IDENTIFYING DISTINCTIVE GROUPS IN A COLLEGE APPLICANT POOL

Robert Lay, John Maguire, and Larry Litten

In a time when most post-secondary educational institutions must distribute limited resources more efficiently, segmentation analysis affords a way to direct planning to yield strategic benefits. The Automatic Interaction Detector is recommended to the institutional research community as an analytical tool for effectively identifying distinctive subgroups within an educational market. In this application, AID is used to segment the Boston College applicant pool according to subgroups' relative probabilities of enrolling. The findings illustrate that AID can make a useful contribution to market research and that the technique has broader applicability—to other student groups and to other educational policy questions.

The utility of a strategy that segments a market into discrete groups for particular treatment has been appreciated for a number of years outside higher education. More recently, academic market researchers have come to recognize the potential benefits of segmentation analysis. (See Litten, 1979, for a review of the literature.) Planners have discovered that through an informed division of potential students, policies and procedures may be tailored to the needs and preferences shared by applicants in each segment. At a time when most post-secondary educational institutions must learn to distribute limited resources more efficiently, segmentation analysis provides a way to guide planning to yield strategic benefits—the sum of which may often surpass the gains from a more costly, yet undifferentiated approach.

A difficulty often associated with segmentation analysis in higher education is the lack of a systematic procedure for defining directly the characteristics that

Robert Lay, Director of Enrollment Management Research, Boston College; John Maguire, Dean of Admissions, Records and Financial Aid, Boston College; Larry Litten, Associate Director, Consortium on Financing Higher Education, Cambridge, Massachusetts.

An earlier version of this paper was presented at the Twenty-First Annual Forum of the Association for Institutional Research, Minneapolis, May 1981.

Research in Higher Education Vol. 16, No. 3, 1982
© Agathon Press, Inc. 0361-0365/82/080195-14$1.50

optimally segment a market. The norm has been to follow more or less educated assumptions regarding which characteristics define the most meaningful segments. The principal statistical techniques used to date—multiple regression, multiple discriminant analysis, and the like—reveal the contributions that variables make to predict behaviors or group membership, but do not efficiently identify specific groups with distinctive attributes.[1] In this article we will discuss an effective analytic technique for segmenting a heterogeneous group of students and examine one application of the technique—segmenting an applicant pool.

THE AUTOMATIC INTERACTION DETECTOR

The Automatic Interaction Detector (AID) was introduced as a data analytic technique by Morgan and Sonquist in 1963. At first AID was applied to social science research questions, but it soon found an application in business as a marketing research tool.[2] The technique's appeal lies in the simplicity of its method and in the ease of its interpretation.

AID searches a data set for distinctive segments using much the same logic as a good researcher might, but with the capacity to evaluate thousands of possible solutions. To begin, the algorithm finds the segmenting variable that produces the maximum separation between two subgroups on a selected dependent variable. When a segmenting variable has more than two values, AID searches for the best division of its scale. Given a 4-point scale, for example, AID would evaluate the splits between: 1 and 2-4, and between 1-2 and 3-4, and between 1-3 and 4.

After the best split is found, the procedure is reapplied independently to each of the identified subgroups. Through successive iterations, a set of mutually exclusive segments is identified.

One of the major advantages of the technique is that intermediate results convey useful information on how segments are formed. When displayed diagrammatically, the steps of the process are easily traceable either backward or forward. And because the accompanying statistics are frequencies and means or percentages, the findings are readily understandable by those untutored in multivariate analysis.

AID AND OTHER MULTIVARIATE TECHNIQUES

The Advantages of AID

True to its name, AID excels at revealing interactions among variables. Sonquist (1970) has found that AID can represent interactive models clearly and accurately even in the presence of noise. Although multiple regression, analysis of variance, and discriminant analysis can incorporate interaction terms, beyond three or four predictors the specification of such terms is cumbersome. AID can easily accommodate interactions among 40 predictors or more (depending solely upon the limits of the computer program). This facility to assess the relative importance of a large number of interactions gives AID its special advantage for finding the optimal segmentation of a market.

This advantage derives from the unique way AID measures effects. While other analytic methods typically assess the magnitude of effects *over the entire sample,* AID measures influences that are important within specific subgroups. For example, AID could find that a specific academic program is very attractive to commuters but has little appeal to the larger group of resident students. Yet a technique such as discriminant analysis might easily pass over this interaction in favor of a predictor that has a lesser impact on commuters but that has a larger effect overall. For this reason AID is better suited, as an analytical tool, for informing a differentiated marketing strategy.

Unlike the restrictive assumptions required by most other multivariate methods, AID needs minimal distributional and measurement assumptions and no linearity assumption. With this flexibility, AID is applicable to most student data bases, including information that academic offices routinely collect, as well as data from studies originated by academic marketers.

The Disadvantages of AID

AID's biggest disadvantage is that it is a-theoretical. Control can be exercised by specifying which variables are available for inclusion, but the danger is always present that misinterpretation of the substantive significance of a predictor may occur. This shortcoming is shared with other stepwise techniques, however, and does not diminish AID's applied research benefits.

As with other stepwise procedures, AID may be expected to capitalize on chance associations within a sample. That is, the selection at each step of one variable over another may be idiosyncratic due to sampling variability. Yet while the overall branching structure of the solution is likely to differ slightly from sample to sample, the definition of final segments has greater reliability (Sonquist, 1970). Analysis should, however, be validated through replication. To further reduce the likelihood of idiosyncratic solutions, it is desirable to perform all analyses on samples of 1,000 observations or larger.

Sonquist (1970) has also shown that AID is sensitive to skewness (greater than 20:80) in either the endogenous or exogenous variables and to the number of categories in the scales of predictors. It is recommended that variables be transformed to even out distributions and to equalize the number of categories.

DATA SOURCES

Data are from two sources: (1) the application form from all Boston College applicants accepted for fall 1980 entry (n=4,400) and (2) responses to the Admissions Research Questionnaire from two mailings in June and July of 1980 to the same accepted applicants. The response rate to the questionnaire was 77 percent (88% for matriculants and 66% for nonmatriculants).[3]

METHODS

The OSIRIS version of AID that was employed (Institute for Social Research, 1973) includes a number of binary segmentation options.[4] The THAID subroutine was chosen since it is appropriate for nominal level dependent variables. Furthermore, THAID with the simple distance criterion *delta* characteristically selects segments of roughly similar size. This property is preferred to sums-of squares solutions that tend to select small outlying subgroups. The identification of smaller groups often improves the ability to predict individual behavior but is not well suited to the policy aims of market segmentation.[5] To adjust for the disparity in sampling rates between matriculants and nonmatriculants (see Data Sources above), observations are weighted to reflect their correct representation in the overall accepted applicant pool.

Two models are estimated. For both, segmentation is optimized on Admissions Yield (the probability of matriculation). In the first model, the accepted applicant pool is segmented using those predictors known at the time of application. See Table 1 for a list of these variables. This analysis illustrates how market segmentation can provide a basis for specific treatment of groups from readily available information. In the second model, the accepted applicant pool is segmented using all the predictors known at application, plus those measured on the 1980 Admissions Research Questionnaire. As may be observed in Table 2, these additional variables are primarily perceptions and evaluations of various aspects of Boston College. This analysis illustrates how AID can provide insights into the process of college choice. In particular, it highlights some of Boston College's strengths and weaknesses and shows how opportunities for improvement may be identified.

Model 1

Before interpreting the full AID model, let us examine how AID performs its tasks by first reviewing Figure 1. In the interests of economy of presentation and interpretation, only the first two iterations of the THAID solution are diagrammed. Segmentation begins at the first iteration by dividing accepted applicants into two subgroups (segments) according to their SAT scores. THAID searched all the predictors in Table 1, tested all the possible splits, and determined that the largest weighted difference (*delta*=.20) in admissions yield was between those who scored below 1100 (Segment B) and those who scored 1100

TABLE 1. Variables Eligible for Inclusion in Model 1.

Variable Name	Value Labels
Sex	1 = male, 2 = female
Alumni	1 = from alumni family, 0 = not
Athlete	1 = in university athletic program, 0 = not
Race	1 = Black, 2 = American Indian, 3 = White, 4 = Asian, 5 = Hispanic, 6 = Other
Faculty	1 = from faculty family, 0 = not
Entrance college	1 = Arts and Sciences, 7 = Management, 8 = Nursing, 9 = Education
Dorm application	1 = resident, 2 = commuter, 3 applied as resident, but accepted as commuter
SAT scores	1 = 400-899, 2 = 900-999, 3 = 1000-1099, 4 = 1100-1199, 5 = 1200-1299, 6 = 1300-1600
High school percentile	1 = below the 80th, 2 = 80th to 85th, 3 = 85th to 90th, 4 = 90th to 95th, 5 = 95th or higher
Financial aid applicant	1 = expressed intention to apply for financial aid, 0 = did not
Geography	1 = Greater Boston, 2 = rest of Massachusetts, 3 = NY, NJ or CT, 4 = rest of United States, 5 = foreign students

or above (Segment C) on their combined verbal and math SATs. In the second iteration, THAID is reapplied independently to Segment B and to Segment C. The maximal split of those who scored below 1100 on their SAT's (B) is between applicants to the Schools of Management or Nursing (D) and applicants to the Schools of Arts and Sciences or Education (E). Illustrating that the same predictor may be selected more than once, Segment C again is best divided by applicant's SAT scores (see Segments F and G in Figure 1).

Each box thus represents a segment that is precisely identifiable by the predictors that lead to it, and the branches detail the process by which segments are selected. On the second line of each box is the estimate of segment size and on the third line is the proportion of the segment who matriculated.

The pattern of branching is informative. If the pattern is symmetrical, i.e. the same predictors are selected at each iteration, then no significant interaction effects exist. Asymmetrical branching, then, is *prima facie* evidence of significant interaction effects.[6] In Figure 1, the asymmetry between the second-iteration branches (from Segment B and from Segment C) suggests that an interaction *may* exist between SAT scores and school applied to. In other words, school applied to has an effect on the admissions yield that is apparently of greater importance among applicants who score below 1100 on the SATs (Segment B) than among those who score higher on the SATs (Segment C).[7]

TABLE 2. The Additional Predictors Available for Inclusion in Model 2.

Variable	Description
Applicant rating	Overall rating of applicant by Admissions staff on a scale of 1-10: 1=top 1%, 2=next 4%, 3=next 10%, 4=next 10%, 5=next 10%, 6=next 15%, 7=next 15%, 8=next 15%, 9=next 10%, 10=next 10%
Boston	Rating of "Boston area as a place to attend a college or university" on a scale of 1-5: 1=poor, 2=fair, 3=good, 4=excellent, 5=the best
Jesuit effect	Response to the question, "Do you think Boston College's Jesuit tradition affects the quality of education provided?" on a scale of 1-5: 1=detracts considerably, 2=detracts somewhat, 3=no effect, 4=improves somewhat, 5=improves considerably
Postgraduate plans	Response to, "Do you plan to attend a professional/graduate school after graduation from college?": 1=no, 2=unsure, 3=yes
Financial aid	Evaluation of these attributes of Boston College on a 1-5 scale: 1=unsatisfactory, 5=excellent
Distance from home	
College faculty	
Social activities	
Teaching reputation	
Parents' preference	
Size of school	
Quality of students	
Admissions personnel	
Attractiveness of campus	
Variety of courses	
Specific academic programs	
General reputation	
Reputation of alumni	
Athletic programs	
Honors programs	
Location of campus	
High school counselor's rating	
Religious opportunities	
Housing opportunities	
Admissions literature	

TABLE 2. *(Continued)*

Variable	Description
Employment opportunities after graduation at Boston College	

Figure 2 is a diagram of the full THAID solution. The boxes have been aligned with the Admissions Yield scale at the left so that segments of higher and lower yield are easily compared. The diagram begins with three segments (rather than two) because SAT score was selected again in the second iteration for one of the first iteration groups (cf. Figure 1) and is treated here as an elaboration of the first to conserve space.

Clearly, applicants' SAT scores provide the best single way to divide accepted applicants according to their probability of enrolling. The yield is lowest for applicants who score above 1200 (.33), reflecting competition for these students,

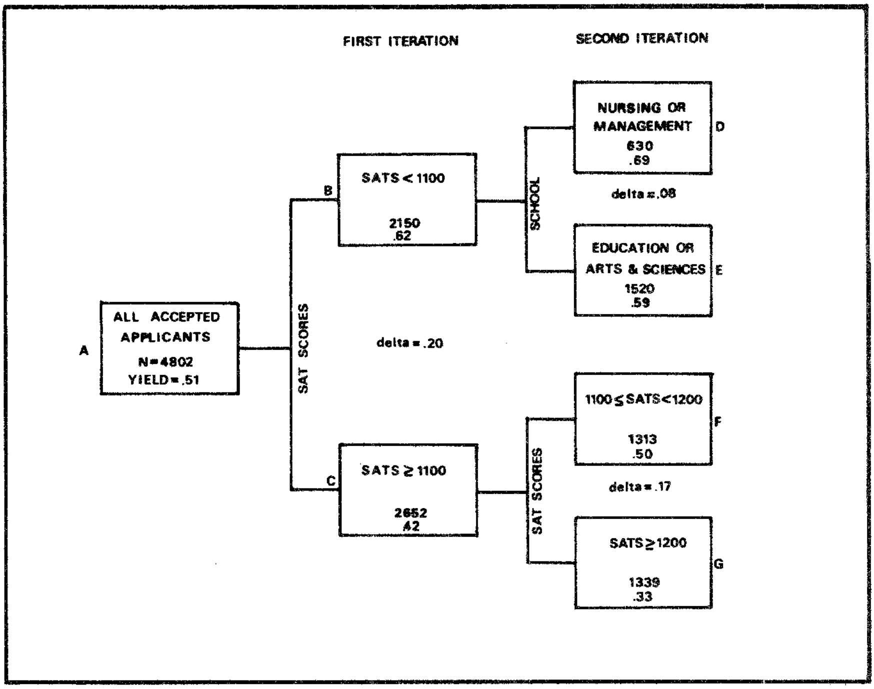

FIGURE 1. First Two Iterations of AID Segmentation Using Predictors Known at Application.

Note. See Table 1 for definitions of variables eligible for inclusion.

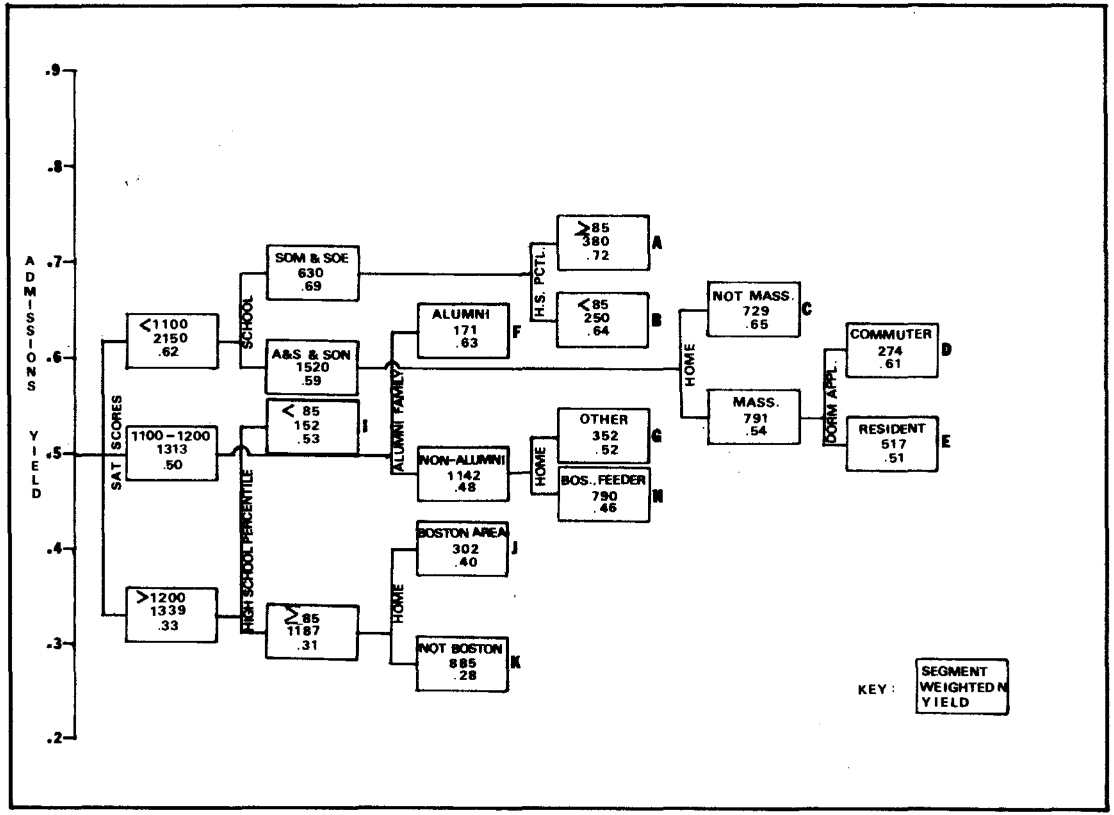

FIGURE 2. AID Segmentation Using Predictors Known at Application.

Note. See Table 1 for definitions of variables eligible for inclusion.

and is highest for those who score below 1100 (.62). Significant interaction effects are apparent by the pronounced asymmetry among the branches originating from the three SAT segments. Thus this analysis is likely to give information that most other techniques would have missed.

Two principal segmentation strategies exist through which the academic marketer can influence the number and mix of prospective applicants: find more students similar to those who presently exhibit high yield rates or improve the yield rates of other groups through promotion, program development, pricing, or combinations of these factors.

In Table 3, the 11 segments at the terminus of each branch are grouped by the responsiveness of segment members to Boston College's offer of admission. Segments A, B, C, F, and D are already highly responsive (yield greater than .60). One strategy for a marketer who wishes to increase enrollments would be to locate and attract prospective applicants with the characteristics of these high-yield groups. Notice that segment F is the only one in the highly responsive group that falls into the middle SAT segment, indicating that alumni families are a significant factor in influencing better-scoring children to attend.

Among the moderately responsive (yield between .50 and .60) segments I, G, and E, the correlation with level of SAT scores breaks down. That is, other defining characteristics take on explanatory importance. For instance, applicants who do not perform as well in high school are more likely to attend Boston College even though their high SAT scores predict otherwise. The policy implication is that if the university wants to enroll more high SAT scorers, then more effort should be placed on attracting the lower achieving applicants similar to those in Segment I. The information on Segment G also may be revealing something encouraging. The yield among medium level SAT scorers is moderately good outside of geographic areas on which Boston College has put primary emphasis. If feasible, effort might be redistributed toward attracting relatively good students in these other areas. Segments I, G, and E might also be responsive to an increase in their yields through more effective, targeted promotional efforts, through program/pricing changes, or all of the above.

Finally, consider the mildly responsive (yield less than .50) segments: H, J, and K. This grouping isolates those segments that Boston College has least success enrolling. K is a relatively large segment of desirable students who manifest a low yield. A low yield might indicate an opportunity for Boston College to improve, *if* these students were being ''neglected.'' This is not the case, however, as Boston College and other schools recruit these students vigorously. Thus the cost of raising that yield is likely to be high. If small increases in yield are deemed worth the cost, then more research should be done to understand how the interests and perceptions of segment members may be better matched with what Boston College has to offer, either through making them more aware of Boston College's present offerings or by offering programs that would be more appeal-

TABLE 3. Responsiveness of Segments Identified in Figure 2

Segment	n[a]	Characteristics	Admissions Yield (Y)
		Highly Responsive ($Y>.60$)	
A	380	SAT score less than 1100, Management or Education applicant, and above 85th percentile in high school class	.72
B	250	SAT score less than 1100, Management or Education applicant, and below 85th percentile in high school class	.64
C	729	SAT score less than 1100, Arts and Sciences or Nursing applicant, and not from Massachusetts	.65
F	171	SAT score between 1100 and 1200 and from Alumni family.	.63
D	274	SAT score less than 1100, Arts and Sciences or Nursing applicant, from Massachusetts, and desires to commute	.61
		Moderately Responsive ($.50<Y<.60$)	
I	152	SAT score greater than 1200, and below the 85th percentile in high school	.53
G	352	SAT score between 1100 and 1200, from non-Alumni family, and not from Boston, New York state, New Jersey or Connecticut	.52
E	517	SAT score less than 1100, Arts and Sciences or Nursing applicant, from Massachusetts, and desires to live on campus	.51
		Mildly Responsive ($Y<.50$)	
H	790	SAT score between 1100 and 1200, from non-Alumni family, and from Boston, New York, New Jersey or Connecticut	.46
J	302	SAT score greater than 1200, at or above 85th percentile in high school class, and from Boston area	.40
K	885	SAT score greater than 1200, at or above 85th percentile in high school class, and not from Boston area	.28

[a] Estimated by adjusting survey responses by sampling rates. Sums to 4,802 because weighting factors were rounded to nearest integers.

ing. With this knowledge, specific policies may be directed toward meeting their special needs.

An encouraging pattern may be found leading to segment J. Boston College is able to matriculate a higher proportion of high SAT scorers who perform well in high school if they are from the Boston area. The Admissions Office should be advised to do as much as possible to cultivate this slight advantage.

Model 2

In Figure 3, 12 segments are identified in a THAID model that includes both characteristics of applicants and their perceptions (see Tables 1 and 2). With the addition of subjective measures the ability to effectively segment accepted applicants is enhanced markedly. This is reflected in the greater range in yields: from .21 in segment L to .97 in segment A.

Parents' Preference is the predictor that first segments this pool, emphasizing the pivotal importance the family plays in college choice among Boston College applicants. If parents are perceived to be neutral (3) or negative (1,2), the percentage who matriculate is only .34. Yet at the other extreme, if parents are perceived to rate Boston College as excellent (5), the matriculation percentage is almost .80. The importance of involving parents cannot be overstressed, if the Admissions Office is to be effective.

In interpreting Figure 3, keep in mind that the branches do not imply chains of causation or the logical priority of one variable over another. These are determinations that must be made by the researcher. One can, however, rank the stepwise predictive importance of variables, and this information can be very helpful. For example, while the yield is only .34 in the lowest Parents' Preference segment, the next branch shows that there is an opportunity for Admissions Personnel to make an impact. If applicants rate Admissions Personnel positively (4-5) then the yield is raised to .50. AID shows how one marketing resource can offset the negative effects of another factor.

Students who believe that their parents strongly favor Boston College and who rate its location, programs, and financial aid highly are about as certain to attend as any college can expect (a 97% yield in segment A). In defining segments C and D, the applicant's gender makes a substantial difference. The reason for lower yields among males deserves further research.

Among those in the middle Parents' Preference segment, size is apparently a concern. Interestingly, even when Boston College is rated poorly the yield is fairly high (.52 in segment G) *if* Social Activities are looked upon favorably. This supports the notion that the ill effects of size are due to the perceived impersonal nature of a larger school. Increased "personalization" of contact with applicants may be an effective way for Admissions to reduce the negative effects of size. This analysis shows that adherence to this time-honored precept is

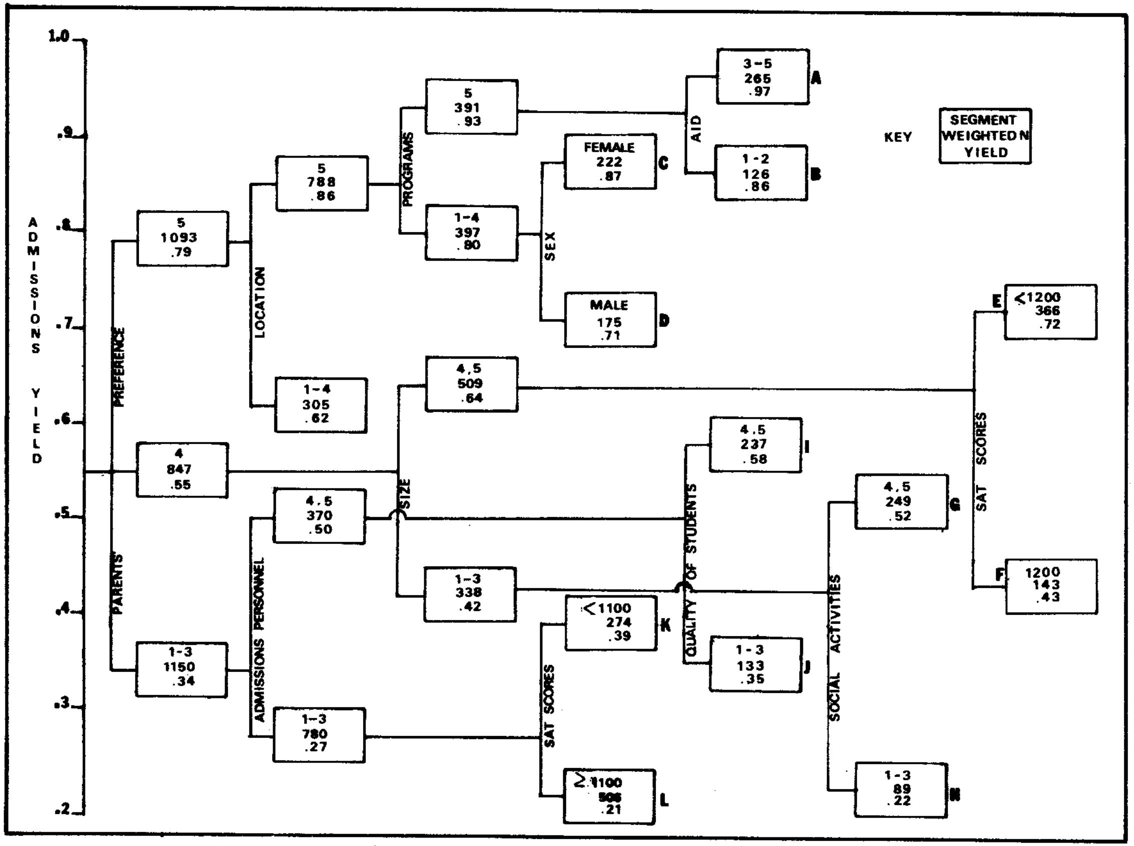

FIGURE 3. AID Segmentation Using Application and Questionnaire Predictors.

Note. See Tables 1 and 2 for definitions of variables eligible for inclusion.

particularly important when parents are not totally enamored of Boston College.

It is also interesting to note that student "quality," through either perceptions of Boston College students or in the characteristics of applicants, again explains low yields in the segments in the lower half of Figure 3 (see segments F, J, and L). These students might be attracted by special, high-quality programs (such as College Honors Programs) that exist for academically superior students. It might be worthwhile to place present Boston College students in these programs in contact with these prospective students.

IMPLICATIONS

AID can yield significant policy benefits by itself. The technique, however, should be viewed as an efficient first step to a full-blown market analysis. Its greatest benefit in this regard is that once distinctive segments are identified, research may be directed to areas of specific opportunity or need. In most cases this implies an intensive study of the images that members of specific segments have of the college and of the appraisal process they go through in selecting a college.[8]

In addition, Sonquist (1970) recommends that the findings from AID be used to reestimate a multiple classification analysis model taking account of interactions to estimate the magnitude of effects across the entire sample. Likewise, one might want to choose a set of segments identified by AID and apply a discriminant analysis to provide a fuller view of the dimensions that differentiate the segments.

AID is clearly a flexible technique that deserves the serious consideration of anyone doing research for an admissions office. But just as this analysis was performed on accepted applicants, AID could be used to segment the inquiry pool. Such information would be invaluable for shaping a policy to convert qualified students who seek information about a college into the desired kinds of applicants.

AID is also applicable to other student groups and to other policy questions. For example, one could segment present students with respect to their persistence rates. With knowledge of the characteristics of subgroups most likely to leave school, policy might be directed toward increasing students' probability of persisting. AID may be able to identify interactive influences that other studies have missed.

Student activities, residential life, alumni relations, career planning and placement, development, academic affairs and individual academic departments all make policy decisions concerning ways to better meet the needs and preferences of different types of students. AID has the potential for providing these and other campus agencies with insights that inform decision making in ways that other analytic techniques cannot.

NOTES

1. The authors know of only three applications of AID to higher education research, all in unpublished reports: Strommen (1976), Whittstruck and Inguanzo (1980), and Institute for Research on Social Behavior (1980).
2. See references listed on pp. 231-234 in Sonquist, 1970, and on pp. 225-257 in Fielding, 1977.
3. More information on the questionnaire or on its administration is available upon request.
4. In OSIRIS IV, AID subprograms are called SEARCH. Contact the Institute for Social Research, University of Michigan, for more information.
5. See Fielding (1977) pp. 225-234 for a review of split criteria options.
6. Asymmetry alone is not sufficient to demonstrate the existence of interaction effects. Analysis of variance (or some other multivariate technique) may be specified to test for the statistical significance of identified interactions above that of main effects.
7. Again, these statements cannot be definitive because the importance of a predictor, in this case "school," could be of equal importance in both segments, yet be overshadowed by a larger split achieved by another predictor—here, a further split of segment C on SATs. A comparison of Admissions Yields by School between segment B and segment C would need to be done to confirm the existence of an interaction.
8. See Maguire and Lay, 1981, for specific methods that may be applied to model and compare the processes of image making and decision making across market segments.

REFERENCES

Fielding, A. Binary segmentation: the automatic interaction detector and related techniques for exploring data structure. In C. A. O'Muircheartaigh and C. Payne (Eds), *The analysis of survey data,* Vol. 1. New York: Wiley, 1977.

Institute for Research in Social Behavior. Retirement plans and related factors among faculty at COHFE institutions. Report for Consortium on Financing Higher Education, 1980.

Institute for Social Research. *OSIRIS III,* Release 2 Edition. Ann Arbor: University of Michigan, 1973.

Litten, L. Market structure and institutional position in geographic market segments. *Research in Higher Education,* 1979, *2,* 59-83.

Maguire, J., and Lay, R. S. Modeling the college choice process: image and decision. *College and University,* 1981, *57,* 123-139.

Morgan, J. N., and Sonquist, J. A. Problems in the analysis of survey data: and a proposal. *Journal of the American Statistical Association,* 1963, *58,* 414-434.

Sonquist, J. A. *Multivariate model building: the validation of a search strategy.* Ann Arbor, Michigan: Institute for Social Research, 1970.

Strommen, M. P. A survey of images and expectations of LCA colleges. Research report to the Joint Committee of the Division for Mission in North America and the Council of LCA Colleges, 1976.

Whittstruck, J., and Inguanzo, J. Research report to Nebraska Post-Secondary Education Commission, 1980.

Received September 18, 1981.