Street, Austin, TX 78774.
SAS/ASSIST currently has no interface to the Format procedure.

2. A new standard classification codeset could be invented, with PERSPECT ensuring that it is comprehensive and can be broken back down for analysis later.

3. The oft-used Dummy Variables can be defined using PERSPECT, and placed in a SAS dataset without explicitly adding statements like those shown in DummyPgm.SAS.

4. Limited "data entry" can set up and done by a novice SAS user. The overhead of PERSPECT makes this a somewhat costly alternative to creating your own Window and Display definitions, but some of the interface and programming techniques are transferrable.

5. Pieces of the System that automate management of Filename and Libname definitions can be applied to any project.

These and other functions can be used alone or tied to other user applications. The specifications given for this system can be completely re-coded and readily extended using Screen Control Language or integrated into the SAS/ASSIST environment.

An effort has been made to code PERSPECT to maintain a degree of portability. The system will no doubt change as it is tailored to our current application. The current documentation and code for PERSPECT are available from the author upon request:

* Vincent Tillman, Statistical Analyst
  TN Department of Revenue
  Management Information and Planning
  Andrew Jackson Bldg., Room 1210
  500 Deaderick Street
  Nashville, TN 37242.

End Notes
SAS and SAS/ASSIST are registered trademarks of SAS Institute Inc., Cary, NC, USA.