DOCUMENT RESUME

ED 264 755 HE 018 928

AUTHOR	Saavedra, Pedro; Kuchak, JoAnn
TITLE	Quality Control Analysis of Selected Aspects of Programs Administered by the Bureau of Student Financial Assistance. Error-Prone Model Derived from 1978-1979 Quality Control Study. Data Report. [Task 3.]
INSTITUTION	Applied Management Sciences, Inc., Silver Spring, Md.
SPONS AGENCY	Office of Student Financial Assistance (ED), Washington, DC.
PUB DATE	29 Aug 80
CONTRACT	300-79-0742
NOTE	53p.; For related documents, see HE 018 926-927.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Accountability; College Students; Dependents; *Disclosure; Eligibility; Error Patterns; Federal Programs; *Financial Aid Applicants; Grants; Higher Education; Interviews; Management Information Systems; *Models; Program Administration; *Quality Control; Research Methodology; Selt Supporting Students; Statistical Analysis; *Student Financial Aid
IDENTIFIERS	*Basic Educational Opportunity Grants; *Fraud

ABSTRACT

An error-prone model (EPM) to predict financial aid applicants who are likely to misreport on Basic Educational Opportunity Grant (BEOG) applications was developed, based on interviews conducted with a quality control sample of 1,791 students during 1978-1979. The model was designed to identify corrective methods appropriate for different types of applicants. During interviews, applicants provided documentation regarding the information they provided on the Student Eligibility Report (SER). An analytical method known as sequential search analysis or automatic interaction detection (AID) was used to classify applicants into groups. About 32 percent of the students reported themselves as independent and almost 68 percent as dependent on the SER. The sample included 52.3 percent misreporters: 20.1 percent underclaimers and 28.4 percent overclaimers. Ten dependent and seven independent groups emerged from the model. The 17 groups are discussed, and data for each group are provided on Student Eligibility Index values, assets, taxes, and percent of error. Misreporting patterns of the student groups are analyzed, and edits for underclaimers are briefly discussed. Recommendations for program improvement and for further research are offered, and the research methodology is detailed. (SW)

Applied Management Sciences

Quality Control Analysis of Selected
Aspects of Programs Administered by the
Bureau of Student Financial Assistance

Error-Prone Model Derived from 1978-1979
Quality Control Study Data Report
[Task 3]

Applied Management Sciences, Inc. ◆ 962 Wayne Avenue ◆ Silver Spring, Maryland 20910 ◆ (301) 585-8181

Prepared for:

Office of Student Financial Assistance
Department of Education

G-170

ERROR-PRONE MODEL DERIVED FROM 1978-1979 QUALITY CONTROL STUDY DATA REPORT

August 29, 1980

In Response to:

Contract No. 300-79-0742

APPLIED MANAGEMENT SCIENCES

G-170

QUALITY CONTROL ANALYSIS OF SELECTED ASPECTS OF PROGRAMS ADMINISTERED BY THE BUREAU OF STUDENT FINANCIAL ASSISTANCE

This report was prepared pursuant to Contract No. 300-79-0742. The names of the persons employed by the contractor with managerial or professional responsibility for such work are as follows:

Ms. JoAnn Kuckak, Project Director

Dr. Pedro Saavedra, Technical Manager
(Data Analysis and Report Preparation)

Applied Management Sciences, Inc.
962 Wayne Avenue, Suite 701
Silver Spring, Maryland 20910

Office of Student Financial Assistance
Department of Education

TABLE OF CONTENTS

EXECUTIVE SUMMARY

LIST OF EXHIBITS AND TABLES

EXECUTIVE SUMMARY AND RECOMMENDATIONS

Background

The Basic Educational Opportunity Grant (BEOG) is the largest of the student financial aid programs administered by the Office of Student Financial Assistance (OSFA). BEOG was authorized by Title IV of the Higher Education Act of 1965, and is second only to the Guaranteed Student Loan program in total student compensation. Students who receive BEOGs are also eligible for other types of State and Federal financial aid; thus, the BEOG serves as a cornerstone of aid to students who are eligible based on a formula which determines financial need. The result of this formula calculation is a student eligibility index (SEI) which, together with cost of education at the institution the student plans to attend, and the student's enrollment status (full time or part time), determines the amount of the BEOG to which the student is entitled.

One of the OSFA's management objectives is to reduce the amount of student misreporting on BEOG applications, or to increase the rate of awards based on accurate information. To meet this objective, OSFA has instituted several procedures for quality assurance. During the 1978-1979 academic year a quality control study was conducted in which grant recipients were contacted and interviewed in an effort to determine the degree to which they had misreported in their applications. A certain proportion of applicants is selected randomly every year for a pre-award validation of selected items from their application by the financial aid administrator (FAA) at their institution. A previous study developed a model designed to identify error-prone applicants as detected through the pre-award validation process for the 1979-1980 academic year. This investigation has developed a similar model designed to identify error-prone applicants as detected through the quality control interview conducted during the 1978-1979 academic year.

The Error-Prone Model

The application of a sequential search technique to the quality control study sample was successful in identifying error-prone applicants and segregating them into "identifiable" groups. Highlights of the model follow:

- Seventeen groups have been identified. They differ from each other in
  - -- the proportion of applicants who seriously misreport.
  - -- the average amount by which the SEI is off.
- The model was more effective in identifying applicants misreporting to their disadvantage than in identifying applicants misreporting to their advantage.
- One group was identified in which 41.0 percent of all applicants misreported dependency status. The total figure is 3.9 percent of all applicants and 9.8 percent of independent applicants.
- Nontaxable income was found to be a greater and more systematic source of error than previous studies had indicated.
- Information was found suggesting the possibility that upperclassmen have up to now systematically underreported summer and part time earnings, though the magnitude of these omissions is not large.
- Applicants entitled to the maximum awards in general misreport less than applicants who are entitled to smaller awards. This result is consistent with other studies.
- Some indication exists that applications processed after the end of August differ from those processed earlier and include slightly greater proportions of misreporters. This would affect somewhat the generalizability of aspects of the VEAPS error-prone model.

The data used in this investigation has the strong potential for antifactual results, and thus all findings should be viewed as tentative. Recommendations for program management and recommendations for further research are presented in Chapter 4.

1
INTRODUCTION

1.1 Purpose and Scope

The Office of Student Financial Assistance has contracted with Applied Management Sciences, Inc., to provide analyses of pre-existing information on Basic Educational Opportunity Grant applicants which will yield findings for management of the program, as well as findings specific to the characteristics of students who misreport on their applications. In order to provide information needed by OSFA, three separate ~~error-prone models designed~~ to identify applicants likely to misreport on the basic grant applications will be conducted:

- An error-prone model predicting applicant error detected through validation.
- An error-prone model predicting applicant error detected through an IRS tape match.
- An error-prone model predicting applicant error detected through a quality control verification interview.

The first model was presented as part of the Secondary Analysis of Validation, Edits and Application Processing Systems (VEAPS) Task 1 Report. The second model will be included in the IRS-BEOG Data Tape match Task 2 Report to be presented later this year. This document presents the third model, as requested in an optional task under Contract No. 300-790742 (AMS G-170).

An overview of the study is found in Chapter 1, along with the necessary background and the methodology used. Chapter 2 presents the model in its final form. Chapter 3 describes each of the groups into which the applicants were classified as a result of the analysis. Chapter 4 presents a summary of the findings, presents recommendations and compares the results to those of the VEAPS error-prone model.

1.2 Key Study Objectives and Questions

The broad objectives of the error-prone modeling process undertaken in this investigation are:

- to investigate the characteristics of students most likely to misreport information on their application
- to provide a model that is sensitive to future changes in the BEOG program (such as different restrictions in eligibility determination)
- to identify effective predictors of misreporting which may be used in the IRS model and in future investigations
- to provide a means of identifying such applicants and of minimizing the impact of their misreporting

The specific objectives of the error-prone modeling component include the following:

- to provide a means of estimating the likelihood that an applicant is misreporting
- to provide a method of selecting applicants for validation
- to provide information for management improvements specifically related to:
  - -- the validation system
  - -- the edit system
  - -- the original application forms
  - -- reduction of drop-outs among students entitled to funds and
  - -- a global quality assurance strategy

In more informal terms, the model presented in this report attempts to address the comment, "different corrective measures ~~are needed~~ for different kinds of applicants," by providing a useful operational definition of the term "kinds of applicants."

1.3 Background of the Study

As part of a quality control study, Macro Systems, Inc. and Westat, Inc. set out to interview 2,309 students representing 1.35 million recipients of Basic Grants in 1978-1979. Applicants were interviewed and asked to provide documentation regarding the information provided in the Student Eligibility Report (SER).

While the documentation provided was recorded, data obtained from these interviews were coded to calculate discrepancies in Student Eligibility Index (SEI) and in award without regard for the nature of the documentation provided. For 27.0 percent of the applicants actually interviewed, information in at least one SER field was unavailable at the time of the interview. The value of the field in the SER was used in such cases to calculate the SEI based on the interview. If in the process of an interview an applicant was found to have misreported dependency status, this fact was recorded, but the applicant's parents were not interviewed, making it impossible to establish whether the applicant was entitled to any grant and if so, to determine the magnitude in dollars or SEI points of the error.

Macro Systems, Inc. provided Applied Management Sciences data tapes with information on the SER values, the verified values and the discrepancies in each field, as well as total discrepancies in SEI and in award dollars. The documentation provided by Macro proved inadequate at first, producing several false starts. In the end, the analysis used three kinds of variables provided by Macro:

- Variables from the SER, including MDE source, process date, type of school, control of school and institution size.
- Discrepancies between the interview and the SER expressed in SEI points, as well as total discrepancies in SEI points and in award and an indicator of dependency status misreporting and of failure to interview.
- Sampling weights.

These are the only variables used in the analysis. Their values were accepted and assumed to be accurate. Whatever artifactual effects influenced these values will have also influenced this study.

The Macro-Westat sample excluded applicants who filed applications after October 1978. Of the 2,309 applicants in the sample, 1,815 provided sufficient information to be used in at least one of the analyses. Of these, 76 had misreported their dependency status and, thus, had to be excluded from some analyses. The 1,815 applicants represented 1,272,512 applicants, 49,110 of which were represented by the dependency status misreporters.

1.4 Research Methodology

A brief description of the research methodology will be presented here. For a more detailed description, see Appendices A and B.

An analytical method known as sequential search analysis or automatic interaction detection (AID) was used to classify applicants into groups which differ as much as possible in a dependent variable relating to their response to validation. The groups are defined in terms of a set of predictor variables. Only variables which could be obtained from the SER, or through knowledge of the institution selected were used as predictor variables. Some of the variables were obtained through the algebraic manipulation of two or more SER fields. AID first splits the sample into two groups which are as different as possible, and continues this process for each resulting subgroup.

The computer program used in the analysis was AID3. This is a different version from the one used in the VEAPS report. The absence of applicants who failed to re-enter the system in the quality control sample made the use of a categorical dependent variable, as in the VEAPS report, unnecessary. Since AID3 permits the use of both continuous and dichotomous variables, it was selected.

Four criterion variables were used. The most useful one proved to be the difference between the SEI from the SER and the SEI which would have resulted had the values obtained in the verification interview been used. This difference was used as a continuous criterion variable without regard for the direction of the discrepancy. A second criterion variable was that portion of the discrepancy which could be attributed to fields subject to validation (regardless of comments) in the 1980-1981 academic years. Both of these criteria ignored dependency status misreporters (i.e., model changers) so two dichotomous variables were used for separate AID analyses, each combining persons for whom the corresponding continuous variable was greater than or equal to 50 with persons whose dependency status model derived from the SER was different from their dependency status model obtained from the interview. Appendix B operationally defines these variables.

Forty predictor variables were included in the AID analyses. Due to the findings of the VEAPS Report concerning PEC A-6, taxes were computed using the adjusted gross income, itemized deductions, marital status and exemption figures reported in the SER. The value of the computed taxes and their discrepancy with the reported taxes, as a percentage of the larger figure and in dollars, were included among the predictors. A new variable marking the presence of savings, investment, farm or business when adjusted gross income was equal to the sum of the earned incomes, was also used in the belief that it indicated likely omission of interest, dividends or profits. This variable is referred to as unreported likely unearned taxable income (LUTI). Exhibit 1.1 presents the list of predictor variables. Since the experience with the VEAPS report indicates that blanks or missing values are not good predictors, assumptions were applied where appropriate.

Four AID analyses were conducted, one for each criterion variable. In each case, the first split was forced on dependency status. The four resulting models were compared and a composite model was formed using the model from the first criterion variable (SEI discrepancy) as anchor with modifications supplied from the other three models (see Appendix B).

While the AID analyses were conducted without sampling weights, all other analyses, including reported percentages, used sampling weights in order to conform to the figures reported in the Macro-Westat report.

EXHIBIT 1.1: PREDICTORS USED IN AID ANALYSIS

1. Marital status
2. Household size
3. Number in household in post-high school education
4. Exemptions **
5. Nontaxable income **
6. Adjusted Gross Income
7. Father's or applicant's earned income
8. Mother's or spouse's earned income
9. Presence of absence of each source of earned income (four values, one for each possible combination
10. Reported taxes
11. Taxes computed from other SER information **
12. Tax discrepancy as a proportion of computed or reported taxes (whichever is greater)
13. Tax discrepancy in dollars **
14. Itemized deductions
15. Medical or dental expenses
16. Casualty-theft losses
17. Unreimbursed tuition
18. Home value **
19. Investment value
20. Home debt
21. Business value
22. Farm value
23. Applicant's resources (for dependents only)
24. Veteran's benefits (amount)
25. M.D.E. source
26. Age **
27. Year in school **

* Appears in intermediate model

** Appears in final model

EXHIBIT 1.1 (continued)

28. Process date **
29. Total income (AGI + NTI + Vet benefits x number of months)
30. Expenses
31. Assets (home, farm, business and investment values minus debts plus savings and applicant's resources) **
32. Nontaxable income as a proportion of total income
33. Presence of unreported likely unearned taxable income (LUTI) (farm, business, investments or savings combined with AGI = Earned income) **
34. Income as percentage of income + assets *
35. Type of school *
36. Control of school *
37. Size of institution *
38. Eligibility index **
39. Expenses as a proportion of income
40. Dependency status **

* Appears in intermediate model

** Appears in final model

1.5 Limitations and Strengths

Unlike the VEAPS report, the sample in this study was too small to permit a replication sample. Thus, it is not known whether any of the results might be idiosyncratic to the sample.

The absence of non-recipients and ineligibles could produce artifactual results. Persons who did not cooperate with the interview present an additional difficulty.

The data collection method with large proportions of undocumented data and the possibility of "false errors" (i.e., data reported correctly the first time, but forgotten or lost and improperly reconstructed at the time of the interview), could lead to questionable results. Also, it should be kept in mind that errors detectable through quality control verification may not be detectable through validation.

On the other hand, the study can produce results which can be incorporated into edits, and possibly validation procedures, provided the results are closely monitored. The focus on recipients is a strength as well as a weakness, since it eliminates the problem of applicants in the sample who have not yet become recipients, but who may later appear in the recipient file. This was a major weakness of the VEAPS report, but an inevitable one if applicants who did not go through with validation were to have been identified.

The groups identified in this report are necessarily less numerous (due to sample size) than those of the VEAPS report, and this makes selection of a small percentage of applicants less feasible. On the other hand, the model is more parsimonious, increasing the likelihood of straightforward interpretation the results.

The next chapter will discuss the major findings of this study.

2
OVERVIEW OF THE MODEL

2.1 General Characteristics of the Sample

A total of 1,791 unweighted cases were included in the analysis, representing 1,272,522 cases in the population. Of these, 32.2 percent reported being independents in the SER and 67.8 percent reported being dependents (all percentages refer to weighted figures). Of these, 3.9 percent were found to have misreported dependency status, and thus had to be excluded from parts of the analysis (such as calculations of mean error and net cost), since data was not collected from the parents of persons claiming to be independents in the SER.

Using the same conceptual definition of over-claimers and under-claimers as was used for the error-prone model of the VEAPS report (i.e., persons whose SEI was off by 50 points or more), this sample included 20.1 percent under-claimers and 28.4 percent over-claimers for a total of 52.3 percent misreporters (including model changers).

The mean error for the entire sample was 224 misallocated SEI points per applicant. These can be broken down into 164 misallocated points attributed to validation fields and 60 points attributed to nonvalidation fields. Also, the mean under-allocation of SEI points was 140 and the mean over-allocation was 84. On the average, each applicant received $49 more than he was entitled to (based on applicant error only), but the mean amount of misallocated dollars (combining overpayments and underpayments) was $134.

These figures are much larger than those found in the VEAPS report. Note that averages are taken over all applicants except model changers. There are various possible reasons for this difference. For instance, the quality Control study pursued fields not tapped by validation. It also treated as factual second estimations for a field that had been estimated in the first place. Quality control data were collected by interviewers trained for a one shot study; validation data are gathered routinely by financial aid administrators who are very familiar with BEOG. .

In addition, both the applicant and FAA have the SER figures available during validation, whereas the quality control interview was conducted without reference to the original SER. This may result in underestimation of error through validation and overestimation through the quality control procedure. For instance, when the applicant reported a value on the application and failed to provide documentation and report any value, the QC interviewer was not aware of the inconsistency and the "verified" value would be zero. The validation process, on the other hand, requires the FAA to question such inconsistencies and obtain an explanation to resolve them.

2.2 Overview of the Model

Ten dependent groups and seven independent groups emerged from the model in its final form. Exhibit 2.1 presents a definition of the seventeen groups, along with the percentage of the population each group constitutes and the percentage of cases where verified SEI was discrepant by at least 50 points from the SEI obtained from the SER, or where the verified dependency status model differed from the SER model. Exhibit 2.2 presents an AID tree diagram, with mean error in over-allocated points and mean error in under-allocated points presented for every split.

The first split was a forced split on dependency status. It was thought best to treat the two dependency status groups separately. For each group, the most decisive split was on SEI, where error was lowest for dependents with SEI not over 100 and for independents with SEI = 0. This pattern is similar to that found in the VEAPS analysis.

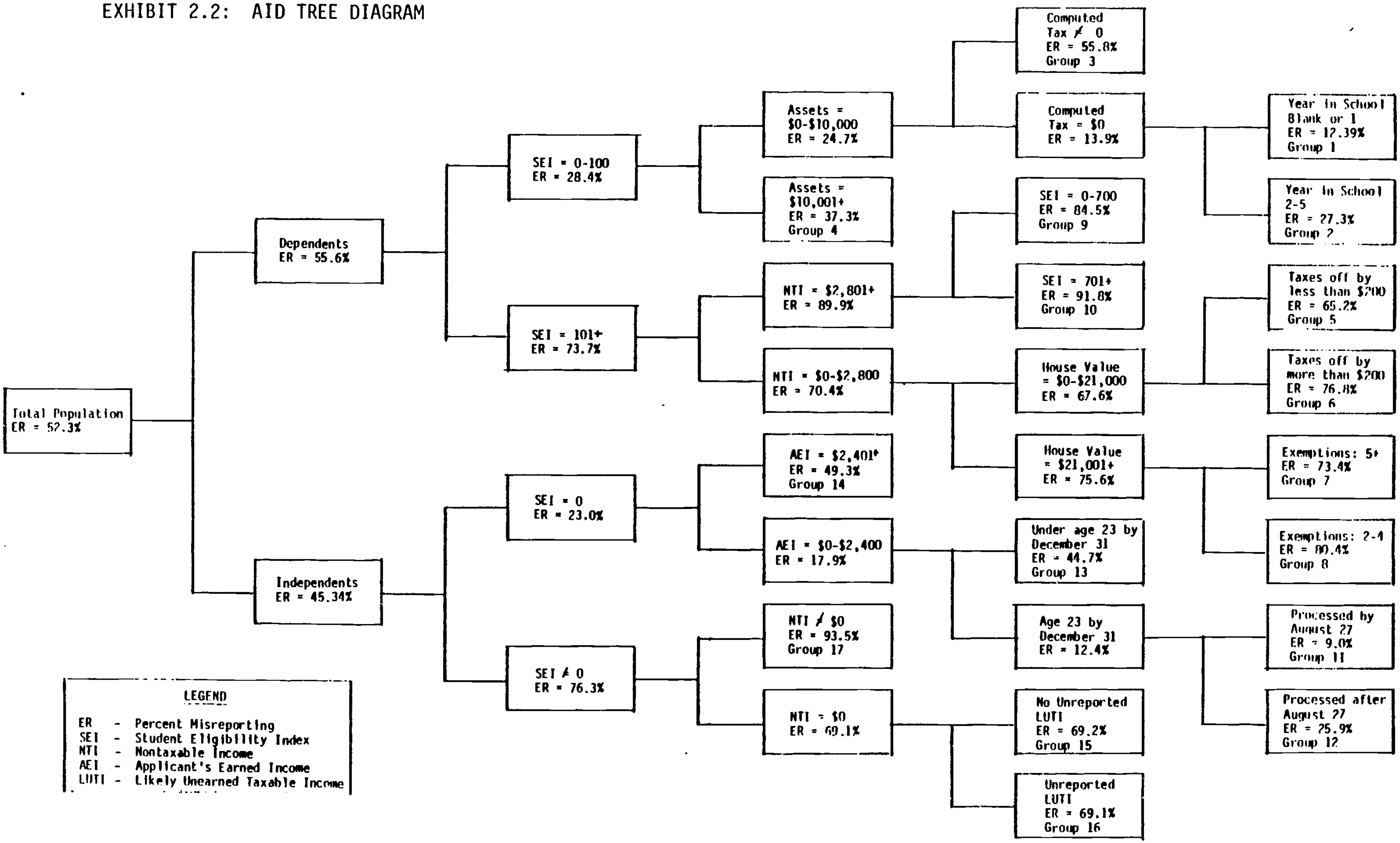
EXHIBIT 2.1: DEFINITION OF THE SEVENTEEN GROUPS RESULTING FROM AIO ANALYSIS

Group	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5	%	%Error
1	Dependent	SEI = 0-100	Assets = $0-$10,000	Comp. tax = 0	YIS = 1 or bl.	8.8	12.3
2	Dependent	SEI = 0-100	Assets = $0-$10,000	Comp. tax = 0	YIS = 2-5	7.7	27.3
3	Dependent	SEI = 0-100	Assets = $0-$10,000	Comp. tax ≠ 0	-----------	2.9	55.8
4	Dependent	SEI = 0-100	Assets = $10,000+	-------------	-----------	7.7	37.3
5	Dependent	SEI = 101+	NTI = $0-$2,800	H.V. = $0-$21,000	Taxes off by under $200	17.3	65.2
6	Dependent	SEI = 101+	NTI = $0-$2,800	H.V. = $0-$21,000	Taxes off by $200	4.3	76.8
7	Dependent	SEI = 101+	NTI = $0-$2,800	H.V. = $21,000+	Over 4 exemptions	7.2	73.4
8	Dependent	SEI = 101+	NTI = $0-$2,800	H.V. = $21,000+	Under 5 exemptions	4.7	80.4
9	Dependent	SEI = 101+	NTI = $2,801+	SEI = 0-700	-----------	3.2	84.5
10	Dependent	SEI = 101+	NTI = $2,801+	SEI = 701+	-----------	4.0	91.8
11	Independent	SEI = 0	AEI = $0-$2,400	Age 23 by Dec. 31	Processed by Aug. 27	10.1	9.0
12	Independent	SEI = 0	AEI = $0-$2,400	Age 23 by Dec. 31	Processed after Aug. 27	2.8	25.9
13	Independent	SEI = 0	AEI = $0-$2,400	Under 23 Dec. 31	-----------	2.7	44.7
14	Independent	SEI = 0	AEI = $2,401+	-------------	-----------	3.1	49.3
15	Independent	SEI ≠ 0	NTI = 0	No Unreported LUTI	-----------	6.2	69.2
16	Independent	SEI ≠ 0	NTI = 0	Unreported LUTI	-----------	3.5	69.1
17	Independent	SEI ≠ 0	NTI ≠ 0	-------------	-----------	3.8	93.5

SEI = Student's Eligibility Index
NTI = Nontaxable income
AEI = Applicant's earned income

YIS = Year in school
H.V. = House value
LUTI = Likely unearned taxable income

EXHIBIT 2.2: AID TREE DIAGRAM


Total Population ER = 52.3%
Dependents ER = 55.6%
Independents ER = 45.34%
SEI = 0-100 ER = 28.4%
SEI = 101+ ER = 73.7%
SEI = 0 ER = 23.0%
SEI ≠ 0 ER = 76.3%
Assets = $0-$10,000 ER = 24.7%
Assets = $10,001+ ER = 37.3% Group 4
NTI = $2,801+ ER = 89.9%
NTI = $0-$2,800 ER = 70.4%
AEI = $2,401+ ER = 49.3% Group 14
AEI = $0-$2,400 ER = 17.9%
NTI ≠ $0 ER = 93.5% Group 17
NTI = $0 ER = 69.1%
Computed Tax ≠ 0 ER = 55.8% Group 3
Computed Tax = $0 ER = 13.9%
SEI = 0-700 ER = 84.5% Group 9
SEI = 701+ ER = 91.8% Group 10
House Value = $0-$21,000 ER = 67.6%
House Value = $21,001+ ER = 75.6%
Under age 23 by December 31 ER = 44.7% Group 13
Age 23 by December 31 ER = 12.4%
No Unreported LUTI ER = 69.2% Group 15
Unreported LUTI ER = 69.1% Group 16
Year in School Blank or 1 ER = 12.39% Group 1
Year in School 2-5 ER = 27.3% Group 2
Taxes off by less than $200 ER = 65.2% Group 5
Taxes off by more than $200 ER = 76.8% Group 6
Exemptions: 5+ ER = 73.4% Group 7
Exemptions: 2-4 ER = 80.4% Group 8
Processed by August 27 ER = 9.0% Group 11
Processed after August 27 ER = 25.9% Group 12
LEGEND
ER - Percent Misreporting
SEI - Student Eligibility Index
NTI - Nontaxable Income
AEI - Applicant's Earned Income
LUTI - Likely Unearned Taxable Income

A second pattern which was consistent across both dependency status groups was that for the higher SEI groups nontaxable income (NTI), and a good predictor of error. Applicants high on both SEI and NTI had the highest error rates, but these were due mostly to errors to the applicant's disadvantage.

Unlike the VEAPS report, MDE source does not appear in the model. This is probably because the VEAPS report concentrated on applicants and whether the applicant re-entered the system or not became part of the dependent variable. The quality control study concentrated on recipients, and MDE did not predict as well for these as other variables did.

Three variables not used directly in the VEAPS report appear in the final model. One was taxes computed from the income, itemized deductions, and exemptions reported in the SER, using only schedules X, Y or Z (which can be programmed more easily and provide a close approximation). The second was discrepancies between reported taxes and computed taxes, a variable closely related to PEC A6 (this variable specifically differentiated among dependents with SEI over 100, NTI not over $2,800 and home value not over $21,000). The third variable was unreported likely unearned taxable income (LUTI). It was reasoned that the presence of savings, investments, a farm or a business should produce income other than earned income. If savings, investments, farm or business are present and AGI equals the sums of portions, then the applicant is said to have unreported LUTI. This variable distinguishes among independent applicants with SEI greater than 0, but no NTI. However, this was a split resulting from the analysis which used misreporting on validation fields as a dichotomous dependent variable. It did not distinguish much on total error, but it did on the nature of the error. Applicants with unreported LUTI (Group 16) were more likely to be over-claimers (48.9%) and to have misreported on fields subject to validation, while those without unreported LUTI were more likely to be under-claimers (33.5%), though an average proportion of over-claimers (27%) was also found in this group.

Age served to define a split, and the younger group (among independents with SEI equal to zero, and earned income not over $2,400) had a high proportion of applicants who misreported their dependency status (41.0%).

The one process date that made a difference was August 27 (process dates which could produce splits were set 15 days apart) corresponding roughly to the beginning of the fall semester. The split was, however, not among the groups with highest error rate.

2.3 Misreporting Patterns of the Groups

The error rates (percentage of applicants misreporting by at least 50 points in either direction or misreporting dependency status) ranged from nine percent for Group 11 to 93.5 percent for Group 17 (see Table 2.3). The groups with the highest percentage of misreporters (Groups 10 and 17) were predominantly composed of under-claimers (67.4% and 60.7% respectively). If one counts model changers as likely over-claimers, three groups (3, 6 and 16) had a majority of its members misreporting to their advantage. These three groups were, in fact, identified through the new tax related variables not used in the VEAPS report. These groups, however, are not the three groups with the highest net overpayments, mean overpayments or mean under-allocated SEI points. Groups 8 and 14, for example, have a lower percentage of applicants misreporting to their advantage, but those applicants that misreport to their advantage do so by larger amounts.

Group 13 was characterized by a large proportion (41%) of applicants who misreported their dependency status. This is one of the few instances where characteristics of applicants likely to be misreporting their dependency status have been identified.

The average discrepancies in SEI between the SER and the verified values ranged from 20.48 for Group 1 to 581.14 for Group 10. Table 2.4 presents various measures of error for each of the seventeen groups. It may be noted that there are greater differences between the groups in SEI

TABLE 2.3: MISREPORTING PATTERNS OF THE SEVENTEEN GROUPS

Group	% of Total	Exact Reporters	Over-Claimers	Under-Claimers	Model Changers
1	8.8	87.7	8.3	2.3	1.7
2	7.7	72.7	17.7	4.8	4.8
3	2.9	44.2	54.0	1.7	0.0
4	7.7	62.7	33.2	4.2	0.0
5	17.3	34.8	33.6	31.3	0.2
6	4.3	23.2	57.6	19.2	0.0
7	7.2	26.6	43.3	30.0	0.0
8	4.7	19.6	46.2	34.3	0.0
9	3.2	15.5	34.4	50.0	0.0
10	4.0	8.2	20.4	67.4	4.0
11	10.1	91.0	6.1	0.0	2.9
12	2.8	74.1	24.0	0.0	1.9
13	2.7	55.3	3.7	0.0	41.0
14	3.1	50.7	32.7	0.0	16.7
15	6.2	30.8	27.0	33.5	8.7
16	3.5	30.9	48.9	13.1	7.1
17	3.8	6.5	22.5	60.7	10.3
Total	100	47.7	28.4	20.1	3.9

TABLE 2.4: ERROR MEASURES FOR THE SEVENTEEN GROUPS[1]

Group	SEI Points Misallocated	SEI Points Under-allocated	SEI Points Over-allocated	Net SEI Mean Difference	SEI Points Misallocated From Validation Fields
1	20	17	3	14	14
2	78	72	5	67	57
3	229	226	3	223	190
4	206	200	5	195	140
5	203	115	88	28	123
6	298	248	50	198	249
7	270	154	116	38	143
8	424	272	153	119	301
9	356	203	153	50	270
10	581	85	496	-411	470
11	65	65	0	65	54
12	191	191	0	191	186
13	54	54	0	54	51
14	344	344	0	344	300
15	281	120	161	- 41	184
16	276	242	35	207	242
17	546	154	391	-237	451
Total	224	140	85	55	164

[1] Model changers excluded from table.

points over-allocated than in SEI points under-allocated. This is partially artifactual, since it is impossible for over-allocated points to exist for an applicant whose SEI is equal to zero. Had ineligibles been included and assigned an SEI eaual to 1,601, groups with points under-allocated equal to zero might have emerged.

In spite of the artifactual nature of some of these results, it remains the case that patterns of under-claiming remain clearer than patterns of over-claiming. Persons purposefully trying to over-claim may use a variety of techniques and thus fall into various groups; a substantial proportion of under-claimers seem to have counted nontaxable income which was not verified. These two factors combined to present a picture of systematic under-claiming and unsystematic over-claiming. The degree to which this picture is accurate or not depends on the degree to which varification of nontaxable income was accurate in the first place. The next chapter will present a group-by-group description of the seventeen groups resulting from this error-prone model.

3

DESCRIPTION AND INTERPRETATION OF THE GROUPS

This chapter describes each group in terms of its misreporting patterns and its SER profile. An attempt will be made to interpret and identify possible reasons for misreporting and to suggest corrective action. Each description should be read in conjunction with the various tables and exhibits.

At this point a few terms should be reviewed. Following the terminology in the VEAPS report, the term over-claimer refers to an applicant whose verified SEI was at least 50 points higher than the SEI derived from the SER. Under-claimer refers to an applicant whose verified SEI is at least 50 points lower than the SEI derived from the SER. The term model changer refers to applicants whose verified dependency status differs from the SER dependancy status. Misreporter refers to over-claimers, under-claimers and model changers. Error rate refers to the percentage of misreporters in a group. Mean error refers to average misallocated SEI points, or the sum of over-allocated SEI points (resulting in underpayments) and under-allocated SEI points (resulting in over-payments). Net cost refers to the difference between what the government should have paid and what it actually paid. Misallocated dollars refers to the sum of underpayments and overpayments. Calculations of discrepancies in SEI points or in cost or misallocated dollars exclude model changers.

Group 1: dependents with SEI not over 100, assets not over $10,000, computed tax equal to 0 and in their first year in school (or leaving year in school blank). This group had the lowest error rate (12.3%) of any dependent group and the lowest mean error of any group (17 misallocated SEI points). No remedial action is required for this group.

Group 2: dependents with SEI not over 100, assets not over $10,000, computed tax equal to 0 and in at least their second year in school. This group has a somewhat higher error rate (27.3%) than the previous one, but remains among the lowest error rates of any group. One area where there appears to be some systematic misreporting is in applicant's resources. One may conjecture that upperclassmen are not reporting all their summer and part-time earnings, a resource which first-year applicants are less likely to have. The new 1980-1981 application form may contribute to reducing this problem.

Group 3: dependents with SEI not over 100, assets not over $10,000 and computed tax not equal to 0. This group has a much higher error rate (55.8%) that the previous two and the second highest proportion of over-claimers of any group (54.0%). A net cost of $97 over-payment per applicant makes this a high misreporting group. This group reports low NTI, and discrepancies in NTI account for much of the error attributed to this group. This group also has the lowest proportion of married parents (32.8%) of any dependent group and mother's portion is almost twice as large on the average as father's portion (mother's portion is also a high source of error).

The major errors of Group 3 applicants seem to be errors of omission. Willingness to disclose figures to the QC interviewer suggests that the right questions might extract some of this information. A detailed sheet listing possible NTI sources and requiring a signed statement certifying their absence might eliminate some of the error. Validation up to now has not paid sufficient attention to NTI (other than social security), and even though starting in 1980, this field will be validated. It is always more difficult to verify the absence of a source of income than its presence. Group 3 requires some remedial action, be it validation or verification by mail.

Group 4: dependents with SEI not over 100 and assets over $10,000. This group has a relatively low error rate (37.3%), but many of the errors seem to be large and often due to fields other than those financial aid officers are required to validate in the absence of a comment. Students in this group have parents whose low income may well be a technicality (e.g., business losses offset gains for a low AGI). An error, whether through carelessness or intent to defraud, is likely to have a large impact. Due to the low SEI of this group, this impact is almost always in the applicant's favor. The low error rate makes validation impractical, but the high impact of the errors made by the minority of misreporters within this group makes some remedial action desirable. The use of more detailed forms for applicants in groups such as this (similar to the IRS long and short forms) might, in the long run, help reduce misreporting for applicants in this group.

Group 5: dependents with SEI over 100, NTI not over $2,800, home value not over $21,000 and computed taxes within $200 of reported taxes. The moderately high error rate in this group (65.2%) is close to evenly divided between over-claimers (33.6%) and under-claimers (31.3%). This is the largest unsplit group and it seems to lack any systematic pattern to its misreporting. Systematic remediation directed at this group seems impractical, beyond those changes directed at the total sample. Further attempts at finding new variables which can split this group should be carried out, but validation would be inefficient unless a decision to validate most applicants is made.

Group 6: dependents with SEI over 100, NTI not over $2,800, home value not over $21,000 and computed taxes and reported taxes at least $200 apart. This group not only has a high error rate (76.8%) but also the highest percentage of over-claimers of any group (57.6%). Almost all of these applicants misreport in fields subject to validation. The net cost to the government of applicants in this group is $181 per applicant, the highest average net cost of any group. These results confirm the findings of the VEAPS report which indicated that discrepancy between reported and computed taxes is among the best predictors of misreporting. Validation seems to be indicated for Group 6.

Group 7: dependents with SEI over 100, NTI not under $2,800, home value over $21,000 and five or more exemptions. Though the error rate for this group is almost as high (73.4%) as that of the previous group, Group 7 has a lower percentage of over-claimers (43.3%) and a higher percentage of under-claimers (30.3%). Much of the error attributable to this group is due to non-validation fields, primarily to home value. Since it is difficult for applicants to determine home value under its present definition, the best procedure for reducing error among Group 7 applicants would seem to be a change in the way information concerning home value is requested. Possibly some formula combining original cost of the home, date of purchase, and location should be used to determine home value in place of the request for a figure which the applicant is unlikely to be able to provide. Since the mean error due to fields which are routinely validated is below the sample mean, the standard validation procedure would not be effective for this group.

Group 8: dependents with SEI over 100, NTI not over $2,800, home value over $21,000 and not more than four exemptions. This group has a relatively high error rate (80.4%) with over-claimers (46.2%) somewhat outnumbering under-claimers (34.3%). It has the highest mean error of any group in which over-claimers outnumber under-claimers. Group 8 errors are spread across various fields including household size, AGI, NTI, home value, savings, applicant's resources and (for under-claimers primarily) home debt and taxes paid. It has the third highest mean net cost per applicant ($136) and should be validated in the absence of other remedial action.

Group 9: dependents with SEI over 100 but not over 700 and NTI over $2,800. This group has the third highest error rate (84.5%) among the seventeen groups, but under-claimers (50%) outnumber over-claimers (34.4%). However, over-claimers in this group have higher errors than under-claimers, so that the net cost of misreporting is $83 per student against the government. NTI is, as one would expect, one of the largest sources of error for members of this group. The group's mean NTI is twice its mean AGI, and since NTI is not reported in tax returns, omissions are more likely to take place. However, errors in AGI are also present (though of lower magnitude) as are errors in home value, primarily for over-claimers. It should be noticed that 35.5 percent of these applicants have widowed or deceased parents, a factor which probably contributes to their misreporting, but while many make small mistakes to their disadvantage, a small number make larger mistakes to their advantage. Unless a larger percentage of applicants than are presently selected for validation are to be chosen, this group should not be, but alternate remedial action such as requiring a more precise breakdown of NTI might be considered.

Group 10: dependents, with SEI over 700 and NTI over $2,800. The presence of high NTI for a high SEI applicant seems to indicate error to the applicant's disadvantage. With a high error rate (91.8%) consisting primarily of under-claimers (67.4% of the total), the average student belonging to this group fails to receive $228 to which he would have been entitled had the information in his SER been correct. The mean error for this group is 581 misallocated SEI points. Both NTI and AGI, as well as home value, contribute to the error, but NTI makes by far the largest contribution. A more detailed examination of the nature of these errors is beyond the scope of this study. Since over half of all applicants reporting NTI are discrepant from their verified NTI values by more than $1,000, this is a field that probably needs tighter edits for the protection of the applicant.

Group 11: independents with SEI equal to 0, applicant's earned income not over $2,400, who have reached their twenty-third birthday by January 1 of the academic year and were processed by August 27. This group has the lowest error rate (9%) of any of the seventeen groups. No corrective action is needed, since verification uncovered few discrepancies for applicants in this group.

Group 12: independents with SEI equal to 0, applicants earned income over $2,400, who have reached their twenty-third birthday by January 1 of the academic year and were processed after August 27. This group has a higher proportion of misreporters (25.9%) than the previous group. While the rate is low, the minority which does misreport, does so by a substantial amount, resulting in a net cost of $101 per applicant to the government. Most of this misreporting is related to AGI and NTI. It should be noted that late applicants were not included in the VEAPS error-prone model study, and that this is an indication that applicants who file later may differ from those who filed earlier. The error rate is low, but the errors are serious enough that some remedial action might be desirable. On the other hand, with the academic year starting or having started, an excess of tight edits might hinder the plans of the majority which reported accurately. Some sort of immediate automatic validation of NTI and AGI only might be considered as one possibility.

Group 13: independents with SEI equal to 0, applicant's earned income not over $2,400 and who had not reached their twenty-third birthday by January 1 of the academic year. This group had a low proportion of over-claimers (3.7%) but the highest proportion of model-changes (41.0%) of any of the seventeen groups. With only 8.4 percent of applicants who claimed to be independents, this group has 34.9 percent of those independents who were later found to be dependents. Validation of dependency status only is recommended for this group.

Group 14: independents with SEI equal to 0 and applicant's earned income over $2,400. While this group has close to an average error rate (49.3%), its mean error is very high (344 misallocated SEI points) and totally attributable to over-claimers. Its net cost ($168 overpayment per applicant) is second highest among the seventeen groups, and highest among the seven groups composed entirely of independents. In addition to AGI and NTI, household size is an important source of error. This group has the largest proportion (40.7%) of married applicants of any independent group. However, the mean household size is 3.3 which may suggest the presence of errors through counting persons not legally a part of the household. It should be also noted that 16.7 percent of the members of this group are model changers. All of this points to a predominance of complex family situations which need to be sorted out prior to an accurate determination of the award the applicant is entitled to.

Group 15: independents with SEI greater than 0, no NTI and no unreported likely unearned taxable income (LUTI). The concept of LUTI was discussed earlier, and will be elaborated upon in the discussion of the next group. Group 15 has a moderately high error rate (69.2%) with under-claimers slightly outnumbering over-claimers. The net SEI difference is slightly to the applicant's disadvantage, but the net cost is slightly to the applicant's advantage. Counting the 8.7 percent model changes found in this group, it is likely that the net effect of the misreporting has been to the applicant's advantage. Household size and AGI have been sources of error, as have been other critical fields not ordinarily validated. While validation of this group would reduce the amount of misallocated dollars, it would not save the taxpayers' money in the long run. Tighter edits are suggested.

Group 16: independents with SEI greater than 0, no NTI and presence of unreported likely unearned taxable income (LUTI). The presence of savings, investments, a business or a farm should indicate a component of AGI other than earned income. If AGI equalled earned income in the presence of any such assets, one might question where interest, dividends, or profits might have gone. This variable did not separate Groups 15 and 16 on total error, but it did so on error due to validation fields. In addition, it produced groups which differ in the direction of their misreporting patterns. Group 16 has an error rate (69.1%) almost identical to that of Group 15, but its percentage of over-claimers (48.9%) is more than three and a half times its percentage of under-claimers (13.1%) with 7.1 percent model changers. Net cost is $103 per member of this group. AGI, NTI and taxes paid are the largest sources of error for this group. This group should probably be validated.

Group 17: independents with SEI greater than 0 and NTI greater than 0. This group is the independent counterpart of Group 10, and confirms the assertion that high SEI and high NTI combined are an indication of an under-claimer. This group has the highest error rate (93.5%), but the largest proportion is due to under-claimers (60.7% of the total). Over-claimers in this group tend to make large errors, so the net cost is

$46 per applicant to the applicant's disadvantage. This group makes more errors in reporting Veteran's Benefits (particularly number of months) than any other group. While AGI is also a large source of error, NTI is the largest source, and thus it is possible that edits to NTI might reduce much of the error.

The next chapter will summarize these descriptions and offer recommendations both for implementation of the results and for further research.

4

IMPLICATIONS, ADDITIONAL RESULTS, AND SUMMARY

This chapter will discuss the explanatory power of the model, present a selection strategy devised from the model, discuss edits for the under-claimers as identified by the model, and discuss attempts at devising alternate models using discriminant analysis. Implications for further research and recommendations emerging from this investigation will be discussed. Finally, the results will be compared with those of the VEAPS analysis and an assessment of the importance of the study will be discussed.

4.1 Explanatory Power of the Model

Several measures of explanatory power can be cited to indicate the power of an error-prone model. The most useful ones are the eta statistic in an analysis of variance (equivalent to R in a multiple linear regression) and the total discriminatory power statistic (Tatsuoka, 1970) in a discriminant analysis.

A one-way analysis of variance has been conducted with the groups resulting from the model as the independent variable and various measures of error as the dependent variables. These measures underestimate the power of the model, since model changers had to be excluded from the analysis because their verified SEI and award was not obtained at the data collection stage.

The following variables were used as dependent variables: net SEI differences, net award differences, SEI points misallocated, SEI points misallocated due to error in fields subject to validation, SEI points under-allocated, and SEI points over-allocated. The groups were most effective in predicting SEI points over-allocated (i.e., assigned to underclaimers) with eta = .52, while SEI points under-allocated was predicted least by the model (eta = .26). Total misallocated SEI points, the dependent variable the model was designed to predict most was predicted to an intermediate degree (eta = .41). Error detectable through validation (eta = .37), net SEI difference (eta = .34) and net award difference (eta = .26) produced intermediate results. These findings confirm the assertion presented in a previous section, that the model is more effective in identifying under-claimers than in identifying over-claimers.

A discriminant analysis was also conducted with dummy variables representing the groups as predictors and misreporting pattern (exact reporter, over-claimer by 50 or more SEI points, under-claimer by 50 or more SEI points and dependency status misreporter) as the criterion variable. The total discriminatory power statistic was .52.

4.2 Selection Strategy Based on Quality Control Study Error-prone Model

Were a strategy for selection for validation to be based exclusively on this error-prone model, the strategy would have to be different depending on whether the major concern was maximizing recovery of overpayments or maximizing detection of over-claimers. For example, Group 3 has a higher percentage of over-claimers than Group 8, but the latter has the higher average overpayment and the higher average under-allocated SEI points. Given a choice between validating Group 3 or Group 8, consideration for deterrence would favor the group with the higher proportion of over-claimers (i.e., Group 3), while consideration for detecting the maximum amount in potential overpayment dollars would lead one to select Group 8.

Selection of Groups 6, 8, and 14 would result in selection of 12.1 percent of all applicants, accounting for 27.6 percent of all overpayment dollars (not counting those due to misreporting dependency status). This combination would maximize identification of potential overpayment dollars (see Table 4.1). Additional validation of just dependency status for Group 13 would further reduce overpayments, though the amount cannot be determined.

Selection of Groups 3, 6, and 16, on the other hand, would maximize the percentage of overclaimers identified. Using this combination selection of 10.7 percent of all applicants would result in identification of 20.4 percent of overclaimers.

4.3 The Underclaimers and Nontaxable Income

Two groups (Group 10 and Group 17) include 7.8 percent of all applicants, but account for 25.0 percent of all under-claimers, and 38.5 percent of underpayments in dollars. These two groups are characterized by the combination of high SEI and high nontaxable income (NTI). While the percentage of over-claimers in these groups is relatively low (20.5% and 22.5%, respectively), over-claimers in these groups misreport by large amounts.

Both the large proportion of under-claimers and the seriousness of what overpayments exist in these groups might be reduced through either of two measures. First, applicants falling in these groups could receive comments suggesting they might be entitled to more funds than their present SEI seems to indicate and that they should bring their records to their financial aid administrator and seek assistance. Second, applicants might be required to itemize non-taxable income. This second requirement would not only reduce carelessness and unwarranted estimation, but would permit identification of inappropriately reported NTI (such as loans, money obtained from a sale where no profit was made, double-counting taxable income, etc.).

TABLE 4.1 AWARD ERRORS IN DOLLARS FOR EACH OF THE SEVENTEEN GROUPS

GROUPS	APPLICANT ERRORS				INSTITUTIONAL ERRORS	
	OVERPAYMENTS		UNDERPAYMENTS		DOLLARS MISALLOCATED	
	TOTAL	PER APPLICANT	TOTAL	PER APPLICANT	TOTAL	PER APPLICANT
1	586,056	5	44,736	0	5,918,546	54
2	1,205,931	13	107,474	1	11,348,354	121
3	3,663,206	99	42,549	1	738,504	20
4	8,669,608	88	154,102	2	3,937,491	40
5	21,755,906	99	10,006,019	46	9,309,491	42
6	11,176,918	204	1,260,212	23	1,296,177	24
7	12,435,102	135	7,572,867	83	4,309,305	47
8	14,979,248	253	7,019,553	118	1,198,887	20
9	5,560,661	139	2,250,303	56	884,478	21
10	4,490,672	92	15,303,916	315	3,255,589	67
11	4,461,521	36	0	0	7,994,282	64
12	3,421,008	99	0	0	2,778,028	81
13	740,734	37	0	0	961,483	48
14	5,569,983	168	0	0	1,431,923	43
15	6,827,267	94	4,845,526	67	3,226,402	45
16	5,016,695	121	884,013	21	3,064,722	74
17	4,152,531	95	6,112,614	140	3,083,558	70
TOTAL	114,713,047	94	55,603,883	45	64,697,228	53

Total of 1,223,391 applicants represented. Model changers excluded.

4.4 Alternate Models Using Discriminant Analysis

Several attempts at using discriminant analysis were made. These attempts, however, must be questioned since the number of va·iables selected was usually large, and even where a smaller number of predictor variables was selected, these had been chosen after determining which were the most promising predictors by examining the AID analysis as well as bivariate tables. Thus, the assumptions needed to establish any inferences from the discriminant analyses were absent. The classification formulas resulting from the discriminant analyses apply to the sample, but cannot be counted on when replicated. Had the sample been divided prior to examination of the data and a discriminant analysis conducted on half the sample, the results could have been cross-validated on the other half. Since the AID analysis, was felt to be more promising, this procedure was not followed.

Upon examination of the results of the various discriminant analyses, it became evident that they constituted no improvement on the AID analysis. Furthermore, as with AID, the discriminant analyses were better able to identify under-claimers than over-claimers. Total discriminatory power for various discriminant analysis runs using SER fields was lower than that of the AID groups, except for instances where the groups were incorporated into the analysis as dummy variables.

4.5 Recommendations for Further Research

The process of investigation of characteristics of error-prone applicants must be a continuous one, since these characteristics may change and every study can provide new insights into variables that can serve as predictors. The following are recommendations which should be incorporated in further statistical analysis of the characteristics of error-prone applicants:

- If, at all possible, ineligibles should be included in the analysis.
- Discrepancy between taxes reported and taxes computed should be used as a predictor variable in further studies.

- Variables suggesting inconsistencies in the tax return should be further investigated. Absence of income other than earned income for persons indicating savings, investments, farm or business proved a useful predictor.
- A separate in-depth study of dependency status misreporters is indicated, since they have different characteristics from other misreporters.
- Greater efforts should be made to interview parents of independents, particularly if dependency status is found in error.
- The one group consisting entirely of upperclassmen exhibited the largest average institutional error (see Table 4.1) of any group, suggesting that institutions are more prone to error with respect to students who have been previously enrolled. This fact should be studied further.

4.6 Recommendations for Program Improvement

Certain facts emerged out of this analysis which can lead to possible actions other than validation. Some of these facts suggest edits, while others suggest changes in the application process or in directives to institutions:

- An edit should be issued to applicants who report savings, investments, farm or business, but whose adjusted gross income equals their earned income, since it is likely that interest, dividends or profits were not reported in this case.
- As discussed in a previous section, edits should be sent to members of Groups 10 and 17 suggesting they seek assistance.
- Nontaxable income should probably be itemized in the application form.
- The one split produced by process date suggests that applicants who apply late are somewhat more error-prone than those who apply early. The distinction was made for groups with low error rate, but some of the misreporters in the late group did so by a large amount. This not only suggests that the VEAPS data may underestimate misreporting (since it used the end of August as a cut-off date), but it also suggests that the validation requirement should be carried out for late applicants at least to the same extent as for early applicants.
- One recommendation that would have emerged from this study had it not been already implemented is a more precise itemizing of applicant's income and resources for dependent applicants. There seems to have been a tendency for upperclassmen to misreport resources, and one surmises that the reason upperclassmen misreport this field more is that they have had

greater opportunity to save from summer and part-time work. Requiring details of applicant's income should reduce the magnitude of this misreporting and the 1980-1981 figures for Group 2 may confirm this.

4.7 Comparison with the VEAPS Report EPM

The error-prone model emerging from this study used different variables and different methods than the error-prome model emerging from the validation study. Methodological and data base differences included:

- This report used 1978-1979 data while the VEAPS report used 1979-1980 data.
- The VEAPS report defined errors as a discrepancy between selection and latest or latest payment transaction for validation applicants; this report defined error as a discripancy between the SER and values obtained through a verification interview.
- This report used AID3, with SEI discrepancy as its major dependent variable; the VEAPS report used THAID, with type of misreporting (including failure to re-enter) as its dependent variable.
- The most effective variable in the VEAPS report, estimation of taxes, was not used in this report since it was not available for 1978-1979 applicants.
- No replication sample was available for this report.
- The sample size for this report consisted of approximately eleven percent of the size of the working or of the replication sample in the VEAPS report.
- Several new variables including taxes computed from deductions, marital status exemptions and income information in the SER were added as predictors.

Due to these differences it is not surprising that very different results were found. This study was more effective in identifying under-claimers than the VEAPS model, but less effective in identifying over-claimers. MDE source played a greater role in the VEAPS report because it was a particularly effective predictor of failure to re-enter the system.

One major difference is due to the fact that every critical field was verified for this report, while financial aid administrators were not required to verify every field as part of the validation process.

One definite possibility which could account for some differences is the presence of spurious discrepancies due to the data collection methodology. It is possible that person who estimated at the time the application was completed merely provided a different estimate in leu of documentation at the time of the interview. The description of the data collection procedures leaves many questions unanswered, but the large role of nontaxable income in identifying under-claimers is the most serious. This may be because many applicants did not document nontaxable income while reporting a different value in the interview than on the original application.

On the other hand, there were several important similarities between the two models. First and foremost is the fact that, except for model changers, low SEI applicants - particularly those with low income and few or no assets - are seldom found to be misreporting. The question of whether they really are not misreporting or whether they are more consistent in their misreporting, and are these less likely to be caught, remains unanswered.

A second similarity is the role of taxes in identifying misreporters. In the VEAPS report estimating taxes or reporting more than 15 percent of one's adjusted gross income (AGI) as taxes when AGI is under $25,000 was associated with misreporting. In the present investigation several tax related variables provided additional predictive effectiveness. Further examination of the tax-income interaction across studies seems appropriate, and may be attempted in the Task 2 IRS match error-prone model.

4.8 Overall Assessment of the Study

The error-prone model analysis of the quality control data has been productive in several regards:

- It has provided information useful in identifying potential under-claimers.
- It has provided an alternate, if less effective model than the VEAPS report for the identification of over-claimers.

- It has provided the first potentially effective means of identifying dependency status misreporters.
- It has provided information supporting specific changes in the application procedure.
- It has provided information which can be used in further studies including the Task 2 IRS match.
- It has provided evidence for the effectiveness of AID as an error-prome modeling technique.

APPENDIX A

GENERAL METHODOLOGY

Error-Prone Modeling Techniques

The development of adequate error-prone models has become an important concern for various agencies. The need to determine which cases are most likely to be misreporting so that correction action-procedures can be instituted is likely to be present for any program engaged in the disbursement of public funds on the basis of stated need and qualifying conditions.

Three major approaches to error-prone modeling have been used by various state and federal agencies. Each has its advantages and disadvantages and each is best suited to different kinds of situations.

The first method, used by the Welfare Departments of South Carolina and the District of Columbia, uses discriminant analysis to obtain a formula which assigns a score to each case. The higher the score, the more likely it would be that the applicant is misreporting. Thus if the agency wanted to select applicants most likely to misreport, it would simply select those applicants to whom the formula assigned the highest scores.

The major drawback of this method is that it ordinarily assumes that a variable will affect all applicants in the same way. If, for example, it turns out that estimating taxes is an indicator of misreporting for dependent but not for independent applicants, discriminant analysis will fail to take this into account. Thus it could easily fail to detect some important combinations of variables which could predict error-proneness. In addition, discriminant analysis would not point to specific areas where an applicant is likely to be misreporting. Since each applicant receives a single score one cannot distinguish those who misreported AGI from those misreporting rates paid. Of course, separate analyses could be conducted to predict misreporting for each specific field, but this method would lack parsimony and would be difficult to interpret. Discriminant analysis, however, may be used in conjunction with other techniques at which point its purpose is not to create a model, but to test one and determine its effectiveness.

The second approach, used by the State of New Hampshire to identify error-prone cases in its Medicaid program, attempts to define a single group most likely to exhibit a high degree of misreporting. Depending on the size of the group, every member or a certain percentage of this group (and this group only) would be validated. Where only a small proportion of all the cases can be validated, and the principal objective is to maximize the savings in actual disbursements from the cases actually validated, this method can be very effective. On the other hand, this approach is likely to overlook groups whose error rate might approach that of the selected group, and which might require a different type of corrective measure. The BEOG program, with the use of edits in addition to validation, a high drop-out rate among applicants selected for validation, the possibility of different treatments for different kinds of misreporters and the dual concern for deterrence as well as savings in disbursements to validated applicants, requires a different approach.

The third approach has been used by the State of West Virginia Aid to Families with Dependent Children program and by the Supplemental Security Income program of the Social Security Administration, and was used in the development of the error-prone model in the VEAPS report. It essentially divides the applicant population into mutually exclusive groups which differ as much as possible from each other in either the mean discrepancy in expected disbursement or in the rate or type of misreporting. This method has the potential to describe each group separately in terms of type of error, and thus to prescribe different types of corrective action for each. It has the further advantages of taking into account effects which apply to only part of the population, and of producing results which can be expressed in simple terms. This method, sometimes called classification analysis, sequential structure search or automatic interaction detection (AID) is the one which will be used in this study.

Overview of the Sequential Approach

The term sequential structure search is the more generic term for a conceptual method of exploratory analysis designed to discover nonlinear combinations among many variables which best predict a single dependent variable. The term Automatic Interaction Detector (AID) is at times used synonymously with sequential structure search, but is often used more precisely to describe the implementation of this concept by the Institute of Social Research (ISR) at the University of Michigan. ISR has developed two programs which will conduct this type of analysis: AID3 which accepts continuous dependent variables, and THAID which accepts categorical dependent variables. This software was used by the Supplemental Security Income program in their development of error-prone profiles and by Applied Management Sciences in the development of the VEAPS report error-prone model. In addition, the West Virginia AFDC program used a software package closely related to THAID. In the subsequent discussion, the term AID will refer to the general technique rather than to a specific program or package.

Techniques such as discriminant analysis or multiple linear regression make the assumption that a given effect will apply equally to all members of the population. The result of either of these techniques is an equation which is meant to predict the dependent variable for every member of the population. AID, on the other hand, does not make the assumption that a predictor variable will affect the dependent variable in the same way for all cases. Instead, AID starts by breaking up the sample into two subgroups selecting that split which produces groups that differ from each other on the dependent variable as much as possible. Each subgroup is then split separately, allowing for different predictor variables to split different previously-formed subgroups.

Interaction effects occur when a variable predicts differently for one group than for another. Ordinarily, discriminant analysis and multiple linear regression do not take interaction effects into account. AID specifically identifies groups (using various combinations of variables) which will differ as much as possible on some criterion variable. Thus AID will be able to identify error-prone cases in instances where, for example, low taxes are an indication of error-proneness among high income applicants, but not for low income applicants. Linear models are oblivious to such relationships.

AID accepts one dependent variable, which may be categorical (such as type of applicant) or quantitative (such as discrepancy in expected disbursement, expressed in dollars). Predictor variables can be monotonic (where the sample or any subgroup can only be subdivided into high and low groups based on some cut-off point) or free (used for categorical, as opposed to quantitative, predictors where any combination of values can be used to split the groups). In either case predictor variables must be coded in terms of a small number of possible values.

Out of the many possible splits defined by predictors, AID selects the one which will divide the sample into two groups as different from each other as possible. The process then continues for each of the two groups into which the original sample was split. When a group becomes very homogeneous, cannot be further split using the predictor variables available, or would yield subgroups under a certain size if it were split, then the process is complete and it becomes one of the groups defined by the model. If an applicant group is both large and heterogeneous, it would be an indication that some additional predictors should be sought and included in the analysis.

Because AID investigates many possible combinations of variables, at times it produces results which are specific to a given sample. Two questions may be asked pertaining to the groups which emerge from an AID analysis: (a) Do the groups have the same characteristics in the population as they appear to have in the sample? and (b) do the groups constitute an optimal classification of the population if one is trying to predict the dependent variable? In order to answer either question one needs to use a second sample randomly drawn from the same population. The first question can be answered by checking whether the subgroups produced by AID from the first sample have similar characteristics in the second sample. The second question requires that a separate analysis be conducted for the second sample. It is quite possible that one would obtain a different solution if variables which are highly interrelated are use (this is similar to the problem of multicollinearity in multiple linear regression). The question of whether a given solution is the best possible, however, is of secondary importance to whether the classification which emerges is effective in predicting error-proneness. Because the sample used in this study is relatively small, these questions cannot be answered. The stability of the model, however, should be verified in subsequent studies.

The AID3 Program

The AID3 program is part of the OSIRIS package, and can also be acquired separately. It is the most popular sequential search program, and is appropriate for continuous and dichotomous dependent variables. AID3 limits predictors to 63, categories in a single predictor to 10 and total number of categories to 400. The definition of "as different from each other as possible" in AID3 is based on a least squares statistic (maximizing the between groups sums of squares). This differs in its effect from THAID (for dichotomous dependent variables which can be analyzed with either program) in that THAID places a greater premium on symmetry than AID3. Thus AID3 can more effectively identify smaller groups during the earlier splits than THAID can.

A minimum number of cases for each resulting group can be specified. In this study, this minimum was set at 50. A reducibility criteria applies to each split, requiring a certain proportion of the variable to be accounted for by the split. The default option of .008 was used in each of the analyses. While it is possible to force a first split, it was found more effective to conduct separate runs for dependents and independents. This loses some information in terms of total effect statistics, but these were unnecessary since they were obtained from SPSS runs.

APPENDIX B

THE FOUR AID ANALYSES

I. Definitions of the Criterion Variables

Your criterion variables were used in four separate AID analyses. They were defined below:

- E_1 equals the absolute value of the SEI obtained from the SER and the SEI obtained using verified information. It is missing for model changers.
- $E_2 = 1$ if E_1 is not under 50 or if the applicant is a model changer, $E_2 = 0$ otherwise.
- $E_3 = E_1\ (a_1+a_2+\ldots+a_k)/(a_1+a_2+\ldots+a_k\ +b_1+b_2+\ldots+b_m)$ where a_j represents the SEI change due to a field subject to validation for the 1980-1981 academic year and b_j represents the SEI change due to a field not subject to validation for the 1980 - 1981 academic year. SEI changes are expressed as positive numbers regardless of their direction. Fields are said to be subject to validation if financial aid administrators are required to validate them for every applicant selected for validation. E_3 is missing for model changes.
- $E_4 = 1$ if E_3 is not under 50 or if the applicant is a model changer. $E_4 = 0$ otherwise.

II. Sources of Splits

Most of the splits were produced by the AID analysis which used E_1 are the dependent variable. The exceptions are noted below.

- The split between Group 1 and Group 2 was produced using E_2 as a dependent variable.
- The split between Group 5 and Group 6 was produced using E_3 as a dependent variable.
- The split between Group 13 and the two preceding groups was produced using both E_2 and E_4 (both of which incorporated model changers) as dependent variables.
- The split between Group 15 and Group 16 was produced using E_4 as a dependent variable.

APPENDIX C

Bibliography

Applied Management Sciences, Inc. Validation, Edits and Application Processing. Phase II and Error-Prone Model Report. Silver Spring, Maryland, 1980.

Macro Systems, Inc. Basic Educational Opportunity Grant Quality Control Study, Volume I. Rockville, Maryland, 1979.

Morgan, J.N., and Messenger, R.C. THAID: A Sequential Analysis Program for the Analysis of Nominal Scale Dependent Variables. Ann Arbor, Michigan: Institute for Social Research, 1973.

Sonquist, J.A.; Baker, E.L., and Morgan, J.N. Searching for Structure: An Approach to Analysis of Substantial Bodies of Micro-data and Documentation for a Computer Program. Ann Arbor, Michigan: Institute for Social Research, 1973.

Tatsuoka, M.M. Discriminant Analysis. Champaign, Illinois: Institute for Personality and Ability Testing, 1970.