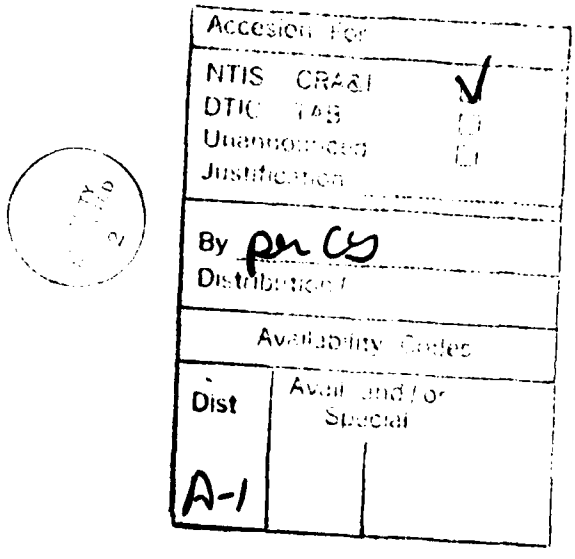
Small-Area Estimates for Military Personnel Planning

Report of a Workshop

Committee on National Statistics
Commission on Behavioral and Social Sciences and Education
National Research Council



NATIONAL ACADEMY PRESS
Washington, D.C. 1989

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This report has been reviewed by a group other than the authors according to procedures approved by a Report Review Committee consisting of members of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine.

The National Academy of Sciences is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Frank Press is president of the National Academy of Sciences.

The National Academy of Engineering was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of & Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Robert M. White is president of the National Academy of Engineering.

The Institute of Medicine was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Samuel O. Thier is president of the Institute of Medicine.

The National Research Council was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Frank Press and Dr. Robert M. White are chairman and vice chairman, respectively, of the National Research Council.

The workshop that is the subject of this report was supported by funds from the Defense Manpower Data Center (DMDC), U.S. Department of Defense.

Available from:
Committee on National Statistics
National Research Council
2101 Constitution Avenue, N.W.
Washington, D.C. 20418

WORKSHOP PARTICIPANTS
June 10-11, 1988

KEN HILL *(Chair)*, Department of Population Dynamics, The Johns Hopkins University
KEN BERLIANT, Program Analysis and Evaluation Headquarters, U.S. Army Recruiting Command, Fort Sheridan, Ill.
MARK CHIPMAN, Navy Personnel Research and Development Center, San Diego, Calif.
MICHAEL L. COHEN, Consultant, Committee on National Statistics
DENNY EAKLE, Directorate for Accession Policy, OASD, Washington, D.C.
ROBERT FAY, Statistical Methods Division, Bureau of the Census
PAUL D. FLOWERS, Headquarters, U.S. Marine Corps, Washington, D.C.
MARTIN R. FRANKEL, Baruch College
DAVID GRISMER, Rand Corporation, Washington, D.C.
DAVID HARVILLE, Department of Statistics, Iowa State University
RICHARD HEATHCOTE, Navy Recruiting Command, Arlington, Va.
JANICE HENDERSON-LAURENCE, Human Resources Research Organization (HUMRRO), Alexandria, Va.
JAMES HOSEK, Rand Corporation, Santa Monica, Calif.
LINDA INGRAM, Consultant, Committee on National Statistics
MICHAEL LAURENCE, Survey & Market Analysis Division, Defense Manpower Data Center, Arlington, Va.
DANIEL B. LEVINE, Consultant, Committee on National Statistics
BETTE MAHONEY, Survey & Market Analysis Division, Defense Manpower Data Center, Arlington, Va.
DAVID MARKER, Westat, Inc., Rockville, Md.
PETER H. MIYARES, U.S. Air Force Recruiting Service, Randolph Air Force Base, Tex.
CARL MORRIS, Department of Mathematics, University of Texas, Austin
GAD NATHAN, National Center for Health Statistics, Hyattsville, Md.
PAUL NICKENS, Defense Manpower Data Center, Arlington, Va.
BILLY NIX, Program Analysis and Evaluation, Headquarters, U.S. Army Recruiting Command, Fort Sheridan, Ill.
STEPHEN J. OREN, U.S. Marine Corps Recruiting Service, New Orleans, La.
JEFFREY PASSEL, Population Division, Bureau of the Census
MARY POWERS, Graduate School of Arts and Sciences, Fordham University
S. JAMES PRESS, Department of Statistics, University of California, Riverside
J.N.K. RAO, Department of Mathematics and Statistics, Carleton University
WESLEY SCHAIBLE, Office of Research and Education, Bureau of Labor Statistics
WILLIAM H. SIMS, Manpower and Training, Center for Naval Analysis, Alexandria, Va.
BRUCE SPENCER, Department of Statistics, Northwestern University
TOMMY WRIGHT, Oak Ridge National Laboratories, Oak Ridge, Tenn.
NAOMI VERDUGO, Manpower and Personnel Policy Research Group, U.S. Army Research Institute, Alexandria, Va.
MEYER ZITTER, Consultant, Committee on National Statistics

COMMITTEE ON NATIONAL STATISTICS
1988-1989

ACKNOWLEDGMENTS

Many people contributed time and expertise to the workshop, and the committee is most appreciative of their cooperation and assistance. In particular, Kenneth Hill served most ably as chair and gave freely of his advice and counsel, and the participants in the workshop contributed their many thoughts and comments to shaping this report.

The report was prepared by Meyer Zitter, Daniel Levine, and Linda Ingram, who also were responsible for the conduct of the workshop. The difficult task of assembling the technical appendix on the possible application of empirical Bayes regression models to small-area estimation was accomplished by Michael Cohen, who also served as consultant to the workshop. Administrative Secretaries Evelyn Simeon and Jennifer Lane assisted with the workshop. Project Assistants Anu Pemmarazu, Carrie Kakes, and Daniel Keen helped to prepare the report for publication, and Administrative Assistant Ann Marie Laskiewicz-Ross provided valuable assistance in making the final editing changes. The report also benefited from the thoughtful comments of reviewers and the editorial assistance of Eugenia Grohman and Christine McShane of the Commission on Social and Behavioral Sciences and Education.

CONTENTS

SMALL-AREA ESTIMATES FOR MILITARY PERSONNEL PLANNING

INTRODUCTION

Since the end of the military draft in 1972, military personnel planners have been faced with the problem of recruiting new enlistees to maintain prescribed military levels and staff the various complex jobs involved in maintaining a modern military. At present, the Department of Defense, through the individual branches of the services, annually recruits approximately 300,000 enlistees without prior military service. Each service's overall recruitment goal is established consistent with its budget and overall defense plan. Each service has its own recruitment goals and targets, maintains a recruiting network, and has its individual procedures for achieving its mission of recruiting the required number of enlistees. The services have also developed various methodologies (and data sets) for estimating the size, composition, and geographic location (at some stated small-area level such as county) of the main population groups from which it draws. Such information is extremely important and valuable for targeting recruitment efforts, establishing recruitment goals, and evaluating the performance of such efforts. (SDW)

Given the need to maintain armed forces at assigned strength levels and the role of appropriate data bases in assisting recruiting efforts and assigning goals, the Defense Manpower Data Center (DMDC) of the Department of Defense approached the Committee on National Statistics (CNSTAT) to convene a group of experts in a workshop setting to assess the merits of various techniques used by the military services to develop small-area estimates of those qualified for military services. An ultimate aim of the workshop as expressed by the sponsor was to decide which among the several methods shall be used to make the estimates, i.e., to provide one common set of "best" estimates that could be used as a starting base for all services. Such estimates would provide a common framework within which the services could establish comparative recruiting goals at the small-area level and design recruiting strategies. Accordingly, CNSTAT assembled a group of experts from a variety of related disciplines, including demographic estimation, statistics, economics, model building, military recruitment, and survey design.

These were joined by representatives of the various services, including military recruiters and research support groups, to provide input on current methods, data sources, and other problems and issues faced by the several recruitment networks in carrying out their mission.

This report reviews the workshop presentations and discussions and details its findings and recommendations. The report is organized as follows: the first section describes the structure and organization of the workshop, as well as the concepts and definitions underlying the main issues. The next section provides an overview of the discussion and summary of the workshop recommendations. The main workshop discussion on methodology, concepts, and definitions follows; the next section reviews and summarizes the deliberations of the two subgroups, on methods and data needs. The report concludes with a section of thoughts by the workshop chair. An appendix completes the volume: it is a discussion of the use of empirical Bayes regression models, which was a major workshop recommendation for investigating a methodology for developing small-area estimates of the population of interest.

STRUCTURE OF THE WORKSHOP

The workshop began with presentations from the Department of Defense, each of the services, and selected research support groups. The presentations covered a number of the main issues, including concepts and definition of both the basic pool from which the military draw their recruits and the concept of propensity to enlist; methods and procedures used in estimating QMA down to small levels of geography (recruiting station, districts, counties, etc.), especially how existing national-level data are used in this process; compiling and developing other types of data useful in setting enlistment goals and carrying out recruiting efforts (e.g. the Recruit Market Network); and the use of the data in the recruiting process. The presentations were informal, permitting intervention and discussion during and after each presentation. The agenda was designed to impart sufficient information for a knowledgeable discussion of the issues at hand.

The workshop next divided into two groups, one to focus on methodology and the other on data requirements; not surprisingly, there was considerable overlap in the discussions of the two groups. The following morning, the workshop reassembled in plenary session, heard reports from each of the groups, summarized the discussion, and agreed on its recommendations.

The key concept on which the workshop focused is known as the qualified military available (QMA), which is generally defined as the number of 17- to 21-year-old male high school graduates (including those with equivalency diplomas) who are mentally and physically qualified for military service and of good moral character. This is the target population of military eligibles to which the services primarily direct their recruiting efforts. Women in the target group are excluded from the determination of QMA, since the services to date continue to be oversubscribed in their recruitment efforts among women.

The criteria are defined as follows: *Mentally qualified* relates to aptitude levels with specific reference to how individuals would score on the Armed Services Vocational Aptitude Battery (ASVAB) or the Armed Forces Qualification Test (AFQT), a subset of the ASVAB.

The distribution of such scores, especially divided between upper and lower classifications or percentiles, is of special concern; *Physically qualified* are those who are medically fit for military service, i.e., they meet the standards of medical fitness as specified by the Department of Defense in its directive of March 31, 1986; *Good moral character* relates to standards of "character and past behavior" and are designed to prevent undesirable people (e.g., criminals) from entering the services.

Another concept important to recruitment efforts is *propensity*, which is defined as the likelihood that a qualified individual will actually enlist in a particular branch of the service. Such information represents an important supplement to the base number of eligibles in the military-potential pool of youth. Only a limited amount of survey-based information reflecting on propensity is available, mainly derived from the Department of Defense-sponsored Youth Attitude Tracking Survey (YATS). The survey obtains information at the national level that research shows provides a measure of the likelihood of young adults enlisting in the military, by service. The survey responses reflect, among other elements, the effects of economic factors on propensity, intention, and attitudes toward military enlistment incentives.

OVERVIEW AND SUMMARY OF RECOMMENDATIONS

As noted, one main objective of the workshop was to consider which methodology could provide a common set of "best" estimates (of QMA) at small levels of geography to be used as a starting base for all services. However, it became increasingly apparent as the discussion proceeded that a dichotomy existed between the thinking of those at the headquarters' level of the services and those at the subnational level. This same situation had been noted by staff in interacting with the different levels in preparing for the workshop. Specifically, although the charge to the workshop dealt with assessing the different methodologies employed by the individual services in developing their estimates of QMA, in fact, those at the local level charged with meeting an established quota had only marginal interest either in the specific methodology or in the QMA estimates derived therefrom. Theirs was a much simpler perception--namely, that the skill required of the local recruiter (at any level below the national) was that of an outstanding salesperson--one who had intimate knowledge of his or her recruiting area and of what had transpired in past years. That meant close association with the high schools and complete familiarity with the size and composition of the class scheduled to graduate in the current recruiting year. Such information, in combination with knowledge of past performance and local conditions, is the main tool of the local recruiter that permit him or her to estimate within a reasonable range the percentage of enlistees that would be forthcoming from the local high schools, some indication of service preference, and military specialty desired. Thus, how well the recruiter could accomplish the mission and the reasonableness of established targets were products of local, specialized knowledge rather than of national data sets.

The comments of those at the local level concerning the quality and utility of QMA were more a reflection of comments from the national level than an expression of their interest in QMA as a tool for carrying out their own mission. In a similar vein, concerns about the quality of the data among those at the national levels within the services seemed to reflect a wisdom handed down from the past, and did not seem to be substantiated by specific examples

detailing either shortcomings or difficulties caused by data inadequacies. Nor was evidence presented to indicate that current systems were inadequate to provide guidance on setting overall recruiting goals for broad areas, such as the first geographic level of the recruit network (brigades, for example, for the Army).

In effect, then, workshop participants found themselves dealing with a somewhat amorphous problem. Indeed they could identify data shortcomings; improvements were possible in the data and, to a lesser extent, in the methods themselves; participants had no difficulty in agreeing on the importance of such improvements in providing information to be used in setting broad regional goals and/or for future planning purposes. It seemed doubtful, however, as to whether these improvements would result in significant or measurable benefit to recruiters at the local level. In fact, any of the methods now in use appear to provide an adequate degree of direction in developing both overall recruitment goals and broad controls at regional or state grouping levels.

It also was not clear from the discussions that any one set of data could uniformly and systematically meet the needs of all service branches. But it was generally understood that a common set of QMA estimates uniformly available at the county (or other small-area) level prepared by acceptable, statistically sound procedures would be a useful product and could serve as a common starting point for the services' recruiting networks. Indeed, a consistent set of estimates would provide benchmark data and a framework within which each of the services could establish comparative and competitive recruiting goals at the small-area level.

In the same vein, no consensus was reached with regard to the concept of propensity. In fact, there were questions raised on whether propensity can even be clearly defined, understood, or quantified locally, and no agreement was reached on how best it could, or should, be used in conjunction with data on QMA.

The discussion clearly brought out the present difficulties in obtaining current data at subnational levels for the elements that comprise QMA, especially for the physical and moral criteria. Nonetheless, participants agreed that efforts should be made to obtain better information for each of the elements in the future and at the same time to explore the development of models that might serve in the interim. Accordingly, the recommendations developed by the workshop reflect both goals--first, that of seeking out additional data sources and, second, that of developing approaches to maximize the value and use of currently available data. In effect, the workshop attempted to strike a balance between the Department of Defense's interest in providing or developing a "best" common set of estimates to be used by all services in setting recruiting goals and the realities of different situations faced by recruiters locally in meeting established goals.

Recognizing this dilemma, the group made only one main recommendation:

> **A regression sample-data method should be investigated as a means of measuring local AFQT distributions (the major measurable component of interest).**

Empirical Bayes regression modeling was specifically recommended, and the details of a plan to test and evaluate this approach using the state of Florida (assuming they carry out their proposal to administer the AFQT to all male high school seniors during the 1988-1989 school year) are spelled out in the appendix.

The group also recognized the varying data needs of different levels of the military recruitment network. At the national and regional levels, the concerns involve setting recruitment goals and evaluating the overall efficiency in achieving these goals; their interest is in a standard set of QMA figures across the country. At the lower levels, however, focus shifts to recruitment efforts to achieve missions and the use of local data (and knowledge) to more realistically assess and set goals.

Accordingly, the group proposed that DMDC:

- Investigate the possibility of using ongoing federal surveys to collect data on elements of QMA, especially those dealing with mental qualification (e.g., proxies for AFQT). Special note was made of the need to explore the usefulness of the National Assessment of Educational Progress (NAEP) and the National Education Longitudinal Survey (NELS), or any other suitable longitudinal study of youth.

- When future surveys or data collection efforts are sponsored by the Department of Defense, they should be planned and designed to measure physical and moral qualifications as well as mental qualifications (for the present, however, since no subnational data are available on the physical and moral qualification components, the consensus was to concentrate on the mental or aptitude component of QMA).

- Investigate the feasibility of using the large number of tests (AFQT) already being given at high schools around the country, while recognizing that at present they represent selective sampling, thus making it difficult to use the results in a generalized program of small-area estimates. To this end, it was proposed that some part of the annual appropriation be set aside for experimentation purposes. Indeed, in light of the potential for developing reliable estimates of QMA at the local level, a reexamination of the allocation of resources (to permit proper sampling) might be in order.

- The utility of other existing data sources that might be used for model-based updating of small-area estimates should be explored. Examples include data on school enrollment and number of high school graduates for high school districts, collected by the National Center for Education Statistics.

- Update the 1980 Profile of American Youth, using an existing longitudinal survey of youth or a similar survey, including the possibility of an add-on to the Current Population Survey, which already collects data on education, to develop proxy measures.

- Undertake work to determine the usefulness of the concept of propensity and develop approaches permitting its measurement.

The workshop had concluded early in its deliberation on estimating the physically qualified that, given the present data systems and the limited funds available, it is virtually impossible to measure or estimate local differentials in expected disqualification rates for failure to meet physical standards. The issue of estimating good moral character was viewed as even more intractable.

ESTIMATING QMA AT THE COUNTY LEVEL

Marine Corps and Navy Procedures[1]

The process of estimating QMA (at any geographic level) is a step-down procedure wherein one starts with an appropriate base population--in this instance, the number of males ages 17-21 and then successively eliminates nonqualifying groups, i.e., non-high school graduates, mentally unqualified, and those not meeting physical standards. The remaining group is classified as QMA. The main methodological issues involve how best to estimate the various dichotomies for areas below the national level. (Note: the problem of estimating those failing the "moral character" criterion was not addressed.)

The Base Population

The local-area estimation of QMA begins with an estimated base population composed of all 17- to 21-year-old males residing within a county. The group is subdivided by age and by racial/ethnic group (black, Hispanic, and white/other). The population estimates exclude institutionalized individuals and those not residing in a household (e.g., living on a military base or college dormitory). The base data derive from the 1980 census files, principally the 5 percent Public Use Micro Sample (PUMS). The 1980 population data were aged annually to generate estimates for future years (e.g., 1988, 1989), although how this was done was not made clear. Bureau of the Census projections were used to correct for migration and immigration changes.

Estimating the Educationally Qualified

From each estimated racial/ethnic group base population, an estimate of those educationally qualified for service was generated. An educationally qualified individual is one who has graduated from

---

[1]This section draws on Navy Personnel Research and Development Center (1987).

high school or has received an equivalency diploma. Estimates of high school completion rates were derived from the educational attainment data contained in the 1980 PUMS file.

Estimating the Mentally Qualified

The educationally qualified population was next apportioned into aptitude qualification groups. Aptitude qualification refers to achievement of a minimum (or higher) score on the Armed Forces Qualification Test. The AFQT scale is segmented into categories based on percentile scores. The categories and their corresponding percentile scores are shown in Table 1. Although the Marine Corps currently classifies as "aptitude qualified" I through IIIb (AFQT percentiles 31-99), the minimally acceptable score has varied over time, so the numbers in specific categories (e.g., I through IIIa, IIIb) were also estimated.

In 1980, the Department of Defense, in conjunction with the Department of Labor, administered the AFQT to a representative sample of 11,878 youths as part of the National Longitudinal Survey (NLS) (U.S. Department of Defense, 1982). From the study, known as the Profile of American Youth (often referred to as the Profile), county estimates of the percentage of individuals in each AFQT category (as defined in Table 1) were developed.

TABLE 1: AFQT Categories/Grades

Category/Grade	Percentile Score
I	93-99
II	65-92
IIIa	50-64
IIIb	31-49
IVa	21-30
IVb	16-20
IVc	10-15
V	01-09

First, a data set was established by merging individual respondent data from the NLS with the corresponding record from the Profile. This large set of variables was then analyzed to identify those that were highly correlated with AFQT scores and were available at the county level from other sources of data (e.g., census files). The most highly correlated variable was race/ethnicity. This finding meant that QMA could not be estimated accurately for a county without accounting for each county's racial/ethnic mix.

At that point, the sample was split into the three racial/ethnic group classifications--black, Hispanic, and white/other. Then, within each group, different combinations of variables that best predicted AFQT scores were identified.

Other variables that were highly correlated with AFQT and available at the county level included level of education (e.g., high school diploma, college degree, parental education), and socioeconomic status (e.g., father's occupation). "Level of education" alone provided maximum prediction for both the Hispanic and white/other groups, while a combination of "level of education" and "father's occupation" best explained black scores.

In order to infer county-level AFQT distributions from a Profile subgroup (e.g., Hispanics), counties were grouped into homogeneous clusters. The Profile individuals from the counties in a given cluster served as a representative sample for that cluster. Clustering was based on census-type variables considered surrogates of the predictor variables. Table 2 relates the predictor and surrogate variables.

All counties, including those without Profile representation, were grouped into a few homogeneous clusters based on the surrogate variable(s) value(s) of each racial/ethnic subgroup. Boundary values were chosen to yield roughly equal-sized Profile subsamples across the clusters.

The distribution of the Profile participants across AFQT categories was computed for each cluster. This distribution was then used to represent the AFQT category distribution of 17- to 21-year-old educationally qualified males in all counties belonging to the cluster. As noted, this process was followed for each of the three racial/ethnic groups.

TABLE 2: Predictor and Surrogate Variables

Ethnic Group	Profile Predictor Variable	Census/Other Surrogate Variables
Hispanic	Level of education	Percent of adult Hispanics in county with 12 or more years of education
Blacks	Level of education and father's occupation	Percent of adult blacks in county with $\geq$ 12 years education and socioeconomic status indicator of county
White/Other	Level of education	High school and college completion rates for adults

As a final step, the county-level AFQT category distributions were normalized to the national distribution.

Estimating the Physically Qualified[2]

Although a relatively large percentage of youth of military age fail to meet the required physical standards of enlistment, reliable, current data reflecting local variations are not available. Most of the available information comes from experience over the years with draftees and with failure rates of military applicants to the all-volunteer force. The latter are deemed biased by self-selection, and the former yielded some unusual results or at least several surprises relative to reasonable expectations. For the purposes of estimating QMA, a number of adjustments were made for those earlier data series and, based on the evidence from these adjustments, a physical disqualification rate of 14 percent was applied across the board--i.e., regardless of ethnic group or geographic location.

Since the studies described above were done, more recent national information based on data collected as part of the National

[2]This section is largely taken from Defense Manpower Data Center (1987).

Health and Nutrition Examination Survey, 1976-80 (NHANES-II) has become available suggesting quite different disqualification levels and significant male/female differentials.

Conducted by the National Center for Health Statistics, the NHANES-II is a nationally representative probability survey designed to collect a broad range of morbidity, health, and nutritional data from the civilian noninstitutionalized population. The medical fitness for military service of the 16- to 24-year-olds in the sample was based on an evaluation of their records and a determination as to whether they met the qualifications of the Department of Defense physical standards for enlistment.

The principal finding of this study was that 18.3 percent of the males and 41.4 percent of females in the general population 16- to 24-years-old was determined to be disqualified for military service under then current medical standards. Obesity was the leading cause for disqualification among both males and females, and the large difference between male and female rates of disqualification is due to methodological differences in the construction of male and female maximum weight standards. When methodologically consistent standards are applied, the large difference in male/female disqualification rates was substantially reduced: the overall male rate increased to 21.6, and the female rate reduced to 24.6 percent. Regardless of methodological variations, in actual practice, according to the discussion, male and female rejection rates for failure to meet physical standards are high due mainly to a failure to meet the height-weight standards.

Questions were raised concerning the differences by race and by ethnicity, but the NHANES-II sample was too small to provide meaningful data on the differentials for these groups. Differences among the services, especially for females, also are not clearly understood, since presumably the height-weight standards are the same for enlistees without prior service. Similarly, usable data on the relationship between physical disqualification and educational level could not be obtained from NHANES-II.

All in all, there was strong consensus that, given the present data systems and the limited funds available, measuring or estimating local differentials in expected disqualification rates for failure to meet physical standards is virtually impossible.

Army Methodology[3]

The method of estimating QMA for the Army at the battalion level is somewhat different, involving essentially a three-step process. (The county is geographically divided into 56 battalions.) First, estimates (and projections) of 17- to 21-year-old males by race and Spanish origin (developed by cohort component demographic techniques) were obtained from the Bureau of the Census. Second, race-specific rates for high school graduation, AFQT category

[3]This section is largely taken from Army Research Institute (1987).

distribution, and enlistment propensity were computed at the Army brigade level (five geographic regions) from several sources, mainly the NLS Profile of American Youth Sample. Third, the brigade rates were applied to all battalions in the brigade. (Table 3 lists the terminology used in the various services' recruiting network areas.)

No attempts were made to apply the rate data to the county level. Small sample size prohibited race-specific estimation at the battalion level and questionable propensity estimates at the brigade level. Some questions were raised about the use of the Current Population Survey (but it could not be linked with test scores) and whether regression analysis had been used to predict propensity. Note was also made of the existence of ASVAB test scores administered in many high schools throughout the country; however,

TABLE 3: Recruiting Networks of the Service Branches

Branch of Service	Hierarchical Geographic Area	Number of Areas
Army	Brigade	5
	Battalions	56
	Recruiting company	262*
	Recruiting station	2,500*
Navy	Recruiting area	6
	Recruiting district	41
	Recruiting zone	300*
	Recruiting station	3,000*
Air Force	Recruiting group	5
	Recruiting squadron	35
	Recruiting flights	211
	Recruiting office	1,250*
Marine Corps	Recruiting region	2
	Recruiting district	6
	Recruiting station	48
	Recruiting substation	579

*Estimate

Source: Information from the Recruiting Commands of the individual services.

the test is voluntary and could be biased because of its selective nature. Thus, there is a reliance on NLS/Profile of American Youth samples. There was also some mention of possible use of the High School and Beyond survey to equate to or model aptitude (mental performance), but no work had been done at this point on this possibility. It also was pointed out that the High School and Beyond universe does not include all age groups within the 17- to 21-year-old recruiting range.

The Army survey methodology related only to estimates of mental aptitude categories. No attempt was made at estimating the physical standard or moral standard components.

Alternate Methodology

Workshop participants also discussed alternate approaches to estimating QMA. An evaluation of the QMA methodology prepared for the Marine Corps was presented, as well as a proposed alternative. Some evidence was presented on the relationship of current QMA and aptitude scores for selected areas--independent cities and counties--for a number of states. Statewide test scores used to measure county-level aptitude were also available for a number of areas. The relationships were illustrated by scatter diagrams that on balance did not provide clear evidence of a strong relationship between the two variables. It was noted, for example, that in the diagrams the horizontal axis was model-based, and the model used for category 1 through 111a QMA put most counties and other entities at the 50th percentile.

A main thrust of the discussion was that estimates of QMA, by themselves, do not provide an adequate basis for targeting and recruiting; simultaneous estimates of the propensity to enlist are required, and a procedure for developing this factor was proposed. A simple definition or criterion, in this instance, was stated as the ratio of the number of accessions in a given area for a specified period to the current estimate of QMA. In the example cited, this ratio ranged from 1 to 10 percent, presumably reflecting local differences in enlistment propensity. However, it was not clear from the discussion whether such a propensity measure merely reflects differential recruiting efforts (such as differences in number of recruiters). One would need to assume constant recruitment efforts (among other factors) if the differences are to be ascribed to variations in propensity and county level recruiting efforts cannot be known. There was concern expressed as to whether propensity can be clearly defined and measured, particularly at the local level. Some participants expressed the view that the measure of propensity is too vague and elusive to be used effectively to modify QMA distributions or to be helpful in setting recruitment goals. One question concerned whether accessions merely represented the assigned quotas, and it was proposed that measures of recruitment effort are needed to reflect the real differences in local propensity. The point was made that it was not necessary to directly measure propensity, which in any case cannot be done.

However, propensity matters a great deal and must somehow be dealt with. One suggested option was to do this indirectly as part of a unified measure of QMA and propensity.

It was suggested that a propensity-weighted QMA was needed to provide an improved QMA estimate. The proposed formula in effect computes the new propensity weighted QMA for an area by assigning to the local area the percentage of all male high school graduates in grade 1 through 11la enlistments of the recent past recruited in that area. Simply put, a single year's enlistments (in all services) best provides a measure for setting recruiting goals for the second year. For small areas an average of the past three years is a better predictor, but it may need to be supplemented with another variable such as employment. For purposes of allocating recruiters and establishing recruiting goals (the main function of QMA), propensity-weighted QMA, supplemented by the local commander's judgment, may be a suitable basis.

Further discussion on the development and uses of QMA reinforced the notion that past experience on the production of accessions was very useful in allocating current quotas. Reference was made to the use of multiple regression techniques to model past production for recruitment goal allocation. The number of high school seniors (and/or graduates) is an important data source for recruiters at local levels. Use is also made of all Department of Defense testing using the Armed Services Vocational Aptitude Battery. It was noted that current estimates of QMA may be used to allocate goals at higher levels of geography but at lower organizational levels, and that the methodology used to distribute the goals to recruiters is left to the discretion of the respective field units. Factors used include size of the qualified market, number of recruiters available, and various exogenous economic factors such as unemployment rates or per capita income. Propensity measures developed from the Youth Attitude Tracking Survey are used in some cases but not others. (YATS asks about interest in the military, and in that respect its use even at larger geographic levels is limited because an indicator of actual intent is needed.) Mention was made that propensity at the local level was estimated by regression techniques using local unemployment rates and previous recruiting records for the area. In fact, past production was mentioned by several speakers as a preferred basis for allocating recruiters and establishing goals.

Data Sources

Workshop participants also heard a description of one of the main data sources available to recruiters. The system, known as the Recruit Market Network System, was established by DMDC and provides data support for all branches of the services. The system is designed as an on-line source of general-purpose data sets at the county level. Data potentially useful for recruitment purposes are obtained from a variety of sources, private and public, and entered into the system. Examples of the data sets found in the system are:

time series of accessions, demographic (census) data, education and enrollment data, and test scores. Some concern was expressed about the lack of knowledge about the reliability of the individual data sets. The extent of use of these data sets was not known, and there was strong agreement that monitoring be undertaken to learn about the specific data used and frequency of use. There was also concern expressed about relevance, i.e., does the system contain the right data and is the interpretation clear? A reexamination of the content would assist in data selection for future years.

SUMMARY

The picture emerging from the discussions was that of a diversity of interests and practices among the various services in the way they go about the business of recruiting and of setting basic recruiting goals. QMA as such is not always used directly in establishing recruiting goals but is one piece of information in the armamentarium of recruiting and recruiters. And it was not clear that any one set of data could uniformly and systematically meet the needs of the enlisting service branches. Indeed, some of the questions and concerns that emerged to guide the small group discussions reflected these diversities and uncertainties:

- Would the military services really use QMA estimates at the county level even if a statistically sound method was used to prepare them and acceptable levels of accuracy were achieved? (One initial reaction was that they would be looked at but not really used, since most local recruiting is done at the high school level.)

- Which is the better concept to use in the estimation process, QMA or QMI? (The latter, Qualified Military Available and Interested, incorporates a weighting of the concept of propensity with QMA.) Can propensity be clearly defined? Is the concept understood, and can it be quantified locally?

- What is known about accepted level of error relative to the uses of the data? What accuracy or reliability is required?

- Is the main interest at the county level, or are counties useful primarily as building blocks to accommodate the different geographical structures of the various services' recruiting networks?

- Can we assume increased accuracy at the higher, aggregated levels, regardless of method?

- Is it feasible to administer ASVAB at local levels, or to somehow make use of the variety of educational data available at the local school level, e.g., Scholastic Aptitude Test scores, the National Assessment of Educational Progress, and the National Education Longitudinal Survey?

- Since present data sets are inadequate for measuring or estimating moral and physical standards locally, can something be done to incorporate these elements into the estimates? Can we consider a simplistic approach such as selecting one set of estimates, arbitrarily chosen, from existing standardized data sets and recommending it for use by all service branches?

- Is a demographic cohort approach useful for providing overall figures on population size for appropriate age groups?

SUMMARY OF SUBGROUP DISCUSSIONS

Methods

The group began by trying to define what population was of interest, could be measured and whose estimate would be used. There was recognition that QMI (Qualified Military Available and Interested) estimates are desired, but general agreement that this could not be measured or even clearly defined. Reliable (or perceived as such) estimates of QMA at the county level would be useful to recruiting organizations. With existing data, however, not all aspects of QMA can be estimated at the county-level with results that truly capture local factors and variations.

With existing data, the focus should be on estimates for males by age (17-21 or other), by race/ethnicity, who are high school graduates, and by AFQT score category. If possible, the full distribution of percentile groupings should be made available to permit reclassification as standards change. Since subnational data on physical or moral qualifications do not exist, the consensus is to start without these elements and concentrate on the mental or aptitude component of QMA. With regard to propensity, there was no consensus on what it is or how it might be defined and measured. It appears to be subject to significant change over time and able to be manipulated by the services. It may actually be the result of recruiting efforts, not an input to it.

Estimates of this population (by AFQT score or equivalent measure) can probably be produced for large areas and populations. Such estimates should be useful for recruiting organizations for various purposes, such as resource allocation. Some consensus should be reached on what areas should be estimated, e.g., all counties of 100,000 or more and groups of counties. Some effort should be made to define a set of "common denominator" areas across the four services.

Estimates should not be produced for very small areas, since poor estimates for small counties (areas) may be misleading and could call into question the credibility of estimates for larger areas (although erroneously). In addition, accuracy of methods would seem to be a function of the amount of money available to spend on methodology. For increased accuracy, one needs to measure the relationship between increases in resource input and increases

in accuracy. At present, the services do not depend on QMA estimates by county. Resource allocation is determined by local knowledge. Personal factors and local data (e.g., high school graduate data) gathered by recruiters is almost certainly more valid for local recruitment purposes and is used extensively by recruiters.

The regression sample-data method should be investigated as a means for measuring local AFQT distributions. Empirical Bayes methodology similar to that used by Fay and Herriot (1979) for small-area income estimates was specifically proposed (see the appendix).

To the extent feasible, future surveys or data collection efforts undertaken by the Department of Defense should be designed to measure physical and moral qualifications as well as mental qualifications.

Ongoing federal surveys should be investigated to explore the feasibility of collecting data on aspects of QMA, especially those concerned with mental qualifications. The particular usefulness of the National Assessment of Educational Progress and the National Education Longitudinal Survey was especially noted. Such efforts will become even more important and useful if the proposed modeling method proves viable.

Data Needs

The discussion started with two broad questions. What are the best data sources to estimate the QMA at the smallest-area level? What methods can be used to get better estimates from available data sources?

While one major reason for seeking the best estimates concerns the allocation of resources, allocating recruiters, and setting local quotas, it is clear that a second consideration, predicting success once in the service, is also a major concern. In this connection, it is necessary to get as accurate an estimate of mental ability as possible. Each of the services perceives the AFVAB as an excellent test for predicting scores in specific service occupations. The question is whether the large number of tests available (2 million annually on a nonrandom basis) can be used to identify those areas with the best pool of available persons.

The large number of tests are available because they are given by self-selected high schools throughout the country. Recruiters use them as a way to gain entry to the schools for recruitment purposes. Some schools use the test, others do not. Some of those who do use it make it mandatory, others voluntary; some give it to seniors only, others to two or three grades. The main virtue is the sheer numbers of data; however, lack of representative sampling makes it difficult to use for small-area estimates of an eligibility pool of youth.

Participants agreed that it would be desirable to set aside some part of the annual appropriation for this effort for experimentation purposes. One possibility would be to analyze areas with complete

coverage (the state of Florida is administering the test to all high school students in 1989) to see how the scores there compare with similar demographic areas with nonrandom test coverage.

This group also agreed that present data sets are inadequate for estimating moral and/or physical standards for local areas.

It was agreed that census data, specifically Summary Tape File 4 (STF4), provide the best starting point for small-area data and that cohort survival techniques for estimating small-area populations, which are now used, are valuable. They can provide good detail on education by age and sex down to the census tract level. It was suggested that subcounty densely settled areas, such as New York, be examined. It was noted that there is considerable detail available on educational status by labor force status for 16- to 21-year-olds, important data for identifying those who might consider the military an option. This would need to be updated with current unemployment rates. Note should be taken of areas with large numbers of college students or institutions such as prisons that are not targeted by recruiters.

The group strongly urged that existing surveys and data sets be explored as sources of additional data that might be used for model-based updating of small-area estimates. These include information on public enrollment and the number of high school graduates for high schools and school districts (from the National Center for Education Statistics); another Profile of American Youth, based on the National Longitudinal Survey or similar surveys including the possibility of an add-on to the Current Population Survey, which already collects data on education (the CPS probably could not be used to administer the AFVAB, but it might be helpful in developing proxy measures); the use of selective service cards; and the use of data for some larger areas--standard metropolitan statistical areas, for example.

CONCLUDING THOUGHTS OF THE CHAIR

The workshop discussion highlighted the various types of data needed and used at different stages and different levels of the military recruitment network. Different purposes require a focus on different measures: targeting efforts, evaluating efforts, and assessing potential for emergency mobilization.

The data needs of an organization (or part thereof) depend on its mission. From the Department of Defense through the individual services, from the regional divisions of the services down to local recruiting offices, the mission changes at each stage, and so do the needs for data.

Defense Manpower Data Center

The Department of Defense is the most centralized group, with concern for all the services as well as the entire country. However, it does not actually recruit anyone. Through its mandate, the Defense Manpower Data Center assists in providing the different services with data and processing support with which to meet their individual recruitment goals, both in terms of quality and quantity, while spending as few resources as possible. DMDC data provide the basis for medium- to long-term planning, monitoring the supply of recruits, estimating future supply, estimating the elasticity of supply to recruitment incentives, and being able, if necessary, to estimate supply under circumstances of general mobilization.

DMDC also plays a role in contributing to the efficiency of the recruiting process. Efficiency can be visualized as having two components, evaluation and improvement. Ideally, evaluation would consist of estimating some ratio such as the number of actual recruits to the number of recruits who could have been recruited given the resources used in the recruitment process. In theory, this latter number can be regarded as the qualified personnel available multiplied by a potential propensity to enlist. However, this propensity can be expected to vary markedly by small areas, for which the qualified personnel cannot readily be measured, so formal

evaluation of this type is impossible. Thus DMDC, for its own use, does not require small-area data for evaluation purposes. Nor, in all probability, does DMDC need to provide information below the broad service recruiting area level to help in the recruitment effort (as opposed to assisting in establishing recruiting goals). As was outlined in the discussion, local recruitment processes depend to a large extent on an intimate knowledge of local high schools and local conditions that can never be estimated centrally with any degree of accuracy.

The Department of Defense (and the individual services) also require information for use in medium- and long-term planning, consisting of estimating how many recruits will be needed and the pool available from which the services can draw in future years. This process requires estimates of QMA and of propensity to enlist, but it does not need to be carried down to the small-area level (though any exercise to estimate the propensity to enlist would probably need to use small-area data). Thus, it seems that for most objectives, small-area data are not required, except for developing estimates of the propensity to enlist.

Furthermore, as noted in the discussion, the development of more useful estimates of QMA requires the ability to better reflect its constituent elements. For example, optimally, one should have available a national distribution of ASVAB/AFQT scores or, failing that, the ability to convert either a wide variety of existing and different measures or some common characteristics as are available into comparable ASVAB/AFQT scores or terciles. Similarly, national measures of the health of youth are required in order to develop estimates of those who will not meet the medical standards established for potential recruits. Finally, some way must be found to estimate the number of youth who will fall short of the required moral standard. The spread of drug activity in young age groups and the heightened concern of the possible incidence of AIDS and the HIV virus in the recruitable population make these latter points even more imperative and compelling.

From the perspective of the DMDC and its interest in the workshop, one might summarize the issue as follows: the purpose of providing estimates of QMA for local areas is to assist regional and national commands in setting recruitment goals across all areas, using a standard set of reliable, high-quality data uniformly developed and consistently applied. Local-area data, such as the number of high school seniors or graduates (and perhaps even propensity attributes), can be used to modify these externally set goals to more realistically reflect local situations. Recruiting efforts can then proceed in a variety of ways to target their populations with a better understanding of the market.

Individual Services--Central and Regional

The individual services are responsible for their own recruitment efforts and could use local-area data for a number of purposes. First, resources in terms of personnel and advertising

could be allocated to maximize recruiting. For this purpose, the important measures for a local area would be the QMA and the elasticity of the propensity to enlist with respect to resources, that is, the extent to which propensity to enlist varies with variations in recruiting efforts. Since the propensity to enlist itself cannot be readily measured, its elasticity cannot readily be measured either.

Second, the efforts of local recruiting officers could be evaluated if the potential number of recruits could be estimated. For this purpose, the propensity to enlist and the QMA are required for small areas. Though it might be possible to estimate the propensity to enlist for small areas, by relating recruiting differentials to independent factors, such as the local unemployment rate, regional dummy variables and tertiary education levels, the QMA itself cannot at present be estimated with any precision for small areas. However, if new methods are developed allowing QMA to be estimated synthetically for small areas, then it would be worth trying to estimate propensity to enlist synthetically for such small areas also, in order to evaluate performance.

Individual Services--Local Areas

It would seem plausible that local-area data would be of most use at the local level, in helping an individual recruiter track down recruits. However, the primary focus of a local recruiter is the graduating classes of a limited number of local high schools, with a secondary focus on the classes likely to graduate in the next couple of years. The knowledge of a conscientious recruiter about these groups in terms of their numbers, likelihood to qualify, and propensity to enlist would be greater than could be provided by any small-area synthetic estimation system. Thus, a good recruiter doing a good job has no need of small-area estimates handed down from above.

Given today's situation, it is our conclusion that some improvements are feasible in the estimation of QMA, and we have set forth several recommendations for these. Significant improvement, however, will only be forthcoming through the development and implementation of large, new, and expensive data collection efforts. Whether such actions are merited, we leave to others; within our scope of competence, we strongly urge continued research into the methodology of QMA, investigation into and exploration of the feasibility of using other existing surveys and sources of data, and an updating of the Profile of American Youth survey. We also urge a continuing examination and evaluation of the many different data sources now used by the individual services, in order to determine their strengths and weaknesses and lead to a standardization of and understanding of what is best utilized. Finally, we would urge an effort at educating all recruiting levels of all the services about the strengths and weaknesses of the data they are using.

With regard to tests already being administered in various high schools, these are used primarily as recruiting tools that generate

lists of names (prospects) for military recruitment. Administering the test to a select sample would only reduce its utility and purpose; thus the services are not likely to change their current practice. However, in light of its potential for developing reliable estimates of QMA at the local level, a reexamination of this allocation of resources might be in order. Other possibilities for estimating mental categories include the use of Scholastic Aptitude Test scores (the Air Force uses SAT scores for officer recruiting only), Educational Testing Service data tapes for linking, use of the National Assessment of Educational Progress, and the National Education Longitudinal Survey.

REFERENCES

Army Research Institute
1987 Projection of the Male Youth Population and Enlistment Propensity, by Army Recruiting Battalion, 1982-1985. Prepared for the Army by Naomi Verdugo and Ray D. Nord. June.

Defense Manpower Data Center
1987 The Medical Fitness of American Youth for Military Service. Defense Manpower Data Center, Arlington, Va.

Fay, R.E., and R. Herriott
1979 Estimate of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association* 74:269-277.

Frankel, Martin, Harold A. McWilliam, and Bruce D. Spencer
1987 Technical Sampling Report, National Longitudinal Survey of Labor Force Behavior, Youth Survey (NLS). August.

Navy Personnel Research and Development Center
1987 Estimating the Youth Population Qualified for Military Service. Ervin W. Curtis et al. August. Personnel Research and Development Center, San Diego, Calif.

Polick, J. Michael, James V. Dertouzos, and S. James Press
1986 The Enlistment Bonus Experiment. Prepared for the Office of the Assistant Secretary of Defense (Force Management and Personnel). Rand Corporation, Santa Monica, Calif.

Research Triangle Institute
1987 Youth Attitude Tracking Study. Wave 17, Fall 1986 report. Prepared for the Defense Manpower Data Center. Research Triangle Institute, Research Triangle Park, N.C.

U.S. Department of Defense
1982 Profile of American Youth: 1980 National Administration of the Armed Services Vocational Aptitude Battery. Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics). March.

1984 Screening for Service: Aptitude and Education Criteria for Military Entry. Mark J. Eitelberg et al., Human Resources Research Organization, Office of the Assistant Secretary of Defense (Manpower, Installation, and Logistics). September.

1988 Manpower for Military Occupation. Mark J. Eitelberg, Office of the Assistant Secretary of Defense (Force Management and Personnel). April.

APPENDIX: THE USE OF EMPIRICAL BAYES REGRESSION MODELS IN ESTIMATING PARAMETERS OF THE DISTRIBUTION OF AFQT TEST SCORES FOR SMALL AREAS

Michael L. Cohen

INTRODUCTION

Empirical Bayes regression modeling is an area of statistics that was developed primarily within the past 10 years. A number of situations have been discovered for which this methodology is considered to be superior to more traditional approaches. The modeling of Armed Forces Qualifying Test (AFQT) scores is possibly another area in which empirical Bayes regression modeling might prove to be efficacious. The following discussion has two purposes: it is meant as a quick introduction to the basic advantages of empirical Bayes regression models to those unfamiliar with this methodology, and it contains recommendations to those statisticians interested in implementing the above methods for predicting small-area estimates of Qualified Military Available (QMA). Finally, some relevant references are included in the text.

To begin, assume that we have a data set of scores attained on the AFQT, collected at the county level on a sample basis. The term *county* as used here is a general term, by which we might mean aggregations of counties when the counties in question are very small and the aggregation is not subdivided by any of the services; we also might mean a splitting up of the counties into pieces for reassembling into a like number of areas that are more homogeneous (say by urban-rural factors) than counties. This does not prevent the formation, after the fact, of actual county-level estimates. For example, if a number of small, similar counties are aggregated for statistical modeling purposes, the resulting estimates can certainly be applied to each county separately. This flexibility for forming homogeneous units for purposes of analysis is potentially of substantial benefit in model development in general, and empirical Bayes regression models in particular.

In developing a model for AFQT scores, we actually have several choices as to the appropriate dependent variable to use. We could take as the dependent variable the median AFQT score in each county, or, alternatively, any percentile. Furthermore, we could take a multiple logit approach, which would attempt to simultaneously

---

Michael L. Cohen, Assistant Professor, University of Maryland, served as consultant to the workshop.

estimate several percentiles. Another possible dependent variable that would be nearly as useful as median AFQT score, and possibly easier to model, would be the percentage of students taking the exam who scored above a certain level. (This latter dependent variable might be beneficially transformed, since it is a frequency and thus likely to have heterogeneous variances.) This decision will have to be made after data collection. Whatever the choice, we denote the dependent variable Y_i, for county i. So, given one above interpretation, Y_i is the median AFQT score for county i from the sample.

We assume that we have also collected additional covariates for the same counties. These might have come from the same survey, from a more complete survey, such as the decennial census, or from administrative records.[1] The median income level for the county and percentage college-educated in the county are suggested as possible variables to use in developing a model of Y_i. Some other possibilities include performance on other standardized tests or administrative data from schools in the county. Recruiters from the services could be asked to assist in identifying additional covariates that would be effective in modeling AFQT. Before defining empirical Bayes regression models, we describe two alternative approaches to modeling AFQT scores.

Synthetic Estimates

If Y_i is considered to be strongly related to certain categorical variables, then one possible model-based estimate of Y_i is the synthetic estimate. With synthetic estimation, one weights the mean of the dependent variable collected at a higher level of geographic aggregation (possibly nationally), for various groups defined by the categorical variables, by each group's percentage of total population in that county. Synthetic estimation has the advantage over other more sophisticated model-based estimates of producing estimates that are generally subject to much smaller sampling errors than other competitors. This advantage is often offset by the frequently higher systematic error (bias) for synthetic estimates.

---

[1]When covariates come from a very much larger sample survey, such as the sample items included in the decennial censuses, the effect of their own sampling errors may be ignored, for practical purposes, and treated in the same form as complete (non-sample) data from an administrative source. When the covariates come from the same sample survey or another sample survey of comparable or smaller size, however, it is generally advantageous to incorporate explicitly the effect of sampling error on the covariates. The discussion of models for this situation is not covered in this appendix but may be found in a number of other papers, including Fuller and Harter (1987) and Fay (1988).

Regression Estimates

If certain covariates related to socioeconomic status, educational level of parents, etc., are found to be effective in predicting Y_i, then one could regress the Y_i on these covariates to produce model-based estimates, denoted θ_i, where $\theta_i = b_0 + b_1X_{1i} + b_2X_{2i} + \ldots$ for some covariates X_1, X_2, ... and associated regression coefficients b_0, b_1, b_2, ... These are not completely standard regression models because the variability of the Y_i has two components, the variance of the model error of Y_i, denoted τ_i^2, and the sampling variance, denoted σ_i^2. In order to proceed, most researchers have made the further assumption that the model error is homoscedastic, that is, $\tau_i^2 = \tau^2$ for all i. Thus the usual mean-square error of this regression is not simply the usual estimate of lack of fit, but is also affected by the sampling variances.

Combination of Regression Estimate and Direct Estimate

For any given county, we can consider using at least two separate estimates of the median AFQT score. A first estimate is the median AFQT score from the scores for that particular county, which we call the direct estimate, denoted Y_i. This estimate is typically unbiased, since it is collected from data for only the county in question. However, since we will often have relatively little data from any particular county, this estimate also has, typically, substantial variability.

We also have the regression estimate or the synthetic estimate, denoted θ_i. This estimate typically is biased since it involves a model making use of data from other counties; depending on the utility of the model, this bias can be negligible or considerable. However, since many more data points are used in estimating the parameters of this model, it is quite possible that the variability of θ_i is greatly reduced from that for the direct estimate, Y_i.

It seems reasonable that these two estimates could be combined in a way that minimized the weaknesses of either estimate. Thus, if the sample was large enough, one would essentially use the direct estimate and, if the model were good enough, one would essentially use the model-based regression estimate. This is the fundamental idea behind empirical Bayes regression modeling. Specifically, these two estimates are combined as follows:

$$\hat{\theta}_i = (1-w_i)\, Y_i + w_i\theta_i, \quad \text{where } w_i = \sigma_i^2/(\sigma_i^2+\tau^2),$$

to form an estimate that, under fairly broad conditions, should perform equal to or better than either of the component estimates. When the weights, w_i, are estimated from the data, the resulting estimate is called the empirical Bayes regression estimate.

The reason for the precise form of w_i above derives from the situation in which one has two unbiased estimates of the same parameter with different variances. The optimal estimate in this situation is the linear combination of the two estimates, wherein each estimate is weighted by the other estimate's variance. This underlies the interpretation given to these methods by Kackar and Harville (1984), and is called the components of variance model. In our example, we have to incorporate the bias into the estimated variability. This is done by replacing variance everywhere by mean-square error.

The use of the above weighted combination also can be justified through a Bayesian approach. This is the paradigm used by most of the researchers in empirical Bayes regression modeling and is therefore relevant. We begin by assuming that we have a process that generates means μ_i such that $\mu_i \sim \Phi(\mu,\tau^2)$, where $\Phi(\mu,\tau^2)$ represents a normal distribution with mean μ and variance τ^2. Also, suppose that $Y_i \sim \Phi(\mu_i,\sigma^2)$. Unconditionally, the $Y_i \sim \Phi(\mu,\sigma^2+\tau^2)$. The problem is to estimate the value of μ_i after observing Y_i, and knowing the value of μ. Basically, the Y_i provide specific information about the μ_i, and μ provides global information about the μ_i. Not surprisingly the optimal estimate uses both to estimate the μ_i, with the direct estimate Y_i weighted more heavily when σ^2 is relatively low, and μ weighted more heavily when τ^2 is relatively low. Bayesian methods direct one to determine the posterior distribution of μ_i given Y_i, and then to use the mean of this distribution, called the posterior mean. That turns out to be:

$$E[\mu_i \mid Y_i] = (\tau^2 Y_i + \sigma^2 \mu) / (\tau^2+\sigma^2). \qquad (1)$$

This is a weighted combination of Y_i and μ, with weights $\tau^2/(\tau^2+\sigma^2)$ and $\sigma^2/(\tau^2+\sigma^2)$, respectively. Unfortunately, neither σ^2, nor τ^2, nor μ is known. Empirical Bayes methods estimate these three parameters from the marginal distribution of the data, and then substitute the estimates into equation (1) above. In the case that we are examining, σ^2 is known but is not constant across counties. Thus we have σ_i^2, instead of σ^2. In addition, we do not have a mean μ independent of the counties. Rather, we have a model-based estimate of the mean for each county. Both of these complications were addressed in Fay and Herriot (1979). So we now have $\mu_i \sim \Phi(X_iB, \tau^2)$, where X_iB serves in place of μ (and an estimate of B is needed), and $Y_i \sim \Phi(\mu_i, \sigma_i^2)$ with σ_i^2 observed. The remaining problem is to estimate τ^2. τ^2 enters into the mean-square error of the regression analysis. Carter and Rolph (1974), Morris (1983, related in Fay and Herriot, 1979), and Morris (1983) have approaches (method of moments and maximum likelihood) to this problem.

EVALUATION

Past Experience

There is a lively debate about the efficacy of these models. Often, the variances (σ_i^2 and τ^2) are so disparate that one ends up using either the model-based estimate or the direct estimate, with little or no averaging. (However, that is still better than either estimate.) Clearly, this depends on the sampling plan used. For areas with a large sampling rate, the empirical Bayes regression estimate will be closer to the direct estimate Y_i. For areas with a small sampling rate, the empirical Bayes regression estimate will be closer to the regression estimate.

Rubin (1980) used this methodology with law school data, to little gain. J.N.K. Rao related little improvement in practical situations when using these techniques. Results of little improvement, when they occur, provide only slight improvement above the single estimator Y_i (when $\sigma_i^2/(\sigma_i^2+\tau^2)$ is small) or over the regression estimator θ_i (when $\sigma_i^2/(\sigma_i^2+\tau^2)$ is large). There is nothing wrong with the empirical Bayes approach in such cases (provided a reasonable model is specified), but it differs only slightly from the simpler, standard estimates, and therefore does not justify the additional cost.

By contrast, Fay and Herriot (1979) reported substantial gains. In the adjustment arena, the 1986 Test of Adjustment Related Operations made use of this methodology with a gain that was believed to be around 33 percent (see Diffendal, 1988, for details). Similar gains were seen by Ericksen and Kadane (1985) in the same problem with respect to the 1980 Post Enumeration Program. In general, the gains from combining the direct estimate and the model-based estimate are greatest when the two sources are of roughly comparable accuracy, so that information from both sources can be used to advantage. When either is far more accurate than the other, however, little or no advantage results from combining the estimates instead of using the better of the two alone. Thus, the extent to which the application of this technology will be effective for this problem is not yet entirely clear.

One problem that might arise is if the error variance for the regression model, τ^2, is county-dependent, i.e., if the error variance for county i is equal to τ_i^2. This would be a difficult problem to overcome since there is at present no methodology available to estimate τ_i^2. In principle, however, one can model this effect, but new work would be required.

A Test of the Empirical Bayes Methodology

The workshop was advised that the State of Florida intends to administer the AFQT to all male high school seniors during the

school year 1988-1989. It was agreed that the availability of such a data set would present an unusual opportunity to test out the empirical Bayes methodology. The following plan, therefore, is proposed for the evaluation:

1. Develop a national model of AFQT scores from the NLS/1980 profile,[2] possibly augmented by census data, or administrative data, but only by data sources uniformly available at the county level.

2. Collect 100 percent information on the AFQT distribution for the various counties in Florida.

3. Develop model-based estimates of AFQT, for which the model-based estimates take the parameters from the nationally determined model and plug in covariate information particular to counties in Florida.

4. Compare several direct estimates that would result from various sampling rates and plans (see also point six below), the model-based estimate, and also empirical Bayes compromise estimates all against the truth, which is known since we have virtually complete information for all counties. This raises at least two separate issues. The first question is whether the model-based estimates are effective predictors of median AFQT scores. This would be accomplished by directly comparing the model-based estimates against the full administration of the Florida AFQT scores. The second question is whether the final estimate is markedly improved when the model-based estimates are combined with direct AFQT information. We point out that these data would let us make rather direct analyses to determine proper model specification, the necessary functional forms, tests of homoscedasticity, etc.

5. For the second question: First, given a scientific sample, does the empirical Bayes regression estimate substantially improve on the model-based estimate? In other words, given a sampled data set in which there is no selection bias, how much improvement is obtained from combining the direct estimate with the model-based estimate? Any sampling procedure can be simulated on the 100 percent data set, and the resulting empirical Bayes estimate compared with the truth. For example, for any modification of the sampling scheme used in the NLS, the resulting empirical Bayes regression estimate can be evaluated. (In addition, other proposed estimates, which are not directly related to the empirical Bayes regression estimate, the model-based estimate, or the direct estimate can and should be evaluated.)

6. Second, it is of interest to extend this analysis to areas in which the administration of AFQT is not a representative sample of test takers. AFQT is currently administered to about 2,000,000 high school students on a haphazard basis. The question is to

---

[2]The Department of Labor's National Longitudinal Survey (NLS) of Youth Labor Force Behavior; in 1980, the Profile of American Youth, i.e., a nationwide administration of the Armed Services Vocational Battery (ASVAB), was added to the data collected in the NLS.

determine how best to utilize the large number of administrations of the AFQT to improve the model-based estimates, possibly through the use of empirical Bayes methodology. More specifically, does the nonrepresentative nature of the administration of AFQT prevent use of this resource? One way of attacking this problem is to simulate various selection-biases along with the various sampling plans with the 100 percent data from Florida and gauge the robustness of the various estimates to various forms of selection bias.

How would one define Y_i outside Florida? In the absence of selection bias, one would simply use the sample data at some level of geographic aggregation. The difficulty would arise for areas in which the test had been administered in an uncontrolled manner. Here there are some approaches that might be beneficial. To begin, the reasons for less than 100 percent administration would have to be understood. If the main reason for less than 100 percent administration was lack of interest on the part of the students, grade point averages, etc., could be used to relate the performance of the test takers to the test avoiders. Then the problem is identifying the percentiles of a mixture distribution, which has been addressed in the literature. (This problem would require a fairly sophisticated methodology.)

7. It is important to develop surrogates for the AFQT, by calibrating other intelligence tests to the AFQT scale. Therefore, it would be important to link NLS88 information, and possibly the National Merit Scholarship Qualifying Test (NMSQT), the Scholastic Aptitude Test (SAT), Minnesota's test, the Iowa test, etc., to the AFQT, along with a determination of the correlations that exist between these various tests. It would also be beneficial to not only link the tests that would ordinarily be administered in Florida during the subsequent year, but to take the opportunity that Florida provides to consider the administration of additional tests for possibly larger sample sizes.

8. Depending on the results, consider using empirical Bayes methodology throughout the nation.

At this point, there would be benefits gained from comparing the model's estimates with the current QMA in Florida, by county, as well as with the percent of military applicants with AFQT scores above the median for the last three years.

Final Points Concerning the Test

There are some difficulties that must be considered in embarking on this course.

1. *Florida is statistically unusual.* Florida has a larger number of Hispanics than the country overall, and the state's economy is unique. The state population has a peculiar age distribution that results in covariates with unusual values, e.g., socioeconomic status information. All these considerations are cause for concern as to the generalizability of the Florida experience. One way of overcoming this problem is to search for counties outside Florida where the AFQT was administered on a 100

experience. One way of overcoming this problem is to search for counties outside Florida where the AFQT was administered on a 100 percent basis and then determine whether the model developed was also effective in those areas.

2. *Variables.* We still are not sure that the appropriate variables have been identified. Possibly High School and Beyond or NLS data could be used to help determine which variables would be useful in modeling AFQT scores.

3. *Computer programming.* The algorithm to estimate τ^2 is a nontrivial search procedure. One of the problems encountered is that there is some chance of estimating a negative τ^2. Before recommending an estimate, we should be clear as to how the technology for computing the estimate can be acquired. There now exist fairly robust, simple but iterative ways to estimate τ^2. Morris (1983) contains one such example, for known variances σ_i^2.

4. *Raking or iterative proportional fitting.* At a final stage, the distribution of AFQT for counties might be altered so that at a more aggregated or national level the answers would be identical to the national distribution. This procedure is also fairly sophisticated algorithmically. The discrepancy between the aggregated estimates and the national totals is also a measure of goodness-of-fit of the models used and should be studied with this in mind.

5. *Model heterogeneity.* There is the question of whether one would use one model, or whether one could develop several models, say 10 or so, and then applying one to Florida. Expecting one national model to be generally applicable, however, may be expecting too much. Possibly there is a model that would work in the Southeast, whereas another would be more effective in the Midwest, etc. This should be investigated.

6. *Variance estimation.* Each of the estimates described in the above discussion has an associated parametric variance estimate, based on various assumptions. In performing an evaluation of the various estimates, it would also be worthwhile to empirically determine the validity of these estimates of variability. These variance estimates, if validated, could be used to help set reasonable recruitment goals for each county through the construction and use of lower confidence bounds. This is especially valuable for the empirical Bayes regression estimate, since the error in estimating τ^2 can have an appreciable impact on the estimation of its variance (see Rao and Prasad, 1987).

The estimation of the prediction error of empirical Bayes regression models is an area of current research. Substituting the estimate of τ^2 and σ_i^2 into the formula for the variance of the posterior mean results in estimates that are too small (see Morris, 1983, for details). Another way to see this problem is to note that the prediction error for both the direct estimate and the model-based estimate is relatively easy to compute. However, the weights are also data-dependent and have variability that is sometimes ignored. There are several ways to address this difficulty. One

possibility is to use cross-validation, whereby a data point is not used to develop parameters that are used to estimate the prediction for that same data point. Morris (1983) presents a Bayesian approach to this problem.

7. As in Freedman and Navidi (1986), the assumptions underlying the empirical Bayes regression estimates need to be made explicit and shown to be appropriate. These include the homogeneity of the error variance and the linearity of the regression model. Sensitivity of the empirical Bayes regression estimate to the choice of covariates should also be studied. One way of studying the validity of these assumptions, as well as gain information about the variance mentioned in (6) above, is to use cross-validation.

8. The investigation of the possibility of modeling the heterogeneity of error variance, i.e., the τ_i^2, could be carried out in Florida. This would remove the necessity of assuming homogeneous error variances.

9. The above effort has concentrated on models developed at or near the county level. The opportunity that Florida provides can als be utilized to investigate the efficacy of models developed at lower levels of aggregation. These more micro models might have the advantage of having more heterogeneous situations, allowing better parameter estimation. These micro models could have both individual-level variables as well as county-level variables.

10. Finally, it will not be clear when this study is completed precisely how often the model will need to be respecified. This would entail several experiments of the sort envisioned for 1989.

REFERENCES

Carter, Grace M., and Rolph, John E.
1974 Empirical Bayes methods applied to estimating fire alarm possibilities. *Journal of the American Statistical Association* 69:880-885.

Diffendal, Gregg
1988 The 1986 test of adjustment related operations in Central Los Angeles County. *Survey Methodology* 14:71-86.

Ericksen, Eugene P., and Kadane, Joseph B.
1985 Estimating the population in a census year - 1980 and beyond. *Journal of the American Statistical Association* 80:98-109.

Fay, Robert E.
1988 Empirical Bayes Estimation for Multiple Characteristics. Paper presented at the annual meetings of the American Statistical Association, New Orleans, August 22-25, 1988.

Fay, Robert E., III, and Herriot, Roger
1979 Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association* 74(366):Part I, 261-277.

Freedman, David, and Navidi, W.C.
1986 Regression models for adjusting the 1980 census. *Statistical Science* 1:3-39.

Fuller, W., and Harter, R.
1987 The multivariate components of variance model for small-area estimation. Pp. 103-123 in R. Platek, J.N.K. Rao, C.E. Sarndal, and M.P. Singh, eds., *Small-Area Statistics*. New York: John Wiley.

Kackar, R., and Harville, David A.
1984 Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association* 79:853-862.

Morris, Carl N.
1983 Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association* 78:47-54.

Rao, J.N.K., and Prasad, N.G.N.
1987 On the Estimation of Mean-Squared Error of Small Area Predictors. Submitted to *Journal of the American Statistical Association*.

Rubin, Donald B.
1980 Using empirical Bayes techniques in the law school validity studies. *Journal of the American Statistical Association* 75:801-816.