

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.				
1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE August 2003	3. REPORT TYPE AND DATES COVERED Technical Note	
4. TITLE AND SUBTITLE The U.S. Army's Health Risk Appraisal (HRA) Survey, Part I: History, Reliability, and Validity			5. FUNDING NUMBERS	
6. AUTHOR(S) L. Senier, N.S. Bell, S.R. Strowman, C. Schempp, P.J. Amoroso				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Military Performance Division US Army Research Institute of Environmental Medicine Natick, MA 01760-5007			8. PERFORMING ORGANIZATION REPORT NUMBER TN03-6	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) US Army Medical Research and Materiel Command Fort Detrick Frederick, MD 21702-5012			10. SPONSORING / MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES				
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.			12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words) The U.S. Army offered a health risk appraisal (HRA) from 1988 to 1998 as part of a comprehensive health promotion program. This report establishes the utility of the Army's HRA as a research tool. The Army initially adopted a modified version of the Rhode Island Wellness Check and later implemented a customized HRA, based on the CDC/Carter Center's HRA and original items. This report reviews what could be found in the open literature about the psychometric properties of the HRA items. The literature suggests that the utility of the data gathered varies widely. By combining HRA data with other Army data, however, it is possible to evaluate the psychometric properties of self-reported health habit data within the military a population that is often understudied, but is also more ethnically diverse than the U.S. population at large. In this way, the HRA database could make a substantial contribution to the literature and can inform not only health promotion efforts within the military, but research efforts in the civilian world as well. Finally, this report reviews some of the important lessons learned in the implementation of the Army's health promotion program and in the development of the HRA questionnaire.				
14. SUBJECT TERMS Army, military, health risk appraisal, survey, reliability, validity, TAIHOD, risk taking, epidemiology			15. NUMBER OF PAGES 111	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED	18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED	19. SECURITY CLASSIFICATION OF ABSTRACT UNCLASSIFIED	20. LIMITATION OF ABSTRACT UL	

USARIEM Technical Note TN03-6

**THE U.S. ARMY'S
HEALTH RISK APPRAISAL (HRA) SURVEY, PART I:
HISTORY, RELIABILITY, AND VALIDITY**

Laura Senier
Nicole S. Bell
Shelley R. Strowman
COL Catherine Schempp
COL Paul J. Amoroso

Military Performance Division

August 2003

U.S. Army Research Institute of Environmental Medicine
Natick, MA 01760-5007

CONTENTS

List of Figures.....	v
List of Tables.....	v
Acknowledgments	vi
Executive Summary	1
Chapter 1: Development of the Army's Health Risk Appraisal Questionnaire	3
The HRA as Part of the Army's Health Promotion Program	3
Development of the Army's HRA Questionnaire.....	7
Chapter 2: The HRA Questionnaire as a Research Tool	11
Reliability and Validity Concepts.....	11
Reliability	11
Test-Retest Reliability	11
Alternate-Form Reliability	12
Internal Consistency	12
Interobserver Reliability	12
Validity	13
Face Validity	13
Content Validity	13
Criterion (Empirical) Validity	13
Sensitivity and Specificity	14
Reliability and Validity of HRA Risk Estimation Scores	15
Reliability of HRA Risk Estimation Scores	15
Validity of HRA Risk Estimation Scores	16
Implications for the Army's HRA Data	22
Chapter 3: The HRA Items	24
Exercise	26
Reliability and Validity	26
Implications for the Army's HRA Data	27
Diet	28
Stress	29
Motor Vehicle Safety	29
Reliability and Validity	30
Implications for the Army's HRA Data	33

Alcohol Consumption	34
Consumption of Alcoholic Beverages	35
Implications for the Army's HRA Data	39
The CAGE	39
Implications for the Army's HRA Data	42
Other Alcohol-Related Problems	43
Diabetes	43
Reliability and Validity	43
Implications for the Army's HRA Data	44
Hypertension	44
Tobacco	46
Reliability and Validity	47
Implications for the Army's HRA Data	49
Periodic Health Exams	50
Women's Health	51
Reliability of Cancer Screening Practices	52
Reliability of Self-Report of Age at Menarche	55
Validity of Cancer Screening Practices	55
Validity of Self-Report of Pap Smear History	58
Validity of Self-Report of Mammography History	62
Validity of Self-Report of Family History of Breast Cancer	65
Implications for the Army's HRA Data	65
Men's Health	66
Summary	67
 Chapter 4: Lessons Learned, Conclusions, and Recommendations	 71
Lessons Learned From the Army's Health Promotion Program.....	71
Proponency and Ideology	71
Implementation Issues	73
Research Findings and Program Justification	74
Inadequate Program Funding and Decentralized Program	
Administration	75
External Pressures.....	76
Lessons Learned in the Development of the HRA Questionnaire.....	77
 References	 80
 Appendix A: The HRA Questionnaire	 87

LIST OF FIGURES

<u>Figure</u>		<u>Page</u>
1	Development of an Installation Health Promotion Program	5
2	Health Promotion Process	7

LIST OF TABLES

<u>Table</u>		<u>Page</u>
1	Summary of Studies on Reliability and Validity of HRA Risk Scoring	17
2	Summary of Studies of Test-Retest Reliability of Self-Report of Physical Activity	27
3	Test-retest reliability of the alcohol items on the Army's HRA among active-duty Army soldiers (N = 766), with 7-30 days between surveys	36
4	Associations between self-reported alcohol use and subsequent adverse health and occupational outcomes among active-duty Army soldiers taking the HRA, January 1, 1991 – December 31, 1998	38
5	Summary of Studies Evaluating Test-Retest Reliability of Self-Report of Current Smoking Status and Number of Cigarettes Smoked Per Day	47
6	Summary of Studies Evaluating Test-Retest Reliability of Self-Report of Pap Smears, Mammograms, and Clinical Breast Exam	54
7	Studies Evaluating Validity of Self-Report of Women's Health Screening Practices	56
8	Quantitative Measures of the Validity of Self-Report of Pap Smear History	61
9	Quantitative Measures of the Validity of Self-Report of Mammography	64
10	Summary of Reliability and Validity of the Self-Reported Health Habit Data Gathered by the Army's HRA	68

ACKNOWLEDGMENTS

We were aided in the preparation of this report by the generous assistance of many people who have dedicated their careers to the establishment of a comprehensive military health promotion program. The current stewards of the HRA, Ms. Judith Harris and LTC Sharon Reese, of the U.S. Army Center for Health Promotion and Preventive Medicine's Directorate for Health Promotion and Wellness, patiently answered numerous questions and provided some background and historical information that was useful to us in the development of this report. We also thank the current managers of the HRA database, Mr. David White and Mr. Joseph Powell of the U.S. Army Medical Information Systems and Services Agency, for providing us with the electronic files of the HRA historical database. Working with the actual data has improved our understanding of the questionnaire's strengths and limitations immeasurably. We owe an enormous debt to all of the former HRA project officers: MAJ (ret) Ken Bush, LTC (ret) Denise Chandler Phillips, COL Sandra Goins, LTC Mary Sanders, and COL Catherine Bonnefil, each of whom granted us telephone interviews and responded to numerous e-mail requests for information and clarification about the history of the program, the details of the questionnaire, and the management of the data. Although there are some pieces of this puzzle that still elude us, the conversations we had with the early project officers were enormously helpful to us in shaping our understanding of the program. Two members of the original Surgeon General's Task Force on Fitness, MAJ (ret) Sandra Yaney and COL (ret) Jeanne Picariello, helped us understand how the foundations of that program laid the groundwork for the eventual Army-wide program. We also benefited from the technical expertise of COL Jerel Zoltick, one of the original proponents of the Over-40 Cardiovascular Screening Program. COL Scott Goodrich and COL Kenneth Hoffman provided information about the HEAR. At every stage of this process, we have been helped by the indefatigable COL Charles Scoville, who was a member of the Surgeon General's Task Force on Fitness. He was able to provide copies of many early documents that shed light on the early days of Army health promotion and, through an extensive network of personal acquaintances, put us directly in touch with many of the people who have been so helpful to us as we worked on this report. LTC Laurence Branch lent his considerable expertise in survey research to a careful reading of an early draft of this report.

In our review of the reliability and validity of individual survey items, we were assisted by survey researchers from a variety of civilian and federal agencies. Ms. Susan Liss, NHTS Project Manager at the Federal Highway Administration, and Ms. Nancy McGuckin, Travel Behavior Analyst, directed us to published and unpublished data on self-reported travel and vehicle use. Dr. David Moriarty of the Centers for Disease Control and Prevention provided us with information on the CDC's HRA, particularly their items on stress. Dr. Ed Hutchins of the Healthier People Network, and Dr. Robert Cabral, formerly of the Rhode Island Department of Public Health, kindly spent time with us in telephone interviews and told us some of the details of the selection and implementation of these instruments by the Army. Dr. David Manglesdorff, Dr. Louis Banderet, and LTC Charles Engel provided advice and encouragement in our ultimately unsuccessful search to find the source of the HRA's

items on stress and, in the absence of definitive source information, advised us on what conservative conclusions we might reach about the quality of these items. We are indebted to COL (ret) Paul Harig and MAJ (ret) Rufus Howe for helping us understand how trends in health promotion swayed the development of the HRA instrument and affected its implementation. Dr. Robert Karch of the American University's National Center for Health Fitness provided details about the ARSTAF (Army Staff Headquarters) Corporate Fitness project. Above all, we thank him for putting us in touch with Dr. Maura Stevenson, whose dissertation on the first decade of the Army's health promotion program was essential in clarifying our conceptual understanding of the context in which the Army's health promotion effort was developed.

This work was supported in part by grants from the National Institute on Alcohol Abuse and Alcoholism (#R01 AA11407 and #R01 AA13324) and the U.S. Army Medical Research Acquisition Activity (DAMD 17-98-1-8610).

EXECUTIVE SUMMARY

The U.S. Army offered a health risk appraisal from 1988 to 1998 as part of a comprehensive health promotion program. Although health risk appraisals are typically designed and used solely for educational and diagnostic purposes, and not to gather information for research purposes, the Army's Health Risk Appraisal (HRA) has yielded an enormous database of self-reported information about health habits that is potentially useful for both surveillance and research efforts. This report documents the history of the Army's HRA and establishes its utility as a tool for epidemiologic research.

The Army used several different iterations of a health risk appraisal questionnaire during the life of the program. It initially adopted a modified version of the Rhode Island Wellness Check, and then, in 1990, implemented a customized health risk appraisal based on items from the Center for Disease Control and Prevention/Carter Center's HRA (CDC/Carter Center HRA) and new items authored specifically for the military. It does not appear that the Army ever undertook any formal efforts to evaluate the psychometric properties of the individual survey items. The HRA database represents the best single source of data on health habits of active duty Army soldiers, but before any HRA data can be profitably used in surveillance or research, a thorough understanding of their strengths and limitations is needed. In the absence of any Army-led studies of the reliability or validity of the Army's HRA, this report reviews what could be found in the open literature about the reliability and validity of the HRA risk estimation scores and the responses garnered by individual items.

The quality of the data gathered by the Army's HRA data varies, at least for purposes of epidemiologic research. In some cases, the literature indicates that certain items perform fairly well, and may be useful in surveillance and research. In some cases, the literature suggests that other items may be useful in combination with other data on health habits (e.g., the seat belt item may be useful in combination with other items in assessing risk-taking propensity). In other cases, however, there is serious doubt as to whether some items produce reliable and valid responses; these items from the HRA may not be of sufficient quality for epidemiologic research without corroboration from other sources or adjustment for potential misclassification. The Army's HRA database could make a substantial contribution to the literature about reliability and validity of self-reported health habits. It could be combined with other Army data sources to evaluate the reliability and validity of self-reported health habit data within the military population—a population that is not only often understudied, but also has a greater percentage of members from minority racial and ethnic backgrounds than the U.S. population at large. Efforts to evaluate the reliability and validity of data collected by the Army's HRA can inform not only health promotion efforts within the military, but can inform research efforts in the civilian world as well.

There is much to be learned from the Army's experience with the HRA, and many lessons that can be applied to the development of future questionnaires or health behavior surveys, whether in military or civilian contexts. The final chapter of this report reviews some of the important lessons to be learned from the implementation of the

Army's health promotion program and from the development of the HRA questionnaire. The Army learned a great deal in launching its health promotion program including, for example, important experiences in the design, development, and implementation of health habit survey instruments, and valuable experience in analyzing the data gathered with such tools. This report concludes by reviewing some of the things that the Army could have done to improve development of the instrument and articulates some lessons they might apply to the development of future survey instruments.

CHAPTER 1: DEVELOPMENT OF THE ARMY'S HEALTH RISK APPRAISAL QUESTIONNAIRE

The U.S. Army offered a health risk appraisal from 1988 to 1998 as part of a comprehensive health promotion program. Health risk appraisals generally comprise three components: (1) measurement of risk factors for the individual based on life style habits, personal medical history, and family medical history; (2) use of the individual's risk factors to predict his or her risk of death (usually expressed as a risk of death within a specified time frame or as a "recalculated age"); and (3) feedback to the individual on ways to modify lifestyle behaviors to reduce the risk of disease, injury, and death (9). Although health risk appraisals are designed as educational and diagnostic tools and not to gather information for research purposes, the Army's Health Risk Appraisal (HRA) has yielded an enormous database of self-reported information about health habits that is potentially useful for both surveillance and research efforts.

This report documents the history of the Army's HRA and establishes its utility as a tool for epidemiologic research. A companion report (12) describes the generalizability of HRA survey responses and tests for sampling or response bias by describing the demographic characteristics of active-duty Army soldiers who completed an HRA and comparing them to the Army at large.

The first chapter of this report briefly describes how the HRA functioned in the broader context of the Army's health promotion program and reviews the development of the Army's HRA questionnaire. Later chapters review what is known about the validity of the HRA risk assessment scores, the reliability and validity of the individual items, and some lessons learned in the Army's experience with health promotion and health habit questionnaires such as the HRA.

THE HRA AS PART OF THE ARMY'S HEALTH PROMOTION PROGRAM

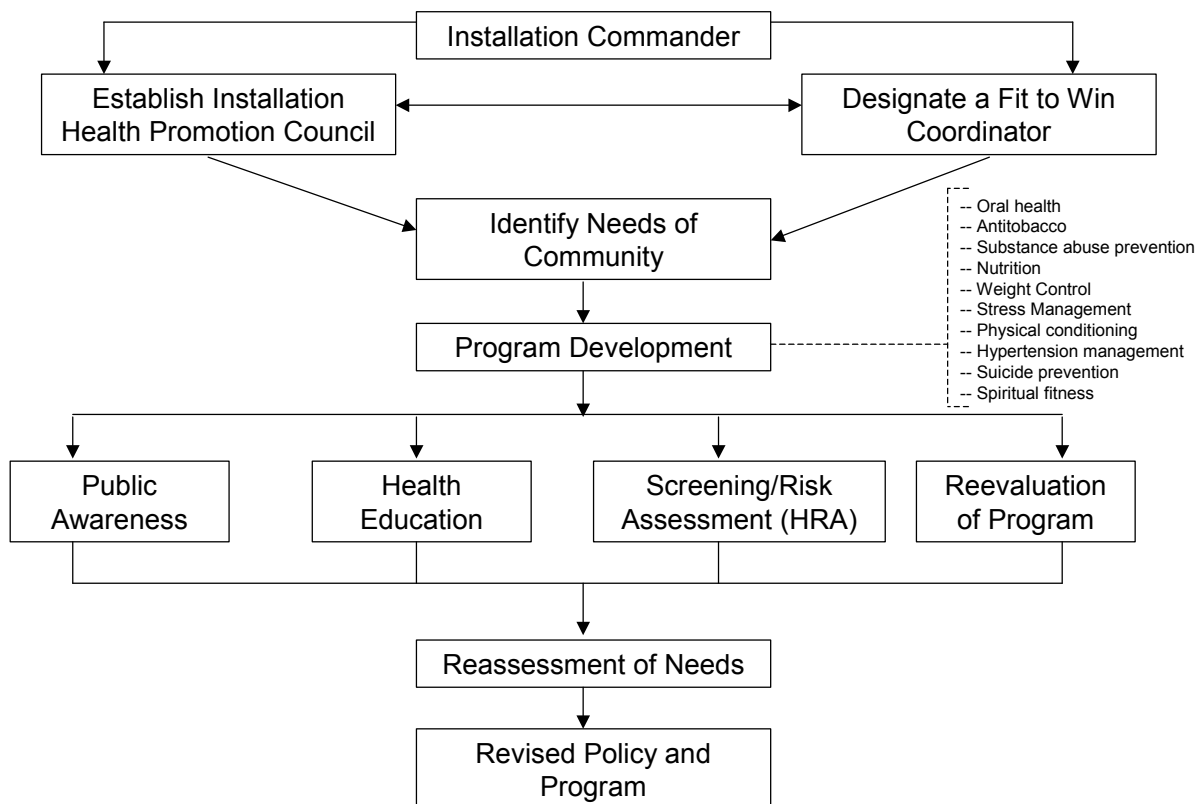
The Army's health promotion program was mandated by Department of Defense (DoD) Directive 1010.10, issued on March 11, 1986, to take effect June 1, 1986 (39). This Directive required all DoD agencies (i.e., all branches of military service, reserves, and defense agencies) to establish health promotion activities, and specifically called for health screening, health education on a variety of topics, and the promotion of a healthy work environment (e.g., it superseded previous DoD requirements about smoke-free workplaces). This Directive targeted six priority areas of health promotion activity: smoking prevention and cessation, physical fitness, nutrition, stress management, alcohol and drug abuse, and early identification of hypertension. In implementing their individual programs, DoD agencies were allowed to address additional goals if they chose to do so, but at a minimum, the programs they put in place had to include components in these six core areas.

In response to this requirement, the Army enacted Army Regulation (AR) 600-63 in November of 1987, outlining the specifics of the Army's health promotion program (41). This regulation placed responsibility for the health promotion program with the

Office of the Deputy Chief of Staff of Personnel (ODCSPER). According to AR 600-63, the Army's health promotion program was designed to address ten specific health promotion objectives (i.e., tobacco control, physical conditioning, weight control, nutrition, stress management, alcohol and drug abuse prevention and control, early identification of hypertension, suicide prevention, spiritual fitness, and oral health). In addition, the regulation asserted that, "health promotion necessarily includes other related activities . . . such as physical and dental examinations, health risk appraisals, physical fitness facilities, recreation and leisure education and activities, as well as initiatives to promote social and emotional well-being (41)."

While the ODCSPER identified these specific priority areas as the focus of the Army's health promotion activities, the design and delivery of specific interventions occurred at individual bases or installations. Figure 1 shows the development of an installation health promotion program, and how screening and health education were intended to function in such a program. In this model, local responsibility for health promotion activities was shared by a "Fit-to-Win" coordinator and a health promotion council, under the supervision and ultimate authority of the installation commander. Aggregate data were to be provided to the installation commander to facilitate development of targeted interventions based on the needs of the local population. By allowing commanders to customize a health promotion program within their command, the program could be more responsive to the needs of the units or the individual soldiers. Figure 1 outlines a basic process of needs identification, program development and implementation, reevaluation, and revision as the means to establishing such a program.

Figure 1. Development of an Installation Health Promotion Program



The regulation thus specifies that overall responsibility for the health promotion program rests with the ODSCPER, with technical assistance from the Office of the Surgeon General (OTSG), but that actual implementation should be executed on individual bases by the local command. This arrangement was intended to leverage both the authority of the ODSCPER and the expertise of the OTSG, with the end result being a customized program tailored to the needs of the local population. As we will explore later in this report, however, ideological differences and competition between ODSCPER and OTSG for control of various program elements would ultimately hinder the implementation of the health promotion program in some important ways. Furthermore, although the OTSG provided funding so that installations could hire a community health nurse to administer the HRA program, the ODSCPER did not provide any additional funding to hire Fit-to-Win coordinators or to fund health promotion activities. As a result, funding for health promotion initiatives varied widely across the major Army commands; in some cases, this may have impacted the overall success of the program.

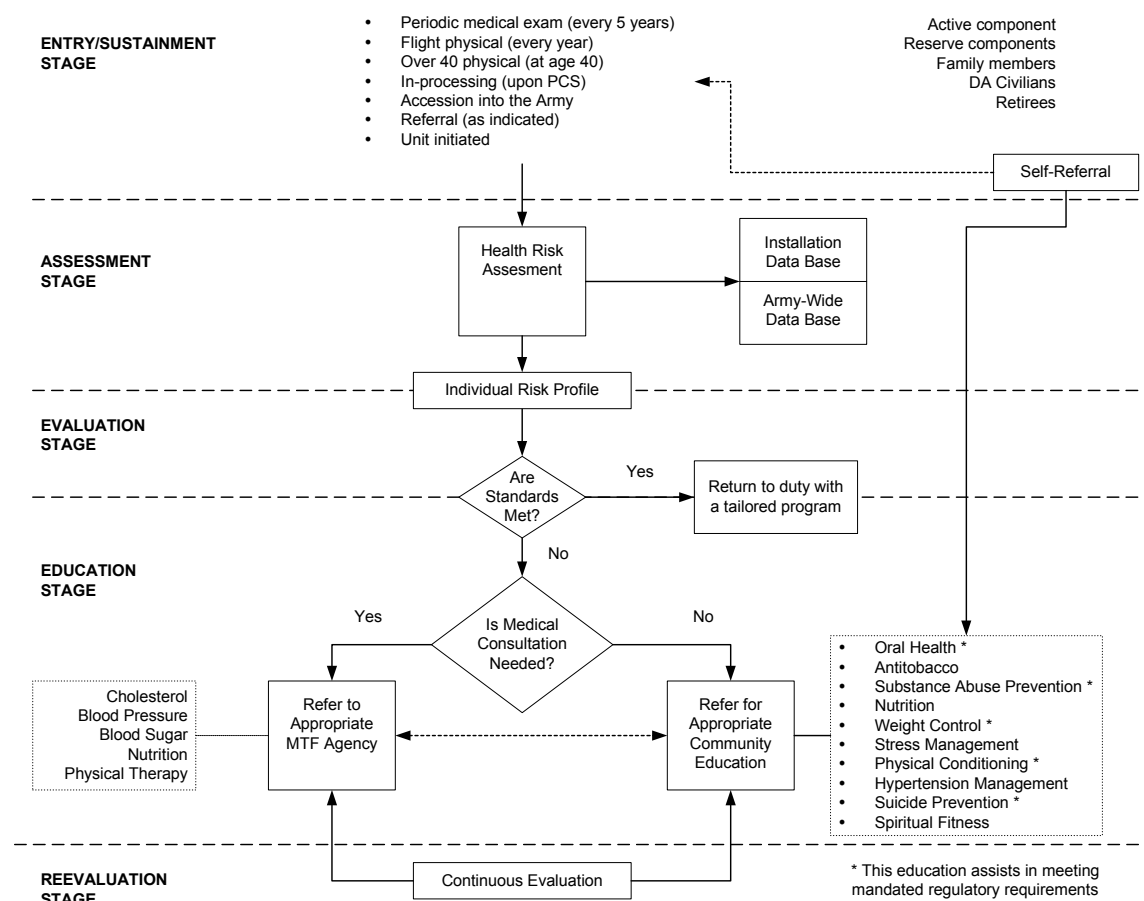
The Army's health promotion program was originally designed to include three types of screening and risk assessment tools: general health risk appraisal, cardiovascular screening, and fitness evaluation. Only the HRA and the cardiovascular

screening component elements were ultimately implemented. The data collected from these tools were to be used for program and resource planning, making comparisons about the health status of beneficiary groups, evaluating intervention programs, and assessing trends in health behaviors.

Figure 2 shows the health promotion process at the level of the individual. Eligibility for the health promotion program extended to active duty and reserve soldiers, family members, civilian employees of the Army, and retirees. The typical entry point into the health promotion program for soldiers (and their families) was accession into the Army, but there were also other means by which people could enter the health promotion program (e.g., periodic medical exams, inprocessing to a new assignment). Participants may also have self-referred into the process (e.g., by presenting for care at a health clinic that offered the HRA or even by specifically asking to take an HRA) or have been directed to the program by someone in their chain of command. In the early years of the program, it was assumed that most soldiers would take the HRA as part of a routine physical exam (109), although it would ultimately become more common for soldiers to take it as part of inprocessing to a new base or duty assignment.

The first step in the health promotion process was the administration of the HRA questionnaire (see Appendix A). This screening instrument queried the respondent on various health habits and behaviors and generated an individual risk profile. The HRA was typically administered by a community health nurse who briefed the soldiers on the purposes of the questionnaire and reviewed the critical items that must be completed. On the basis of the individual's risk profile, the HRA respondent received a customized report documenting the most immediate risks to their health. This report may have included medical or behavioral interventions, if warranted (e.g., a soldier may have been referred to a medical treatment facility for management of hypertension, or to an education program such as smoking cessation or weight control). Participants were to be reevaluated after the medical or behavioral interventions and, if they required additional intervention, be referred again as necessary (41).

Figure 2. Health Promotion Process



PCS = Permanent Change of Station; MTF = Medical Treatment Facility

AR 600-63 enumerated, as one of the responsibilities of the OTSG, the planning, implementation, and evaluation of “an automated health risk appraisal with procedures for administration and for processing and compiling the data at HQDA (Army headquarters), MACOM (major Army command headquarters), installation or community, and unit levels.” Figure 2 shows that individual HRA survey results were to be maintained in databases at both the installation and Army-wide levels. Although required by regulation, it is unclear whether these Army-wide databases were maintained, as we have not been able to locate an electronic repository of pre-1990 HRAs.

DEVELOPMENT OF THE ARMY’S HRA QUESTIONNAIRE

The Army had been conducting various health promotion activities throughout the 1960s, 1970s, and 1980s. In June 1987, when the DoD issued Directive 1010.10, requiring all of the services to design comprehensive health promotion programs, the Army formalized its activities in AR 600-63, and consolidated its various health and wellness programs under the ODCSPER. As part of this effort, the Preventive Medicine Division of the OTSG was tasked with the responsibility of selecting a health risk appraisal questionnaire (109).

The development of the health risk appraisal instrument has been identified as one of the most contentious points in the history of the Army's health promotion program (109). In the late 1980s, the Army was working simultaneously on two different components of the health promotion program: the health risk appraisal and a physical-fitness screening program for soldiers over age 40.

Typically, soldiers complete semiannual physical fitness tests that include two-minute timed tests of maximal sit-up performance, push-up performance, and a two-mile timed run. Prior to 1981, soldiers over age 40 were exempt from this fitness-testing requirement, but this exemption was eliminated by a new DoD Directive on physical fitness and body fat requirements, issued in 1981 (40). This caused great concern among Army physicians who feared that this requirement might place soldiers at risk of cardiovascular-related deaths during physical fitness testing or during regular physical fitness training.

In addition to the semiannual fitness test, soldiers typically undergo periodic physical exams at induction into the Army and then every 5 years starting at age 20. The Surgeon General tasked two cardiologists in the Preventive Medicine Division with responsibility for developing a screening process that could be administered as part of the periodic physical exam soldiers underwent at age 40. The objective of this screening program was to estimate coronary risk for an individual at age 40 in order to determine whether they should participate in the semiannual fitness test, and to then update that analysis every 5 years. Phase I of the Over-40 Cardiovascular Screening Program exam consisted of screening for risk factors established by the Framingham Heart Study (sex, age, systolic blood pressure, cholesterol, smoking status, resting electrocardiogram, and glucose tolerance). If a soldier met a certain risk profile based on their Framingham risk score, they were referred for other evaluations and interventions as necessary (e.g., treadmill test).

While the proponents of the Over-40 program were proceeding with this approach, the health risk appraisal selection committee was simultaneously developing plans to administer the HRA that was required by the health promotion program through periodic physical exams. Even though all parties concerned belonged to the Preventive Medicine Division of the OTSG, they differed widely in their philosophical approaches to health risk appraisal and in selection of an appropriate survey instrument. The cardiologists in charge of the Over-40 program favored selection of a health risk appraisal that used a risk estimation methodology based on the Framingham heart study data. The health risk appraisal selection committee, on the other hand, viewed the risk estimation methodology with skepticism, dubbing it "pseudoscience," and instead favored an HRA that would give "simple congratulatory messages for positive health behaviors and messages of concern for negative behaviors (109)."

By 1985, the OTSG's health risk appraisal selection committee had decided, over the objections of the Over-40 program team, to adopt the Rhode Island Wellness Check (RIWC) questionnaire as its Army-wide vehicle for health risk assessment (109). In selecting an instrument, the committee focused on two areas: how labor intensive it would be to implement the instrument, and whether or not the instrument gave the

respondent “appropriate” messages about health objectives. The RIWC instrument appealed to this committee partly because they believed it met their criterion of low labor intensity (it was readily available, had been optimized for administration via a computer-scannable form, and had computer software so that the questionnaire could be easily scored). They also approved of the “output messages,” because the RIWC does not express risk as a recalculated age, but instead compares the respondent’s scores to mean scores for people of his or her sex and age. This version of the HRA was pilot-tested at six U.S. bases in 1986 (Forts Jackson, Lewis, Bliss, Carson, Bragg, and Leavenworth) (109).

The Army did have some prior experience with an HRA based on risk estimation methodology, however. In the early 1980s there had been several exercise-related cardiovascular deaths that occurred during physical training, thus bolstering concerns that the Army’s physical fitness requirement might place some soldiers at risk of cardiac arrest. In approximately 1982-1983, the Army used the Center for Disease Control and Prevention’s (CDC) HRA at the Command and General Staff College at Ft. Leavenworth, Kansas, to see if it was useful in detecting prevalence of cardiovascular risk factors in a group of soldiers under age 40 and in identifying specific health conditions for individual follow up. This program evaluated the utility of the CDC’s HRA as both a primary cardiovascular screening tool and a method of initiating a comprehensive risk intervention program.¹ The coordinators of this program ultimately concluded that it was not cost-effective to screen all Army soldiers for cardiovascular disease because of the high proportion of false-positives in a population under age 40 (109).

Meanwhile, health risk appraisal methodology was also enjoying a surge in popularity in the civilian sector. In the mid-1980s, the CDC and the Carter Center at Emory University embarked on a collaborative effort to update the CDC’s health risk appraisal questionnaire and risk algorithms. As a result of this work, the CDC’s public domain health risk appraisal was updated and the Carter Center obtained permission to offer a version of that health risk appraisal to corporate clients.

In 1988, shortly before the Army launched its health promotion program, the Army’s health risk appraisal selection committee decided that they wanted to use the CDC’s instrument instead of the RIWC version (109). This not only represented a major shift in ideology for this committee, but also greatly increased the complexity of the implementation, as the Army had already purchased computers and card scanners that would work with the RIWC-based instrument (109). Shortly thereafter, the Army contracted the Carter Center to modify the CDC/Carter Center’s health risk appraisal for use by the Army, on the provision that they adapt the program components (e.g., questionnaires) to work with the computers and card scanners already purchased (109). This version of the health risk appraisal questionnaire was ultimately implemented in the fall of 1989 (122). It subsequently underwent minor revisions in 1992. Chapter 3 describes the 1992 version of the Army’s HRA form in greater detail.

¹ MFR, CPT Sandy Yanne, September 1983.

In the early stages of the health risk appraisal program there were procurement difficulties in getting the computer equipment (e.g., scanners) needed to process the health risk appraisal distributed to all Army installations. Dates of program initiation thus varied from installation to installation. Moreover, it is doubtful, given the degree of the logistical complexities involved, that all Army bases implemented new versions of the questionnaire at precisely the same time. We do not know what instructions were given to health promotion coordinators regarding transitions between versions of the health risk appraisal form, but it is probable that some bases adopted the newer versions of the form immediately while other bases may have exhausted their existing inventory of forms before using a newer version of the form. For these reasons, care should be taken in interpreting the composite risk assessment scores from the Army's HRA data, as the methods of calculating overall risk profiles are very different between the RIWC and the CDC/Carter Center's versions of the health risk appraisal.

The Army offered the HRA to active-duty soldiers for more than a decade, finally ceasing formal requirements for the program in late 1998 (although it is still in use at a few active duty installations and is being used by reserve components). The resulting databank of HRA survey responses contains a wealth of historical information about health habits and risk behaviors that may assist researchers in the study of health and wellness among Army soldiers. Before proceeding to use this information in quantitative research, however, an assessment of the psychometric properties of the questionnaire is appropriate. The next chapter introduces some basic concepts about reliability and validity, and reviews what is known about the validity of the individual HRA items, as well as the risk scores calculated from the HRA.

CHAPTER 2: THE HRA QUESTIONNAIRE AS A RESEARCH TOOL

The Army's HRA is an important resource for researchers interested in studying the effects of behavior on health. A large number of soldiers took HRA surveys while the program was in effect. Even the most conservative estimates put this figure at close to half a million individual active duty soldiers (12). The military is typically excluded from surveys of health habits conducted by civilian health agencies, such as the CDC's Behavioral Risk Factor Surveillance System (BRFSS). Thus, this bank of HRA survey responses can potentially provide researchers with valuable information on the prevalence of certain health habits and risk factors in a young, active, healthy, and largely understudied population. When combined with other sources of data, such as inpatient and outpatient hospitalization records, casualty records, disability evaluations, and accident reports, it is possible to study associations between these health behaviors and a wide variety of health outcomes, from chronic diseases to acute injuries. Moreover, there are some 90,000 thousand active duty soldiers who have taken the HRA more than once during their military careers, allowing for the assessment of changes in health behaviors and how such changes might impact health outcomes. In order to gauge the HRA's utility in describing soldier health behaviors and risk factors, however, it is important to first assess the reliability and validity of the questionnaire.

RELIABILITY AND VALIDITY CONCEPTS

Reliability

Reliability measures the extent to which a survey (or a particular survey item) produces consistent and stable responses over time (15). That is, a reliable survey administered to the same individual or group of people at two different times should result in the same set of responses. Reliability of responses to HRA items is important partly because an unstable instrument may interfere with the correct calculation of a risk assessment score, and it is this score that will determine whether the participant needs and gets referred to interventions that will benefit their health. Reliability may also be important if HRA scores are used to gauge the efficacy of the health promotion program; an unstable instrument would produce fluctuating pre- and post-program scores, and would make it impossible to parse out what degree of change is due to the success or failure of the program and what degree of change in scores is attributable to flaws in the questionnaire (103). Poor reliability can also attenuate correlations between the survey and other measures, and thus could be a problem when using survey responses for research purposes.

Test-Retest Reliability. In this type of assessment, the same survey is administered twice to the same group of people and a correlation coefficient (e.g., kappa or κ statistic, Pearson's r) is calculated to assess the level of agreement between the first and second sets of responses (15, 71). If the Pearson's r exceeds 0.70 or if the κ statistic exceeds 0.40, the test-retest reliability is judged to be fairly high (94, 71). The amount of time that elapses between the first and second administration is critical to the

assessment of reliability (15, 71); if the second test is administered too soon after the first, respondents may recall their first set of responses and simply repeat their answers to the first survey, making the two sets of responses appear to be more alike than they actually are. On the other hand, if too much time elapses between the first and second test, respondents may actually change their behavior, and thus give a different but still truthful set of responses on the second test (60). The different responses do not, in such a case, mean that the instrument is unreliable, but it is nearly impossible for the investigator to discern whether the differences are due to actual behavior change (true variance) or to the instability of the instrument (error). Another issue to consider with test-retest assessments is the proportion of successfully completed second surveys.

Alternate-Form Reliability. Alternate-form reliability is similar to test-retest reliability, but prompts respondents to answer similar forms of the same question on the same survey (15, 71). In this approach, a survey questionnaire will include two versions of the same question, but with different wording. Sometimes the wording of the question is changed, and sometimes the wording or order of the response set is changed. In another type of alternate-form reliability, respondents are administered the survey twice, but the items differ on the two surveys (although they are measuring the same constructs). In one of the more common analytic approaches (the so-called split halves approach), the total number of items on a given survey is divided in half, and then the scores on the two halves are correlated (15, 71). To ensure the accuracy and relevancy of this method of assessing a survey's reliability, it is important that the alternate forms of the question are at the same level of complexity with respect to grammar and vocabulary (71).

Internal Consistency. Another means of assessing the reliability of a survey is to look for internal consistency among the responses to various items (71). A series of items that are all designed to measure the same thing, or different facets of the same thing, should produce similar responses. For example, if a person reports that they consume a large number of drinks per week, one might also expect them to report that their friends worry about their drinking. We may also expect them to be more likely to report that they are trying to cut down on their drinking than a respondent who reports comparatively fewer drinks per week. The correlation between similar items is usually measured and expressed as the coefficient alpha, or Cronbach's alpha (α) (71). Alpha is calculated based on the number of items and the average intercorrelation between items. As either of these increases, α will also increase.

Interobserver Reliability. This type of reliability is not germane for self-reported questionnaires, but when data are being collected by trained observers, it is useful to measure how closely the assessments of two observers match for a particular individual subject (71). It is especially important to measure interobserver reliability when the observers are making subjective assessments. In the case of the Army's HRA, interobserver reliability might have threatened the overall reliability of responses if the persons administering the questionnaire coached respondents in different ways prior to administration of the HRA. Although the community health nurses who administered the HRA all received similar training on how to administer the survey, there is some anecdotal evidence that other parties (e.g., NCOICs, unit leaders) may have

consciously or unconsciously exerted peer pressure on soldiers to influence responses (for example, discouraged soldiers from truthfully reporting unhealthy habits, such as smoking, on the HRA). Because there were so many parties administering the HRA, and because these reports of influenced responses are purely anecdotal, it is hard to know how widespread this phenomenon was and whether and how it may have impacted the overall reliability of responses.

Validity

Validity is a measure of how accurately the survey or survey item measures what it intends to measure (15). For example, if you are trying to assess alcohol consumption, are you accurately measuring the amount of alcohol actually consumed, or do your questions actually gather information about some other type of behavior, such as purchasing patterns? Validity may be threatened by many factors including questionnaire wording, and recall and selection biases. For example, are responses to the alcohol item uniformly lower than actual consumption for the total population of respondents, or only for some subset of this population?

Face Validity. The simplest type of validity, face validity, refers to the extent to which the survey items appear to be logically related to the behavior or characteristic they are supposed to be measuring (15). If you are surveying people about wealth and poverty, for example, an item asking about annual income has better face validity than an item asking about how much is spent monthly on going to the movies. A survey may have good face validity, but still may not necessarily demonstrate empirical validity. For example, a survey asking about self-reported dietary habits may demonstrate good face validity but still not be highly correlated with body fat or future physical fitness test performance. Similarly, a survey that seems not to have good face validity may in fact still be correlated with another important outcome or variable of interest.

Content Validity. Content validity refers to how well a survey covers the domain of interest, as evaluated by a group of experts in that field (15). In designing a survey instrument to assess a complex topic, it is useful to think of that topic as having various facets, and to write a variety of questions to address each facet of that topic. For example, in assessing health, you might want to write several questions to gather information on various aspects of health, such as exercise habits, tobacco and alcohol use, preventive health practices, and diet. Once you have constructed your questionnaire, it is helpful to show it to several experts in that field. Experts should judge the quality and relevancy of the items on the survey and suggest additional items that might be important. A panel of subject matter experts should include people with different areas of expertise (e.g., for a health questionnaire, you might want to include a physician, a nurse, a physical therapist, and a nutritionist on your review panel).

Criterion (Empirical) Validity. This type of assessment compares the performance of a survey instrument against another criterion to see how well the two measures correlate (15). Criterion validity can be either predictive or concurrent (15). To assess whether a survey item has predictive validity, you might gather information on self-reported drinking and driving behavior among a group of people, and then

survey them for a period of time to see whether they experience future hospitalizations for alcohol-related conditions or motor vehicle-related injury hospitalizations. Concurrent validity is assessed by correlating responses to a survey with an independent criterion, especially one regarded as a “gold standard,” and which is measured at the same time.

Concurrent validity can also be assessed within a group of survey items that all aim to measure an underlying construct (15). Suppose, for example, you have a survey that attempts to assess alcohol-related problems, consisting of several items that measure different aspects of alcohol consumption and related behaviors. If one of those items has been shown to correlate closely with an external measure (e.g., if the risk of having a diagnosis of cirrhosis of the liver correlates closely with the number of drinks per week reported by respondents), then you could use Cronbach’s α to assess concurrent validity, in much the same way you would to assess internal consistency. The difference is that in internal consistency, the α expresses how well the items relate amongst themselves; if, however, one of those items has been validated with an external measure, Cronbach’s α may also be used to judge the concurrent validity of the group of items with that external measure.

Sensitivity and Specificity

The utility of a screening measure is often measured by its sensitivity and specificity, or its ability to correctly classify respondents. This relates in large part to the empirical and face validity of the items contained in the HRA survey. At issue is how well a test detects a disease or behavior when it is truly present, and how likely it is to indicate it is present even when it is not. Sensitivity is the probability that the test will be positive given the presence of the disease, as confirmed by a supposedly definitive diagnostic test (82, 92). Closely related to this measure is the false negative rate, or the proportion of people who truly have the disease but obtain a negative result from a test or screening measure (the false negative rate is equal to 1-sensitivity). Specificity is the probability that a test will be negative given the absence of disease (82, 92). The false positive rate is the proportion of people who do not have the disease but obtain a positive result from the test screening measure. In the ideal world, diagnostic tests and screening measures would be both highly sensitive and highly specific (82). In reality, this is seldom the case, and compromises must be made between sensitivity and specificity. In general, highly sensitive tests are preferred when the consequences of not detecting a disease are dangerous, such as treatable cancers. Highly specific tests are preferred when false positive results are harmful or may cause distress to the individual, such as in the early days of the HIV/AIDS epidemic, when there were no effective treatments. In the case of the HRA, good sensitivity would be demonstrated by accurately identifying respondents at risk via their self-reported behaviors (e.g., good criterion validity). Good specificity is demonstrated when *only* those individuals whose behaviors actually place them at risk are targeted for intervention or counseling.

RELIABILITY AND VALIDITY OF HRA RISK ESTIMATION SCORES

The rationale for health risk appraisals was developed and popularized in the 1980s, during a time when health care costs were spiraling upwards rapidly. Managed care organizations and corporations were actively seeking ways to control the escalation of costs. Many people hoped that the combination of health risk assessment and health promotion programs would be useful in halting this inflation, and there was pressure during this time to make health risk appraisals available so that they could be implemented in health promotion programs. Edington et al. speculate, in their review of the literature, that this pressure may have, “rush(ed) technology into practical application ahead of basic testing (45).” For this reason, studies of the reliability and validity of health risk appraisal methodology have been sparse, and have tended to focus on the accuracy of the calculation of risk scores and technical problems in the estimation of risk rather than on reliability and validity of individual items (45). The algorithms and computations that lie behind most health risk assessments (including the Army’s) generally draw upon three sources of information: death certificate data for average probability of dying from every cause of death for every combination of age, sex, and race; epidemiologic and clinical data assigning values (debits and credits) for health habits; and self-reports of these risk factors (45). The few studies that have assessed methodological issues have typically focused either on the reliability and validity of risk estimation, or on the efficacy of health risk assessment results as an educational tool to promote healthy behavior change. Table 1 summarizes the results of the reliability and validity studies of a variety of health risk appraisals. Fewer studies have examined reliability and validity of individual items, but we will review this body of literature in the next chapter.

Reliability of HRA Risk Estimation Scores

In their review of the literature, Edington et al. state that early studies of the test-retest reliability of many health risk appraisal questionnaires (not only the CDC’s or Army’s versions) showed weak correlations (45). They go on to note that this is probably not surprising, since most health risk appraisals are long, and it would be unreasonable to expect that people would answer such a detailed battery of 35-70 questions in exactly the same way twice. People may change their responses to items as they learn new information about their medical history, but changes in responses may also reflect true behavior change (e.g., a person may receive a report from a health risk appraisal that tells them they are at risk of cardiovascular disease and may make changes to their exercise habits or diet because of this information, or they may mature or “age” out of the behavior).

Paradoxically, unreliable (that is, inconsistent) responses on individual items may not necessarily compromise the calculation of a valid risk score. Edington et al. note that results of most risk calculation algorithms are minimally affected by minor changes in responses (45). Although it is widely accepted that behavior impacts health and longevity, there are still many unknown factors (such as genetics or environmental exposures) that also play a role in determining the course of morbidity and mortality (92). Health risk appraisals are necessarily limited by the extent to which they can

quantify only the known or predictable effects of behavior on health; until more is understood about the role of these unknown factors in the course of human health and disease, the algorithms that lay behind risk assessment scores will necessarily be incomplete, and unable to completely account for all of the factors that may accelerate or forestall disease.

Edington et al. also note, however, that the relative stability of health risk appraisal results in the face of minor reporting changes may not hold true for younger populations (45). Risk score calculations may be more highly influenced by inconsistent answers to certain types of questions, such as questions on motor-vehicle risk behaviors or alcohol consumption. Because these behaviors may be the principal risks reported by otherwise young and healthy people, reliability of composite risk scores may be more of an issue in younger populations than it is for older adults. Many of the studies that have been conducted to assess reliability of these instruments have limited themselves to respondents between the ages of 25 and 60. Therefore, care should be used in extrapolating the results from composite risk scores either to young adults (such as those who primarily comprise the U.S. Army population), or to elderly adults.

Validity of HRA Risk Estimation Scores

Validation studies have typically focused on how accurately the risk estimation algorithms predict mortality in a group of people over a period of time (usually 10 to 20 years) or against some other predictive model. Table 1 summarizes the studies that have examined reliability and validity of these risk algorithms.

Table 1. Summary of Studies on Reliability and Validity of HRA Risk Scoring

Author	Instrument Tested	Purpose of Study	Methods	Results	Conclusions
Smith (1987)	41 different HRAs - Type I: probability based, calculates 10-year mortality risk - Type II: probability based, calculates 8-year morbidity risk - Type III: self-scored; categorizes risk of CHD or MI as low-medium-high - Type IV: Life-expectancy model, self-scored, adjusts age based on reported risks - Type V: General health status, categorizes respondents as low-high risk	Validation of prediction of 10-year coronary heart disease mortality	- Developed two sets of logistic regression equations based on Framingham Heart Study 1956 exam data on 3,604 people and the Risk Factor Update Project (RFUP) - Took 240 test cases from the Framingham cohort (all white, ≥ 35 years old, missing data imputed) and computed HRA scores from 41 different HRAs - Compared results of HRA scores to logistic equations developed for Framingham and RFUP to assess the ability of the HRAs to predict mortality accurately	- Type I & II HRAS correlated most closely with Framingham and RFUP estimates - Most HRAs predicted higher risks than criterion models; most HRAs overestimate risk, even though they rank ordered the risk factors properly	Three factors affect the validity of HRA scores: - Sophistication of the algorithm (number of risk factors included) - Instruments with finer gradations of risk scores produced more valid risk scores than those with cruder categorizations - Instruments that included age and gender in the calculation of risk produced more valid estimates than those that did not

Author	Instrument Tested	Purpose of Study	Methods	Results	Conclusions
Foxman (1987)	CDC HRA	Validation of HRA risk age, by observed vs. predicted mortality (i.e., HRA risk score) in a subsample of Tecumseh Community Health study	- Used the CDC HRA to calculate risk age and 10-year all cause risk of mortality for 3,135 members of Tecumseh cohort (limited to white smokers or never smokers aged 25-60) - Categorized people by difference between age at baseline and risk age, then calculated proportion surviving 20 years for each age-sex group - Developed logistic regression equation to predict odds of mortality for a 1% increase in HRA predicted mortality	- As difference between chronologic age and risk age increased, observed proportion of people who had died also increased - Each 1% increase in HRA risk score was associated with 33% increase in mortality, controlling for age-sex-race predicted mortality	In this cohort, CDC HRA risk scores were more accurate in predicting 20-year mortality than typical age-sex-race predictions

Author	Instrument Tested	Purpose of Study	Methods	Results	Conclusions
Smith (1989)	4 HRAs: ▪ CDC's HRA ▪ Arizona Heart Institute's Heart Test ▪ American Heart Association's RISK0 ▪ Blue Cross/Blue Shield's Determine Your Medical Age	Test-retest reliability	▪ Subjects aged 25-65, no history of CHD, diabetes, or hypertension; N=338; selected randomly from community ▪ Subjects were reinterviewed 7-12 weeks after first HRA (time 1); 55% repeated the same HRA at time 2. ▪ Calculated test-retest correlation scores for responses on individual items and for HRA risk scores ▪ Developed regression models to evaluate impact of length of time between HRAs on reliability of responses	▪ Test-retest correlation coefficients for items on family history, smoking status, and relative weight $\geq .75$ for all four instruments ▪ Correlation coefficients for risk scores: CDC HRA $r = 0.84$, Arizona Test $r = 0.84$, BCBS $r = 0.99$, RISK0 $r = 0.76$ ▪ No appreciable change in correlation coefficients when analyses were restricted to participants who reported that their behavior had not changed	▪ Correlation coefficients for items on physical activity, diet, and stress were far less consistent between baseline and follow-up survey than for items on family history and smoking status ▪ Correlation coefficients for self-scored HRAs were lower than others, but improved when corrected for computational errors by participants ▪ Inconsistencies in responses more likely due to instability of participant response than to actual behavior change ▪ Length of time between surveys had little effect on reliability of either overall risk scores or responses on individual items

Author	Instrument Tested	Purpose of Study	Methods	Results	Conclusions
Smith (1991)	4 HRAs: - CDC's HRA - Arizona Heart Institute's Heart Test - American Heart Association's RSKO - Blue Cross/Blue Shield's Determine Your Medical Age	- Accuracy of respondents self-reported risk factors - Accuracy of HRA estimates of CHD mortality (based on self-reported risk factors) - Impact of errors in self-reports and respondents computational errors on validity of HRA risk score	- Subjects aged 25-65 recruited randomly; N=732 - Comparison of self-reported health behaviors with physiologic measurements or other gold standard - Investigators compared risk score obtained from HRA with interview data on behavior; risk score corrected for computational errors; and score that would have been obtained if risk had been calculated on basis of physiologic measurements rather than self-reports - The three sets of risk scores were correlated with logistic models predicting 10-year coronary heart disease risk for each respondent, based on NHANES I Epidemiologic Followup Study (NEFS) data	- Correlations on comparison of self-reported cigarette smoking and relative weight were fairly high ($\geq .6$) for all instruments; reports of physical activity, blood pressure, and serum cholesterol were less so - CDC's HRA had highest correlation between self-reported score and logistic estimate; however, 10-yr risk of CHD from this instrument was consistently higher than estimate from NEFS model	- Self-reported assessments of smoking status and BMI appear to be accurate for use in HRAs - Low accuracy of measures such as blood pressure and cholesterol suggest that HRA scores should be based on actual physiologic measures rather than participant self-reports of these factors - The validity of self-reported HRAs is compromised by participants' computational errors and lack of awareness of physiologic measures
Gazmararian (1991)	CDC's HRA Carter Center's HRA	Comparison of average HRA-predicted 10-year mortality risk from all causes and risk age from the two instruments	- Used the CDC and Carter Center HRAs to calculate risk age and 10-year all cause risk of mortality for 3,135 members of Tecumseh cohort (limited to white smokers or never smokers aged 25-60) - Compared differences between actual age and risk age from both HRAs - Constructed ROC curves to compare HRA-predicted risks by 10-year mortality rates, for men and women	CDC's HRA consistently overestimated predicted risk of mortality for both men and women; Carter Center HRA overestimated risk of mortality for men but underestimated risk for women	- Difference between actual age and risk age was less for Carter Center HRA - For some men in the sample (especially younger men), the Carter Center HRA was no better than chance at predicting 10-year mortality risk

The general consensus from studies of the validity of health risk appraisal scores seems to be that these instruments, while imperfect, perform fairly well in classifying individuals into low-, medium-, and high-risk groups, and in estimating relative risks of mortality that are associated with various health habits. They are thought to be less effective, however, in predicting individual risk of dying (45). Several of the studies described above indicate that appraisal algorithms based on the CDC's instrument (i.e., the Geller-Gesner credit-debit method of assessing risk) produce risk estimates that are very close to the criterion models (within 1%, but perhaps with a tendency to overestimate risk of mortality). It should be noted, however, that the fact that HRA scores for a random sample of participants correlates closely with the overall mortality calculations from the criterion models speaks only to the overall performance of the model. It is not unreasonable to expect that individuals with particular high- or low-risk health habits may obtain widely differing results (98).

The studies reviewed above indicate that assessment instruments that rely upon actual physiologic measurements of clinical parameters (e.g., blood pressure, serum cholesterol) produce more valid estimates of risk than assessment instruments that rely on participant self-reports. For example, Smith et al. found that less than one-third of respondents reported systolic blood pressure readings within 10 mm Hg of the actual readings taken by field technicians, and that only four percent gave cholesterol levels that were within 20 mg/dL of their actual cholesterol level (104). Some instruments, such as the CDC's HRA, handle missing values by imputing the population norm (45, 98, 103). This poses a particular problem for high-risk individuals who may be unaware of their true risk status. Take, for example, the case of a hypothetical patient with high cholesterol. If the participant does not know his or her true cholesterol level, and the health risk assessment process does not include a blood test to determine it, the computerized algorithm enters the average cholesterol level for a person of the same age, race, and sex in its place. This may produce a false negative result. The inability of people to accurately self-report such clinical data, or even to correctly guess whether their levels are higher or lower than normal, indicates that administration of health risk appraisals should be accompanied by clinical screening whenever possible. If the assessment process cannot correctly identify high- and low-risk individuals (sensitivity and specificity), its utility in promoting behavior change will be undermined.

There have been a few studies that have examined how incorrect reports by respondents may impact the correct calculation of risk scores. In their review of the literature, Edington et al. cite several studies that found that self-reports of physical activity levels correlate well with physiologic measures such as resting heart rate, resting blood pressure, and maximum oxygen uptake (45), suggesting that substituting participant self-reports for actual clinical values may not have an adverse impact on overall scores. Smith et al. found that self-reported data on smoking status and body mass index were consistent enough to be useful in computation of risk estimation scores, whereas their study of test-retest reliability cast doubt on the utility of the physical activity, diet, and stress items (103, 104). Smith et al. have also pinpointed problems in instruments that are self-scored, as participants may make computational errors that would produce invalid risk scores (103, 104). Fortunately, the Army's HRA

and the CDC's HRA, on which it is based, are both scored by a computer and so are not susceptible to this type of threat to validity.

Implications for the Army's HRA Data

How are we to interpret the results of these studies with respect to the Army's HRA? The algorithms that form the foundation of the CDC's HRA are based on epidemiologic work from studies that focus on adult populations (such as the Framingham Heart Study). Indeed, most of the validation studies reviewed in Table 1 restricted their analyses to persons between the ages of 25 and 60. In contrast, the Army has a large proportion of soldiers under age 25 (approximately 40%) (123). Moreover, many of the validation studies done to date have examined how well health risk appraisals predict mortality from coronary heart disease; it is unclear how valid the risk estimations are for other causes of death. In an editorial in the *American Journal of Public Health*, Victor Schoenbach described work by Chaves et al., who found that the top five causes of death were ranked in different orders depending upon the health risk assessment instrument used (98). In their review of the literature, Edington et al. described work by Elias and Dunton, who found that risk age calculations were fairly reliable across most age groups, except for younger age groups, whose mortality risks are often associated with risk behaviors such as driving and alcohol consumption (45). In comparing the CDC's HRA with the Carter Center's HRA, Gazmarian et al. found that the ROC curves for women were fairly similar from the two instruments, but the ROC curve for men derived from the Carter Center's HRA crossed the chance line, indicating that for some men in the sample, the survey performed no better than chance in predicting 10-year mortality risk (55). Moreover, the authors assert that the Carter Center HRA performed particularly poorly among younger males. Given that there is doubt about how accurate the risk estimation scores may be for young adults, and given that the Army comprises mostly younger males, it may not be advisable to use these risk estimation scores in research.

The Army's experience with the health promotion program and the HRA is sparsely documented, so it is uncertain how effective the HRA was in raising awareness of health risks or in promoting behavior change among soldiers. Although there have not been any longitudinal studies assessing the long-term impact of the program on soldier health, a series of cross-sectional analyses were done for the years 1991-1995 (88). Analysts examined Army HRA responses by year from 1991 to 1995 and compared the results to the Healthy People 2000 objectives and to DoD health promotion objectives. These analyses showed that the Army population achieved some of those objectives (most notably in the areas of fitness, nutrition, use of seat belts, and preventive health services), while their status with respect to other objectives either remained the same or worsened (for example, smoking, prevalence of overweight, and total cholesterol levels). A study by Yore et al. used a combination of cross-sectional and longitudinal analyses to examine health behaviors reported by active-duty Army soldiers (123). They also found that soldiers surpassed Healthy People 2000 goals for fitness and certain dietary objectives, but that they fell short of attaining objectives in the areas of smoking, alcohol use, or safety-related behaviors. It is interesting to note, however, that neither of these assessments relied upon the composite risk scores

generated by the HRA software, but instead analyzed responses to individual items. This is probably the wisest choice of action, as the validation studies reviewed above indicate that the reliability and validity of individual items on HRAs may vary widely, and may negatively impact the overall quality of the risk estimation scores. Furthermore, the work of Rao and Yore show that analysis of responses to individual items may prove useful in assessing changes in health behaviors among soldiers. In the next chapter, we will examine the individual items on the Army's HRA in more detail, and review the formal evidence that exists with respect to their reliability and validity.

CHAPTER 3: THE HRA ITEMS

As noted in Chapter 1, the Army launched its health promotion program in 1987. Between 1987 and 1998 there were at least three distinct versions of the HRA form in use by the Army (see Appendix A). The earliest version of the form we have been able to locate is dated March 1988. We have also located a version dated May 1988, but the only difference between this version and the March version is a change in the title (from Health Risk Appraisal Assessment to the U.S. Army Wellness Check). We believe this to be the version based on the RIWC. The next version is dated August 1989; this represented a major update to the form, as the questions are different and appear in a different order than on the 1988 form. This version does not have a DA form number on it, but says HSC Form 592 (Test) at the bottom. It is not clear how widely this test version of the HRA form may have been used, or for how long. The next version is dated October 1990; the questions on this version are also in a slightly different order and use different wording from the August 1989 version. The last known version of the HRA is dated February 1992, and is a minor update to the 1990 version of the form. The major changes that were made at this time were a change to the Privacy Act statement (making responses optional and allowing soldiers to skip individual items without disciplinary repercussions) and the removal of a skip pattern in the alcohol items.

Because the 1992 version was in use for the longest period of time during the life of the health promotion program, this chapter describes the items on the February 1992 version of the HRA, introduces the major topic areas covered by the HRA health habit items, orients the reader to major issues in assessing health behavior via self-reporting, and reviews what is known about the reliability and validity of the individual items.

The Army's HRA questionnaire comprises 75 items (DA Form 5675, 1 Feb 1992). Items 1-14 collect basic demographic and administrative information (such as rank, branch of service, duty status, and unique identifying information such as name and Social Security Number). Items 15-17 include self-reported anthropometric information on height, weight, and frame size. Items 70-75 gather clinical information (e.g., blood pressure, fasting glucose). The remaining items (items 18-69) form the core of the HRA and ask about health behaviors.

It does not appear that the Army ever published any findings related to the reliability or validity of the HRA questionnaire or any of the items on it. As noted previously, however, some of the items on the HRA appear on other questionnaires and may have been tested for reliability or validity in other settings before being picked up or adapted by the Army for use in the HRA questionnaire. In many cases we could not find any evidence that the exact item had been evaluated with respect to reliability and validity. In some cases, however, we found studies of similar items and have presented data from those studies here, as they are often the only evidence we have to indicate how reliable or valid self-reports of the health habit in question may be. In addition to searching for general articles on reliability and validity of HRA items, we found many studies assessing reliability and validity of items on the CDC's Behavioral Risk Factor

Surveillance System (BRFSS). The BRFSS is one of the most popular and widely used tools for gathering information about health status and health behaviors. Although the mode of administration is different (the BRFSS is a telephone survey, whereas the Army's HRA is a paper-and-pencil questionnaire), and the wording of specific items differs, these studies of the BRFSS items give us at least some rudimentary information about reliability and validity of respondents' answers on items about health behaviors.

There are some important caveats to the interpretation of these studies of items that are similar but not identical to the Army's HRA items. Variations in the findings regarding reliability and validity could be related to sample selection, to the instrument itself, or to the mode of administration, and may be influenced by factors such as the race or ethnicity of the respondent or concerns about anonymity. Several of the studies reviewed below, for example, indicate that an item or group of items may perform differently among people of varying racial or ethnic backgrounds. This could indicate that people responded differently to various translations of the instrument, or that cultural barriers inhibited them from talking freely about certain topics in a telephone interview. Also, as noted in Chapter 2, many reliability and validity studies have restricted their study populations to adults between the ages of 25 and 60, making it difficult to extrapolate these findings to younger adults. Given that the Army is younger than the civilian population at large, the results obtained with respect to reliability and validity in these civilian studies may not be perfectly and directly applicable to the Army. Therefore, because the Army is more ethnically diverse and on the whole younger than the civilian population, the possibility that the quality of the information gathered by the HRA with respect to health behaviors varies with age and among racial or ethnic subgroups should be taken seriously.

As for privacy and anonymity, it is important to bear in mind that the BRFSS is an anonymous telephone survey, whereas the Army's HRA is administered either as a paper-and-pencil questionnaire or as a computer-based survey, and that the respondent is required to provide unique identification information such as a name and a Social Security Number. Various items on the HRA may be construed as sensitive, especially items about risk-taking behaviors such as self-reported suicidal ideation, and alcohol consumption habits or related behaviors such as drinking and driving. Social desirability theory suggests that people tend to minimize or under-report behaviors that are socially unacceptable, and research has shown that respondents are often less forthright about revealing such truths when they cannot do so privately (20, 97). As noted earlier, the Privacy Act statement on the HRA form was changed in 1992 to allow respondents to skip items, but this change notwithstanding, some respondents to the Army's HRA may have feared negative consequences if they admitted to risky or unhealthy behavior. An exploration of the validity of some of these sensitive items appears elsewhere (12, 13).

While it may be tempting to extrapolate from the studies described below, and attempt to make estimates about the reliability and validity of responses on the Army's HRA, the lack of anonymity and unique demographic characteristics of the Army are just two reasons why it would be inadvisable to do so. Even so, the studies described herein summarize the sparse evidence we do have concerning reliability and validity of self-reported health behaviors. The sections that follow highlight areas where the

civilian literature has, in some cases, documented the psychometric properties of these items fairly well. There are other items where evidence from the civilian literature is either less clear, less relevant to the Army's special needs, or suggests that the Army may want to seek alternative items. If, for example, the civilian literature shows that an item performs poorly among a particular racial or demographic subgroup, and if the Army has a large subpopulation of that subgroup (e.g., young minorities), the Army may need to assess whether the item is performing well enough for its purposes, and if not, revise the item or use a different item in surveillance. Researchers using Army HRA data in epidemiologic work need to know more about the reliability and validity of responses so that they can judge whether the information is of sufficient quality for use in their work, or if they need to supplement with physiologic data or adjust for possible misclassification (17). The following summarizes the extent to which HRA items have been studied for reliability and validity. It is organized around major topical area, in the order in which they are presented on the HRA.

EXERCISE

Items 18 and 19 on the Army's HRA ask about aerobic exercise and strength training activities. It appears that these items were adapted from the RIWC instrument. We have not been able to locate any published studies that assess the reliability or validity of these items.

Reliability and Validity

Although the specific items that are on the Army's HRA have not been evaluated with regard to reliability and validity, there have been other studies that give us some idea about how accurate self-reported aerobic activity might be. Table 2 summarizes the results of studies of the test-retest reliability of the physical activity items on the CDC's HRA and on the BRFSS. The CDC's HRA assesses physical activity in one item, asking respondents to categorize their typical activity with one of three categories: little or no physical activity, occasional physical activity, regular physical activity at least three times per week. There is a brief definition of "physical activity" as "work and leisure activities that require sustained physical exertion such as walking briskly, running, lifting and carrying." Smith et al. reported a Pearson's r on test-retest of 0.65 (95% CI: 0.50-0.76) (103). This score failed to meet the authors' *a priori* criterion of 0.8 for determining reliability of items (although r scores of 0.7 or greater are generally considered to be indicative of good reliability) and also showed the lowest correlation of any of the HRA items evaluated in their study. The BRFSS has a more detailed battery of questions on physical activity, but it is questionable whether these more detailed questions garner more precise and replicable responses. Stein et al. evaluated test-retest reliability of the BRFSS physical activity items in a group of 210 respondents from Massachusetts (106). They compared responses in a typical BRFSS sample and in a sample drawn from census tracts with large minority populations, in an effort to determine whether there may be racial or ethnic differences in consistency of self-reported health behaviors. The exercise item they assessed differed slightly from the Army HRA item on physical activity; they categorized respondents as to whether they had or had not performed aerobic activity at least three times per week for at least 20

minutes per occasion in the past month. They found a κ statistic of 0.45 for the total sample (N=210) indicating only fair reliability, but noted that test-retest reliability varied across racial and ethnic subgroups of their study sample (106). A similar study by Shea et al. evaluated test-retest reliability in a sample of respondents from New York State and found an overall κ statistic of 0.65 (N=145), with slightly less variability among racial and ethnic subgroups than observed by Stein et al. (102). Finally, Bowlin et al. evaluated test-retest reliability of various BRFSS items in a rural population (17). They assessed physical activity by asking about the frequency of different types of weekly activity that caused subjects to work up a sweat and documented a κ statistic of 0.60 (N=628). The authors of these three studies thus all documented modest correlation coefficients for self-reported aerobic activity, although there are some apparent variations among racial and ethnic subgroups. All authors acknowledge the limitations of their respective studies, especially in regard to differing response rates among various ethnic groups, and in that there were demographic factors associated with likelihood of response or completion of a second interview.

Table 2. Summary of Studies of Test-Retest Reliability of Self-Reported Physical Activity

Study	Overall Sample		White non-Hispanics		Black non-Hispanics		Hispanics	
	N		N		N		N	
Smith (1989) ^a	338	0.65 ^b						
Stein (1993) ^c		0.45 ^d	75	0.61 ^d	64	-0.07 ^d	45	0.64 ^d
Shea (1991) ^e		0.65 ^d	49	0.57 ^d	43	0.77 ^d	53	0.62 ^d
Bowlin (1996) ^f		0.60 ^d						

^a CDC's HRA item: categorize physical activity as "little or no," "occasional," or "regular physical activity at least 3 times per week."

^b κ statistic

^c CDC's BRFSS item: regular aerobic exercise, defined as "performed an aerobic activity at least three times per week for at least 20 minutes per occasion in the past month."

^d Pearson's r

^e CDC's BRFSS item: regular physical activity in the past month.

^f CDC's BRFSS item: weekly activity to work up a sweat.

In their validation study, Smith et al. compared responses on the HRA physical activity item to information collected via interview, which they used to develop a measure of kilocalories expended in the previous week, using formulas from the Harvard Alumni Activity Survey Scale (104). They then calculated a Pearson's r comparing the answer to the HRA item with the criterion measure of physical activity in the past week. They reported a negative correlation of -0.48 between these two measures. They concluded that the HRA item on physical activity was too frequently inaccurate to be of use in predicting risk.

Finally, we have not found any studies evaluating reliability or validity of reporting of strength training, and thus cannot speak to the quality of data elicited by that item.

Implications for the Army's HRA Data

These results from these studies of self-reported exercise should be used with caution when assessing the performance of the Army's HRA items regarding physical activity. Although the studies reviewed above showed fair reliability, the mean age of

participants in these studies ranged from 34 to 45, and there were differences in consistency of self-reported activity levels among racial and ethnic subgroups. Therefore, it is not clear that we could generalize from these results to the Army, which is a younger and more ethnically diverse population.

We have not found any studies evaluating reliability or validity of strength training and thus cannot speak to the quality of that item. However, recent work to evaluate the Army's performance in meeting Healthy People 2000 objectives noted that more than 95% of HRA respondents reported participating in strength training activities more than once per week (123). Others have noted that when a population is fairly homogeneous with respect to some characteristic, it becomes difficult to use the κ as a correlation coefficient for the reliability of response (21, 22, 106). The Army's HRA item on strength training may produce unstable estimates if used to assess test-retest reliability of self-reported strength training habits.

DIET

The HRA collects nutrition information in two sections. Items 20 and 21 ask about frequency of fiber and fat intake, and appear to be adaptations of items from the RIWC (with the addition of a response category "at every meal"). The CDC/Carter Center's HRA has similar items that request yes/no responses. Item 22 asks if participants salt their food before tasting, and Item 37 asks about frequency of consumption of well-balanced meals; the source of these items is unknown. Item 38 queries about intake of high-sodium foods and appears to have been taken directly from the RIWC.

We could find no evidence that these dietary items from the Army's HRA or their source questionnaires have been evaluated for reliability or validity. There is a vast body of literature surrounding evaluation of nutritional intake, but many of these nutritional surveys are designed to obtain far more detail about intake of specific types of foods, portion sizes, or frequency of consumption. The 1991 version of the BRFSS, for example, asks 13 questions about consumption of different kinds of fatty foods and six questions about consumption of different kinds of fruits and vegetables. It is interesting to note that when Shea et al. did their study to assess the test-retest reliability of the BRFSS, they chose to substitute the standard BRFSS questions on diet with an even more detailed battery of questions (102). We cannot comment on whether the five diet questions on the Army's HRA elicit reliable and valid responses, because they have not been studied. Because there are so few questions, however, at least in comparison to other dietary assessment instruments, the Army's HRA diet questions, as a group, probably do not garner specific enough information to be useful in epidemiologic research. They may, however, be useful in drawing comparisons to the stated objectives on a health promotion agenda (e.g., by comparing to the Healthy People 2000 recommendations about consumption of fruits and vegetables).

STRESS

The Army's HRA includes 14 items on stress and life satisfaction (Items 39-52). Only two items (personal losses in the past year, and general life satisfaction) appear on the CDC/Carter Center's HRA, and one additional item (hours of sleep each night) is similar to an item from the CDC's HRA, with the exception being that the response scales differ between the two instruments. The source of the remaining stress items on the Army's HRA is unknown. Neither the CDC's nor the Army's HRA includes responses from the stress items in the calculation of the respondent's overall risk score.

Stress surveys can take several approaches. Some evaluate the nature or quality of stressors (e.g., life events questions, as an item asking about losses or misfortunes in the past year) (33, 56, 99). Others assess coping strategies, as an item asking about social support (44). Still others seek to understand the respondent's emotional responses to stressors (e.g., anxiety or depression, as in an item asking about experience of prolonged or repeated bouts of depression) (44). Some life events scales evaluate the individual's response to particular types of adverse events (e.g., unemployment or bereavement), while other surveys focus on the cumulative effect of many life events, both pleasant and unpleasant (34). The influence of life events surveys on the Army's HRA is evident from the inclusion of questions about losses and misfortunes in the past year, as well as major pleasant changes in the past year.

It is likely that the people who constructed the Army's HRA wrote new items specifically for the HRA; if so, there is no documentation that any newly constructed items were assessed for reliability and validity. The result is a combination of measures of stress and distress that addresses many of the major thematic areas in the literature on stress and health without actually replicating any of the standard items used on other published surveys that measure stress and distress. Although there is a considerable body of literature evaluating the reliability and validity of various stress or depression scales, it is difficult to apply to the Army's HRA, because most of the published studies of those stress scales assess the reliability and validity of the overall scale, and it would be difficult to parse out the reliability and validity of any individual item, especially when used in another context. Studies evaluating the psychometric properties of the Army's HRA items on stress are necessary before drawing any conclusions about the utility of HRA response data.

MOTOR VEHICLE SAFETY

The HRA contains five items to assess behaviors related to motor vehicle safety. They ask for estimates of the number of vehicle miles traveled (VMT) per year by car and by motorcycle, about typical mode of transportation, the percentage of time the respondent uses a seat belt, and how closely the person adheres to the posted speed limit. When the Army's HRA was first launched, there were two separate questions about drinking and driving: one that assessed driving after drinking and another about riding with a drunken driver. The HRA was revised in October of 1990, however, and these two questions were, unfortunately, combined. So-called double-barreled survey items (those that ask more than one question but only allow the respondent to provide

one answer which, presumably, would apply to both questions) are difficult to analyze. Civilian studies have shown that teenagers often accept rides from a peer who has had too much to drink because they perceive few alternatives to riding with a drunk driver, and will take this risk even though they understand the associated hazards (114). Ride sharing is common on military installations, especially among younger soldiers who may have limited access to privately owned vehicles. An analysis of 1992 respondents to the HRA found that 11% of the nondrivers reported riding with a drunk driver in the past month (11). Being able to analyze the group of people who report riding with a drunken driver separately from the group who report drinking and driving personally would be valuable in furthering our understanding of the social dynamics of these risks.

All of the Army's HRA items on motor vehicle safety appear in some form on the CDC/Carter Center's HRA. The questions about VMT and seat belt use also appear on the CDC's HRA. Although we could find no studies assessing these exact items, there are several studies that have assessed the reliability and validity of self-reported motor vehicle-related behaviors.

Reliability and Validity

The Federal Highway Administration conducts the Nationwide Personal Transportation Survey (NPTS), which has surveyed drivers five times since the late 1960s. The survey collects information on number and purpose of trips, means of transportation, length of trip in time and miles, day of week and month, number of passengers, and other related variables. Military personnel are excluded from the sample, unless they live in civilian housing. The survey gathers several different self-reported estimates of VMT (e.g., odometer readings at 2-month intervals, estimates of miles traveled in a single day). These measures are used to formulate multiple extrapolated estimates of annual VMT, which may then be compared with the respondent's self-reported estimated annual VMT to check for internal consistency. Unpublished data shows discordance between the annualized estimates of VMT and self-reported estimates, and that these variations may be greater among some demographic subgroups than others². Comparing the annualized estimated mileage based on a typical travel day to self-reported estimates of annual VMT, it seems that men tend to overestimate VMT, and that women tend to underestimate VMT but to a lesser degree than men overestimate it. Of particular interest, however, is that both younger men and younger women (aged 16-19) underestimated VMT. These preliminary analyses have some methodological shortcomings (e.g., the NPTS allows proxy reporting, and some measures of mileage estimates are specific to the car, not the driver, which may lead to an underestimate of teen driving if the teen is using a parent's car), but they may cast enough doubt on the quality of self-reported VMT to warrant further validation.

The BRFSS item on seat belt use is worded similarly to the item on the Army's HRA, except that instead of asking respondents to estimate the percentage of time they buckle up, they are asked to categorize their response into one of five discrete

² N. McGuckin, Federal Highway Administration, written communication, October 11, 2001.

categories (always, nearly always, sometimes, seldom, or never). Stein et al. assessed the test-retest reliability of the BRFSS seat belt item in a sample of 210 respondents from Massachusetts (106). Approximately 60% of the people reported always using a seat belt (60.5% at time 1 and 61.4% at time 2). The overall κ statistic for the entire sample was 0.76; among white non-Hispanics it was 0.81 (N=75), among Black non-Hispanics it was 0.77 (N=64), and among Hispanics it was 0.75 (N=45).

Although the reliability of this item appears fairly high, there is contradictory evidence on the validity of self-reported seat belt use. Efforts to validate self-reports of seat belt use have typically employed two types of direct observation. Direct observation is somewhat limited as a validation technique because it must be restricted to daylight hours, captures information on only one instance of seat belt use and thus cannot estimate the driver's typical practices, and produces subjective estimates of a driver's age and ethnicity (83). It is also difficult to determine seat belt use of passengers in the rear seat. The first type of direct observation study compares self-reported state survey data on seat belt use habits with the results of direct roadside observations. These studies have typically concluded that people tend to over-report seat belt use. For example, Robertson et al. compared CDC data on self-reported seat belt use with observations of actual behavior for 13 states and found that the proportion of drivers self-reporting that they "always" or "nearly always" use seat belts was consistently higher than actual behavior; the median difference between observed and self-reported seat belt use was 21.5% (91). An earlier study compared observed and self-reported seat belt use in 15 states and similarly found that self-reports of persons who "always" used seat belts exceeded observed use by 8% (ranging from 11% above observed use to 24% above observed use across states) (36). When investigators included self-reports of persons who "nearly always" used seat belts, the average discrepancy between observed and actual use increased to 27% (ranging from 12% above observed use to 39% above observed use). The chief criticism that has been leveled against this methodology, however, is that the observed and self-reported populations differ. The second type of direct observation method that has been employed to validate self-reports of seat belt use compares self-reports and observed use in the same population. These studies have also, however, reached the same conclusion: that people over-report belt use. For example, researchers in El Paso observed belt use among patrons arriving at gas stations/convenience stores, then approached the drivers to invite them to participate in the study and answer a brief questionnaire that included one item about seat belt use habits (83). The authors note that other studies of similar methodology had documented discrepancies between observed and reported belt use on the order of 6%-14%. In their study, they found a discrepancy between observed and reported seat belt use of approximately 14% in the overall sample. They further noted that whites were significantly more likely than Hispanics to report always wearing seat belts and were significantly more likely to be observed wearing them at the time of the survey. Among the subsamples of white and Hispanic respondents who reported always wearing seat belts, however, whites over-reported use by 21% and Hispanics over-reported use by 27% (a nonsignificant difference between the two racial subgroups).

In contrast to these studies based on direct observation, a prospective cohort study that examined self-reported seat belt use among active duty Army soldiers found that soldiers who reported lower levels of seatbelt usage are at greater risk for motor vehicle-related injury hospitalizations. In this respect, the Army's HRA item has demonstrated good criterion validity (11).

Smith et al. examined the correlation between per capita alcohol sales data and prevalence of self-reported drinking and driving and found a modest, positive correlation between them ($r=0.51$) (105). Per capita sales explained approximately 26% of the prevalence of self-reported drinking and driving. Robertson took a similar methodological approach to validating BRFSS data on self-reported drinking and driving, but with different data sources. He compared responses on the drunk-driving items in the 1988 BRFSS in 19 states with data from the National Highway Traffic Safety Administration's Fatal Accident Reporting System (FARS) (91). The FARS system captures information on nearly all motor vehicle crashes on public roads that result in fatalities; blood alcohol concentration (BAC) data are available for approximately 80% of all crashes. Although the BRFSS and FARS systems are not capturing the same individuals, it could be reasoned that states with a large proportion of people who admit to driving after drinking too much may also be expected to have high rates of fatally injured drivers with illegal BACs. In fact, Robertson documented poor correlations between these two measures; the percentage of BRFSS respondents who reported drinking and driving accounted for only 20% of the fatally injured drivers with illegal BACs.

In a 1982 review of the literature, Midanik examined studies that sought to validate self-reported alcohol-related problems by comparing self-reports to official records (e.g., hospitalizations, arrests for public drunkenness or driving while intoxicated) (78). Results varied depending on the method of interview, the population under study, the referent time frame, and the definitions being used. Over-reporting seemed especially prevalent in clinical samples of alcoholics and less prevalent in general population samples. Midanik reviewed a study by Locander et al. who found that respondents in a general population sample of persons arrested for DWI tended to distort reporting of drunken driving arrests more than other types of arrests. They also noted a strong effect of interview method on quality of reporting, in that respondents were more likely to under-report DWI arrests than other arrests when information was gathered by self-administered questionnaire than through other methods (e.g., face-to-face or telephone interviews). In the only military study reviewed by Midanik, Polich and Orvis found no evidence to suggest that active-duty Air Force service members were under-reporting DWI arrests. Indeed, Polich and Orvis documented a twofold difference in rate of reporting DWI arrests when comparing self reports to official base records, pointing to a possible over-reporting of DWI among this population (85). Anda et al. examined the correlation between self-reports of drinking and driving on the Michigan BRFSS and police reports of motor vehicle crashes (4). They calculated age-, sex-, and region-specific prevalence estimates of self-reported drinking and driving and an injury crash rate (based on police report of whether alcohol was involved in the crash), and found a strong, linear correlation between self-reported drinking and driving and injury crash rates for drinking drivers (Pearson's $r = 0.96$). Criterion validity of self-reported

drinking and driving was demonstrated in a civilian study that linked motor vehicle records of traffic violations and crashes with health risk survey results for members of a large health maintenance organization (111). The investigators found that respondents who reported drinking and driving had increased risk of traffic violations, and an increased risk of motor vehicle crashes (although this result reached statistical significance among women only). A prospective cohort study by Bell et al. found that soldiers who reported drinking and driving and typical alcohol consumption in excess of 21 drinks per week on the Army's HRA were at increased risk of sustaining a subsequent hospitalization for a motor vehicle injury (hazard ratios of 1.45 and 1.98, respectively) (11).

The HRA items on seat belt use and drinking and driving may be most useful when used in combination with other items as a proxy for risk-taking behavior. Civilian studies have shown that, especially among young drivers, seat belt nonuse clusters with other types of risky behaviors such as driving after drinking too much, driving after using marijuana, speeding for the thrill of it, and having had a driver's license suspended (10, 65, 86). A field study of nighttime drivers in Minnesota found that drivers with BACs $\geq$ 100 mg/dL were substantially less likely to be wearing a seat belt than drivers with lower BACs (51). A study of Army soldiers who responded to the first version of the HRA (May-June 1989) compared health habits of 428 aviators and 899 nonflight personnel with a comparison group of soldiers and with the Army at large (50). They found that aviators were significantly less likely than nonflight personnel to report using seat belts, and that both aviators and nonflight personnel were more likely to drive after drinking or to ride with a drinking driver than either of the comparison groups of military personnel. A 1996 study evaluated responses of all HRA respondents and characterized respondents as hazardous drinkers or nonhazardous drinkers (47). Hazardous drinkers were defined as men who reported consuming $\geq$ 21 drinks or women who reported consuming $\geq$ 14 drinks in a typical week. Hazardous drinkers were less likely to use seat belts and were more likely to exceed the speed limit. Of all the health behaviors studied, hazardous drinking related most closely to driving after drinking. An analysis of Army HRA data for 292,023 soldiers who took the HRA between 1990 and 1998 revealed a similar clustering of high-risk habits among risky drinkers (121). High-risk drinkers (i.e., soldiers who responded affirmatively to two or more of the CAGE items and also reported drinking more than 14 drinks per week and/or driving or riding with a drunken driver at least once in the past month) were less likely to wear seat belts, more likely to report driving over the speed limit, and more likely to smoke than low-risk drinkers.

Implications for the Army's HRA Data

The data from the NPTS on self-reported VMT indicates that the estimates of miles driven gathered by the HRA should probably not be taken as a literal indication of driver exposure. The finding that younger adults were particularly prone to underestimate VMT should be of special concern when analyzing Army data on this variable, as the Army's population is largely comprised of younger males. Although the seat belt item demonstrates fairly high reliability, there is evidence that people are inclined to over-report actual use. Moreover, not all military vehicles have seat belts

available for every seating position in the vehicle, and the HRA item is not designed to assess seat belt use relative to availability of seat belts. For these reasons, it is probably not advisable to rely on responses to the HRA seat belt item as literal indicators of actual seat belt use. The seat belt item may, however, be useful as an indicator of risk-taking propensity. Our own work confirms that soldiers who are heavy drinkers tend to engage in other risky behaviors such as failure to use seat belts, speeding, and smoking (121).

The studies reviewed above indicate that self-reports of drinking and driving behavior may not be reliable enough to use them as measures of exposure, per se. The findings of Polich and Orvis indicating an over-reporting of DWI arrests are interesting; if this group did in fact over-report DWI arrests, they may have done so through confusion over what constituted an arrest (i.e., military vs. civilian arrests); more work is needed to determine how accurately military servicemembers report drinking and driving behavior. As with the studies on seat belt use, however, the HRA item on drinking and driving may be useful as a proxy for risk-taking behavior.

ALCOHOL CONSUMPTION

Items 27-34 on the HRA ask about consumption of alcoholic beverages and alcohol-related problems. In contrast to many other topical areas of the HRA questionnaire, there have been an enormous number of studies evaluating the quality of self-reported data on alcohol consumption and alcohol-related problems. It is beyond the scope of this report to evaluate this literature exhaustively, but what follows is a general overview.

Alcohol intake can be assessed by self-reported questionnaire, face-to-face or telephone interview, or by diary entry. A comprehensive evaluation typically elicits information on the quantity of alcohol consumed and the frequency with which this quantity is consumed. Many evaluations then proceed to ask about alcohol-related health or social problems. Consumption is often measured in two questions asking about volume and frequency; for example, "How often do you drink?" and "How much alcohol do you typically consume on those occasions when you drink?" The principal drawback to this approach is that it does not garner information about variability; some drinking patterns, notably episodic heavy drinking, or so-called binge drinking, are associated with particular adverse health or social outcomes (117, 119). Another drawback to this approach is that it tends to underestimate alcohol consumption. Most respondents simply report the number of drinks they consume on a typical drinking occasion, which may mask episodes of heavier drinking (80). There are other approaches to measuring level of alcohol consumption (e.g., graduated frequencies, recent or typical drinking occasions, social context), but this method, also known as the "usual amount" method, is the one most commonly used in population surveys (80), and most closely approximates the alcohol consumption question on the Army's HRA.

Accuracy of self-reported level of alcohol consumption may be influenced by the design of the instrument, such as length of the recall period, beverage specificity, or mode of administration (49). Moreover, evidence from epidemiologic studies suggests

that the relationships between alcohol intake and various health outcomes can vary substantially among demographic subgroups. Women, for example, are often found to be at greater risk for negative health outcomes related to alcohol consumption, and they may also experience adverse events at lower levels of alcohol consumption than men (59, 113, 124). It has furthermore been observed that women metabolize alcohol differently than men (due to factors such as body weight or lean body mass), leading several researchers to argue persuasively for a gender-specific measure of binge drinking, with a threshold of four or more drinks on one occasion for women and five or more for men (116, 118). Depending upon the research question under study, the implementation of gender-specific questions to assess alcohol intake may be warranted.

Consumption of Alcoholic Beverages

The Army's HRA measures alcohol consumption in one item asking, "How many drinks of alcoholic beverages do you have in a typical week?" A direction line on the 1990 version of the form defines one drink as, "one glass of wine, one can of beer, or one shot of liquor." In 1992, the direction line was revised to "one glass of wine or wine cooler, one can of beer, one shot of liquor, or one mixed drink." The National Institute on Alcohol Abuse and Alcoholism defines a drink as, "one 12-ounce bottle of beer or wine cooler, one five-ounce glass of wine, or 1.5 ounces of 80-proof distilled spirits (53)." Respondents enter their estimate in a two-digit field. This question about typical consumption is preceded by a question about drunken driving or riding with a drunk driver, and then followed by six questions asking about other alcohol-related problems. The 1990 version of the HRA questionnaire had a skip instruction after the consumption item directing respondents to skip the six items on alcohol-related problems if they did not drink. The 1992 version of the form deleted this skip instruction. We have not been able to locate any information documenting the reason for deleting this skip instruction, but its presence on the 1990 version of the form was not entirely appropriate, as the items about alcohol-related problems ask about lifetime incidence of these problems (e.g., "have you ever. . . "). Thus, a subject who had had a drinking problem in the past but no longer drank currently would have inappropriately skipped out of the items concerning alcohol-related problems on the 1990 version of the form. The deletion of the skip instruction from the 1992 version of the form now allows the HRA survey to elicit information on lifetime history of alcohol-related problems from all respondents, regardless of their current alcohol consumption patterns.

The BRFSS, in contrast, assesses alcohol consumption in four questions: (1) have you had any alcoholic beverages in past month; (2) in the past month, how many days per week or per month did you drink alcoholic beverages; (3) on days when you drank, how many drinks did you drink on average; (4) how many times in the past month did you drink five or more drinks on one occasion? Stein et al. examined the test-retest reliability for these items on the BRFSS (106). They calculated the number of drinks a person reported consuming in a month and compared responses between the two administrations of the survey. The Pearson's r for the entire sample was 0.72, but varied among racial and ethnic subgroups studied (0.79 for white non-Hispanics, 0.57 for black non-Hispanics, and 0.60 for Hispanics). In this particular setting, the item

asking about total monthly consumption thus garnered responses that were fairly consistent, although they appeared slightly less reliable among Blacks and Hispanics.

Although there have been few studies of the reliability and validity of specific items on the Army's HRA, the alcohol items have been examined in this regard. Bell et al. analyzed the test-retest reliability of the alcohol items on the Army's HRA among 40,870 nonabstaining soldiers who took the HRA more than once and discovered that the items show good reliability, especially over short intervals (13). These analyses were limited to soldiers who took the HRA twice with a minimum of 7 days between surveys and a maximum of 30 days between surveys. Although Bell et al. found that reliability of the HRA alcohol items declined over time, these decreases in consistency of responses over long intervals could be due to an actual change in drinking behavior rather than to poor reliability of items (60). Bell's work also showed that all of the alcohol items on the Army's HRA demonstrate good internal consistency, with a Cronbach's α of 0.69 (13). A separate analysis of HRA responses taken between 1991 and 1998 confirms these findings (Table 3)³. Measures of reliability are generally good for all items, but especially high for three of the four CAGE items (Cut Down, Eye Opener, and Annoyed) and the item that asks respondents whether they have ever had a drinking problem. The reliability for the continuous measure of drinking quantity (drinks per week) was also good, (Pearson's $r=0.72$) over a short time period (2-30 days).

Table 3. Test-Retest Reliability of the Alcohol Items on the Army's HRA, Among Active-Duty Army Soldiers (N = 766), With 7-30 Days Between Surveys

Item	Measure	Reliability
Drinking quantity	Pearson's r	0.72
Cut down	Cohen's kappa	0.80
Annoyed	Cohen's kappa	0.78
Guilty	Cohen's kappa	0.69
Eye opener	Cohen's kappa	0.79
Friends worry	Cohen's kappa	0.62
Drinking problem	Cohen's kappa	0.76
Drinking and driving	Cohen's kappa	0.70

There are a variety of approaches to validating self-reported levels of alcohol consumption. Smith et al. compared production and distribution statistics in 21 states with self-reported alcohol consumption gathered via the BRFSS in 1985 (105). They used linear regression to explore the relationship between sales of alcoholic beverages on a per-capita basis with self-reported measures of alcohol consumption. There was a strong linear correlation between these two measures ($r=0.81$, $\beta=0.34$; i.e., average per capita increase in consumption was 0.34 gallons for each gallon increase in per capita sales). The authors also examined relationships between per-capita sales of alcoholic beverages and specific drinking behaviors and found linear relationships between these measures (heavier drinking $r=0.74$; binge drinking $r=0.59$; drinking and driving $r=0.51$). Smith et al. concluded that states that had higher per-capita rates of alcohol sales also had higher rates of alcohol-related problems. This method of assessing alcohol

³ Senier L., S.R. Strowman, N.S. Bell. Unpublished research, April 2002.

consumption has its strengths and limitations. Midanik and Room have noted that it is useful in describing trends in the consumption of different types of alcoholic beverages, and in comparing regional consumption patterns, but that, overall, it tends to underestimate individual consumption (80). Interestingly, a similar study of 13 Air Force bases conducted by Polich and Orvis accounted for 83% of the alcohol consumed by comparing base sales records to self-reports of consumption—a far higher percentage of coverage than found in most other civilian studies (85). The authors theorized that this higher coverage rate may have been partially attributable to the sampling frame; most civilian studies that have attempted to compare sales data to self-reports of consumption have failed to capture the heaviest drinkers. Polich and Orvis had a more complete sampling frame and obtained a higher response rate. Also, civilian studies that use this methodology note that incomplete coverage rates may be attributed to wastage, stockpiling, or purchase of alcohol by out-of-state visitors; these phenomena may be less of a factor on a military post.

The CDC's BRFSS shifted from beverage-specific items to grouped-beverage items in the late 1980s. The Army's HRA uses a grouped-beverage item, whereas the CDC and the RIWC HRAs use beverage-specific questions to assess alcohol consumption. Serdula et al. conducted a study contrasting responses on the 1987/1988 BRFSS beverage-specific alcohol items with those on the 1989/1990 grouped-beverage version (101). They noted a decrease in mean levels of alcohol consumption between the two versions of the questionnaire, both in estimating average number of drinks consumed and in classifying drinkers as "heavier" drinkers. They acknowledge a downward secular trend in alcohol consumption during this time period (as evidenced by per capita sales data), but theorize that some of the decline may be attributed to the revised wording of the question. Other studies and reviews have noted that beverage-specific items tend to yield higher estimates of alcohol consumption than grouped beverage questions (49, 80).

The work by Bell et al. described earlier with respect to the reliability of the Army's HRA alcohol items also evaluated the internal and external validity of these items (12, 13). Bell et al. analyzed HRA responses from 404,966 soldiers who took the HRA at least once between January 1991 and December 1998. They dichotomized the Drinking Quantity item (with low-risk drinkers consuming 0-14 drinks per week and high-risk drinkers consuming 15 drinks per week or more,(27)) and the Drinking and Driving item (no exposure versus one or more times per month), and then calculated κ statistics between each of the alcohol-related items. All κ s were positive, although the associations were generally rather weak, ranging from 0.05 to 0.43.

Bell et al. also assessed the external validity of the Army's HRA alcohol items (12, 13). They compared HRA respondents in high and low risk alcohol groups in terms of their risk of one or more subsequent hospitalizations for any of 31 alcohol-related conditions. They evaluated risk for alcohol-related hospitalization over time using Kaplan-Meier survival curves and log-rank tests, constructed for each alcohol item, followed by Cox proportional hazards regression models. The study cohort was followed from the date of their HRA through December 31, 1998, until they experienced an alcohol-related hospitalization, or they left the Army (were censored). They also

compared the risk of discharge from the Army for alcoholism versus other reasons, including honorable discharge. They used univariate logistic regression models to evaluate the relationship between self-reported drinking and risk for alcohol-related discharge as compared to separation from the military for other reasons (including honorable discharge) among soldiers who had both completed an HRA and were subsequently discharged by 1998. Results are shown in Table 4. All of the alcohol items were significant predictors of future alcohol-related hospitalizations. There was a strong linear relationship between self-reported weekly drinking level and subsequent risk for an alcohol-related hospitalization. At greatest risk were those who indicated their friends were worried about their drinking, those who admitted having had a drinking problem, and those who reported consuming more than 21 drinks per week. All measures of self-reported drinking were strongly associated with alcohol-related discharge, and there appears to be a linear increase in risk with successively greater amounts of reported weekly alcohol use. Soldiers reporting they ever had a drinking problem were at approximately five times greater risk for experiencing a subsequent alcohol-related discharge. Believing that friends worry about one's drinking is associated with a five-fold increased risk (RR = 4.92, 95% CI = 4.00-6.04) of discharge due to alcoholism, and reporting feelings of annoyance when others criticize one's drinking is related to a four-fold increased risk (RR=4.36, 95% CI = 3.71-5.13).

Table 4. Associations Between Self-Reported Alcohol Use and Subsequent Adverse Health and Occupational Outcomes Among Active-Duty Army Soldiers Taking the HRA, January 1, 1991 – December 31, 1998

Alcohol Variable	Alcohol-Related Hospitalization ¹ N=404,966		Discharge From the Military for Alcoholism ² N = 222,843	
	Hazard Ratio	95% Confidence Interval	Relative Risk	95% Confidence Interval
Drinking Quantity	1.04	1.04, 1.04	1.04	1.03, 1.04
Drinking Group				
1-7 Drinks/Week	1.19	1.12, 1.27	1.38	1.17, 1.63
8-14 Drinks/Week	2.16	1.98, 2.35	2.44	1.98, 3.01
15-21 Drinks/Week	3.23	2.86, 3.65	3.14	2.34, 4.21
>21 Drinks/Week	6.36	5.79, 6.99	6.04	4.83, 7.56
Heavy Drinking	2.27	2.15, 2.40	2.37	2.07, 2.70
Cut Down	2.94	2.78, 3.12	2.56	2.22, 2.94
Annoy	4.27	3.99, 4.57	4.36	3.71, 5.13
Guilty	3.67	3.44, 3.91	3.07	2.61, 3.61
Eye Opener	3.79	3.51, 4.09	3.74	3.14, 4.46
CAGE	3.94	3.71, 4.19	3.57	3.07, 4.15
Friends Worry	6.24	5.74, 6.77	4.92	4.00, 6.04
Drinking Problem	5.92	5.52, 6.34	4.94	4.16, 5.88
CAGE2	4.00	3.76, 4.25	3.65	3.15, 4.22
Drink and Drive	2.11	1.97, 2.26	2.21	1.88, 2.58

¹ Single variable Cox proportional hazards models. Diagnoses include any of 31 alcohol-related conditions found in any of the eight possible diagnostic fields (primary, secondary, etc.). Study population includes all first-time HRA survey takers who completed at least one HRA between 1991 and 1998.

² Single variable logistic regression models. Discharge for other reasons includes honorable discharge. Study population includes those who both completed an HRA sometime between 1991 and 1998 AND were discharged from the Army by 1998.

Implications for the Army's HRA Data

Although the study by Shea et al. indicates that self-reports of alcohol consumption may exhibit good reliability, the validation studies reviewed above suggest that the Army's HRA question on alcohol consumption may produce underestimates of actual consumption. First, the item has a two-digit response field, limiting the maximum reportable number of drinks per week to 99 (37). Our work using HRA responses from 1991 through 1998 shows that during this time a sizable number of soldiers reported very high levels of weekly drinking; soldiers in the top percentile reported routinely consuming more than 30 drinks per week suggesting that truncating response options may undercount the true upper level of weekly alcohol consumed by some soldiers (12). Anecdotal accounts and epidemiologic research on drinking behavior in the Army have documented that many soldiers drink very heavily (23, 85); there may, in fact, be soldiers who routinely consume more than 99 drinks per week, but the HRA is not designed to capture information on these individuals. Second, as reviewed above, the grouped-beverage item may lead soldiers to underestimate levels of consumption. Third, the Army's HRA captures information on quantity alone, and does not produce an independent estimate of frequency. Fourth, there is no measure of binge drinking (a gender-specific measure of this type of hazardous drinking pattern would be optimal). Finally, the HRA is not taken anonymously, and soldiers may be motivated to under-report actual consumption. Our work has shown that most soldiers do not skip the potentially sensitive alcohol items. Of the soldiers who do skip items pertaining to alcohol use on the HRA, there is a slight tendency to be older (i.e., 36+ years or older), African-American, and of upper enlisted ranks. While it may underestimate the actual amount of alcohol consumed, the HRA nevertheless elicits a wide range of responses. In addition, greater reported alcohol usage has been shown to predict alcohol-related health and occupational problems (13). That the alcohol items show a positive and linear relationship between adverse consequences and successively greater levels of alcohol consumption suggests that even though respondents are not reporting their behaviors anonymously, the alcohol items are still capturing enough variation in consumption patterns that they are useful in epidemiologic research projects that seek to link drinking with adverse health and occupational outcomes.

The CAGE

To assess alcohol-related problems and potential dependent drinking, the Army's HRA uses the CAGE questionnaire. It comprises four questions, "have you ever felt you should cut down on your drinking, have people annoyed you by criticizing your drinking, have you ever felt guilty about your drinking, and have you ever had a drink first thing in the morning to steady your nerves (eye opener)?" The CAGE was developed in the late 1960s as a brief alcohol-screening instrument and was first used to identify alcoholics and heavy drinkers in a general hospital population (46). The authors of the Army's HRA made two small changes to the wording of the questions. In the question on cutting down, they substituted "should" for the more formal "ought to," and in the question on annoyance, they added the word "ever." The authors of the original questionnaire note, however, that, "physician(s) in clinical practice (may) paraphrase the four questions to suit the occasion without significantly altering their

validity (46).” Although the CAGE has certain limitations (e.g., it does not capture current drinking practices, and does not perform with uniform reliability across various demographic groups (28, 32)), it is generally acknowledged to be an easy-to-use and sensitive instrument in detecting alcohol dependence or alcohol-related problems.

The CAGE questionnaire has been studied extensively in the diagnosis of hazardous drinking and alcohol dependence. The original authors tested the questionnaire in a group of 166 male patients admitted to an alcoholism rehabilitation center (46). They sorted the patients into three groups (acknowledged alcoholics, acknowledged heavy drinkers, denied alcoholics) and compared their responses on the CAGE to those of a group of 68 nonalcoholic, nonabstaining male hospital patients. A positive response on one question captured 100% of both the acknowledged alcoholics and the acknowledged heavy drinkers, and 97% of the denied alcoholics, but also captured 18% of the nonalcoholic controls. Raising the cutpoint to two affirmative responses still captured 100%, 97%, and 92% of these three groups of drinkers, respectively, yet captured only 4% of the nonalcoholics. The original author asserts, “the existence of even one affirmative response to the four questions call(s) for further investigation and the suspicion of alcoholism until proved otherwise (46),” although many studies have defined a positive response as two affirmative answers.

Mayfield et al. validated the CAGE among patients hospitalized on a psychiatric ward of a Veteran’s Administration hospital (74). They found that the CAGE had poor sensitivity if positive responses were required on all four questions, but that it had good predictive power at two- and three-question cutoffs ($r=0.89$ for both). The correlation coefficients for the four items were as follows: cut down, $r=0.88$; annoy, $r=0.60$; guilty, $r=0.89$; eye opener, $r=0.83$; suggesting that in this population, the annoy question had the least predictive power. The Mayfield study population was predominantly male (99%), white (77%), middle aged (63% of the sample was between ages 35 and 55), and married (60%). Bush et al. tested the CAGE in a sample of 518 consecutive admissions to orthopedic and medical services of a community hospital and found it had a sensitivity of 85% and specificity of 89% in detecting alcohol abuse or alcoholism (26). These results are similar to those found among patients attending a primary care clinic in London (68). Pileire found that the CAGE identified 74% of moderate drinkers and 94% of excessive drinkers (84).

Other studies conducted in more diverse populations, however, have had less consistent results. Cherpitel et al. have conducted numerous studies evaluating the performance of various rapid screening tools, including the CAGE, in racially and ethnically diverse populations in different parts of the country. They found that sensitivities varied in different racial and ethnic subgroups, in different parts of the country (even when controlling for race and ethnicity), and that sensitivities were consistently lower among women, whites, and injured persons (28-31). Among the racially and ethnically diverse populations taking one of several rapid screening assessments in an emergency room, Cherpitel et al. concluded that none of the instruments evaluated, including the CAGE, detected both dependence and hazardous drinking as well as they detected dependence alone (30). The CAGE has been demonstrated to have poor sensitivity and specificity in elderly populations (1, 73), but

Adams et al. found that sensitivity and specificity were improved if the CAGE was supplemented with items about frequency and quantity of consumption (1). Studies seem to indicate that the CAGE does a fair job of identifying people with advanced alcohol dependence (or who were at one time alcohol dependent), but misses a substantial group of people whose drinking is problematic, particularly the elderly, women, and whites (95). On the other hand, Thompson et al. linked health survey data with official records of traffic violations and motor vehicle crashes and found that the risk of traffic violations was significantly elevated for women but not men (111). This may suggest that although the CAGE may not identify all problem drinkers in all contexts, it may do a better job of predicting certain adverse alcohol-related outcomes among women than men.

Fertig et al. correlated CAGE scores with self-reported hazardous drinking in a group of Army soldiers who took the HRA (48). Hazardous drinking was defined as ≥ 21 drinks per week for men and ≥ 14 drinks per week for women. They calculated sensitivities and specificities for the CAGE and a modified CAGE (the CAGE with the addition of the item about drunk driving and ever having had a drinking problem). At a cutpoint of one, the modified CAGE showed greater sensitivity than the standard CAGE (81% vs 72%); sensitivity for both versions dropped dramatically at a cutpoint of two (41% for the modified CAGE and 54% for the standard CAGE). Furthermore, the authors noted demographic differences in predictive abilities of the two versions of the CAGE. On both versions of the instrument, the cutpoint of one was more predictive of potentially hazardous drinking for women, for officers, for never-married persons, and for younger soldiers. These findings are corroborated by Heck and Williams, who found that using the CAGE at a lower cutpoint or in combination with items about other risky drinking practices or self-reported alcohol consumption was more likely to detect hazardous drinking in a population of college students (61, 62).

Many researchers are inclined to be skeptical about the quality of self-reported data concerning alcohol consumption and alcohol-related problems. Because of these concerns about response bias, there has been great interest in identifying biochemical markers to detect hazardous drinking practices. Researchers have experimented with Breathalyzer tests, urine tests, and sweatpatches. Although some of these methods are useful in detecting acute intoxication, many of these methods are limited in their ability to detect typical drinking practices. It has also been noted that individuals metabolize alcohol at different rates due to factors such as body composition and liver damage, and it is unclear how these variables may influence the validity of biochemical tests. Moreover, some of these methods are expensive and intrusive, and while they may have the cachet of technology behind them, they are not without their own methodological limitations with respect to sensitivity, specificity, and predictive ability. In her reviews of the topic, Midanik has cautioned against adopting these measures as a “gold standard” until more work has been done to assess their utility (78, 79).

More recent work has focused on other biochemical markers such as carbohydrate-deficient transferrin (CDT), γ -glutamyl transferase (GGT), and mean corpuscular volume (MCV). Numerous studies have assessed the performance of these biochemical tests in detecting hazardous drinking relative to rapid assessment

tests such as the CAGE. Bisson and Milford-Ward compared three screening questionnaires and five biochemical tests to determine their sensitivities and specificities in detecting alcoholism (14). Cases were British soldiers under the age of 30 who were admitted to an alcohol treatment unit; controls were also young British soldiers who were selected from nearby Army units. All three of the screening questionnaires tested exhibited superior sensitivity in detecting alcoholics; the CAGE correctly identified 93% of the cases. In general, the five biochemical tests showed greater specificity, meaning that they were less likely to generate false positives than the questionnaires, but they all demonstrated unacceptable low sensitivity (ranging from 4% to 26%). Wetterling et al. took a similar approach in a general patient population and documented disappointingly low sensitivities for all of the assessment methods they evaluated, whether at detecting alcohol dependence or hazardous drinking (120). The questionnaires, however, demonstrated superior specificity and positive predictive value over any of the biochemical tests. Lee and DeFrank compared three rapid screening assessments, two biochemical markers, and self-reported quantity of consumption among students at an allied health school (70). They calculated Spearman rank correlation coefficients between all the measures they assessed and found that the CAGE correlated significantly with self-reported consumption of alcohol for men but not for women. Only one of the two biochemical tests, MCV, was significantly correlated among both men and women, but it did in fact show a higher degree of correlation with self-reported alcohol consumption than did the CAGE. Lee and DeFrank note, however, that both the CAGE and MCV showed much higher correlations among men than among women, suggesting that these tests may not be adequate in detecting problem drinking among women. More recent work by Aithal et al. indicates that although CDT had fairly good sensitivity and specificity overall (69% and 81%, respectively), and although these values were comparable to the sensitivities and specificities obtained by the CAGE, the CAGE had better positive predictive value than the CDT (2). Moreover, the sensitivity of the CDT test varied substantially between men and women (80% vs. 33%). In her review of the literature on this topic, Midanik notes that although the biochemical markers appear to have the cachet of a “gold standard,” in that they are assessing self-reports against apparently objective data, it seems clear from the studies reviewed here that it is premature to conclude that laboratory markers are superior to self-reports in detecting alcoholism or problem drinking.

Implications for the Army’s HRA Data

The studies reviewed above show that the CAGE questionnaire does a fair job of detecting problem drinkers, especially in combination with other alcohol items on the HRA. However, there are potential challenges in using the HRA alcohol items for epidemiologic research. For example, Steinweg and Worth reported that the sensitivity of the CAGE in detecting alcoholism was attenuated when it was preceded by items about drinking quantity and frequency, as is true on the Army’s HRA (108). Furthermore, slight changes in the way respondents were queried about their alcohol consumption in different versions of the HRA survey may bias temporal analyses of trends in drinking. The deletion of skip instructions between the 1990 and 1992 versions of the questionnaire has been found to impact response rates for the CAGE

and other alcohol items (12). It is not clear when the Army began using the 1992 version of the questionnaire; it is dated 1 Feb 1992, but as we stated in Chapter 1, we do not know what instructions were given about transitioning to the new version of the form, and it is likely that both versions were in use concurrently for a significant period of time after February 1992. This issue may render the information on alcohol-related problems documented by the HRA around the time of the change in versions difficult to interpret.

Other Alcohol-Related Problems

The Army's HRA includes two supplemental items asking about alcohol-related problems, which may have been adapted from the Michigan Alcoholism Screening Test (MAST), although the Army uses a slightly different wording. Item 33 asks whether one's friends worry about one's drinking. Item 34 asks whether the respondent has ever had a drinking problem. As noted above, the CAGE does not inquire specifically about current drinking patterns. Respondents may truthfully produce seemingly contradictory responses by reporting a current consumption level of zero but answering yes to two or more CAGE items. Item 34 may have been added to the HRA to improve its ability to identify abstaining alcoholics. The reliability and validity of these items in a military population is unstudied and therefore uncertain.

DIABETES

The HRA includes one item that asks respondents if they have ever been told that they have diabetes and requests a yes or no response. It appears this item was picked up from the CDC/Carter Center's HRA, and it is the same as the one asked on the 1991 version of the BRFSS. Both the CDC/Carter Center and BRFSS versions give possible responses as yes or no, except the BRFSS also includes an option for don't know/not sure or refused. We have found no studies assessing the reliability and validity of the Army's or the CDC/Carter Center's HRA items, but there have been several reliability and validity studies of the BRFSS item.

Reliability and Validity

Stein et al. documented the test-retest reliability of this item in a group of 210 adults from Massachusetts (106). The approximate prevalence of diabetes in this population was 6.2% at the first survey, although it varied among racial and ethnic subgroups, and among Black non-Hispanics it was 10.9%. The test-retest κ statistic for the entire sample was 0.82; among Whites it was 0.85, among Black non-Hispanics it was 1.00, and among Hispanics it was -0.03. The authors note that the result for Hispanics was not statistically significant, probably due to the very low prevalence of self-reported diabetes among this subgroup (4.4% at time 1, and 2.2% at time 2). Shea et al. assessed test-retest reliability in a triethnic population among 145 residents of New York State (102). They calculated an overall κ statistic for the entire sample of 0.60. Among whites (N=49) the correlation score was 1.00; among Blacks (N=43) it was 0.36 and not significant; and among Hispanics (N=53) it was 0.65. Brownson et al. evaluated test-retest reliability in a group of BRFSS respondents in Missouri (24). Only

a small proportion of the respondents (7%) reported having been told that they have diabetes; the κ statistic reported in this study was 0.86. Among subjects who reported having been told they were diabetic, when asked, 92% of them gave consistent responses between the two administrations of the survey (Pearson's $r = 0.99$).

Bowlin et al. conducted a validation study in a sample of 628 BRFSS participants in three rural communities in upstate New York (16). After participating in a telephone survey, subjects were invited for a free physical examination. Investigators compared yes responses on the BRFSS item with a fasting blood glucose test of ≥ 140 mg/dL, a commonly accepted threshold for a diagnosis of diabetes. Sensitivity of self-reported diabetes (that is, the proportion of people correctly classifying themselves as diabetic) was 67% for men and 80% for women; the prevalence of self-reported and actual diabetes for men and women in the overall sample agreed closely (3% and 4% for men, respectively, and 5% and 5% for women, respectively). In a follow-up analysis to the same study, Bowlin et al. sought to determine whether combining repeated measures for a factor improved the validity of the measurement by adjusting for random error (17). Subjects were interviewed by telephone and then invited for a free physical examination. Upon presenting at the clinic, they were reinterviewed and underwent a number of physiologic tests, including blood testing for fasting blood glucose. They documented a κ coefficient for the overall test-retest assessment of 0.79 (95% CI: 0.67-0.91). The analysts experimented with three different methods of combining multiple measures of the self-reports of the risk factor in the telephone and clinic interviews. The strict combination defined the risk factor as present when both the telephone and the clinic interview were positive and absent in other combinations; the loose combination defined the risk factor as present when either the telephone or clinic interview was positive and absent only when both interviews were negative; and the concordant combination only used answers that were the same in both the telephone and clinic interview, whether positive or negative, and discarded discordant pairs. None of these methods improved the sensitivity, specificity, or positive predictive value of the self-reported measure of diabetes (75%, 98%, and 48%, respectively). They also compared the five different methods of self-reporting (i.e., telephone interview, clinic interview, and the strict, loose, and concordant combinations) to compare their relative efficiency in determining the prevalence of diabetes reported in this sample. They found that all of the methods produced similar estimates of the prevalence of diabetes when compared to objective measurements of fasting blood glucose.

Implications for the Army's HRA Data

These studies seem to indicate that the self-reported measure of diabetes on the Army's HRA may demonstrate fairly good reliability among whites, but questionable reliability among other ethnic groups. This item has shown respectable validity. It should be noted that this item asks whether the respondent had ever been told that they have diabetes, not whether they currently have diabetes. This may result in an artificially high rate of reported diabetes, as women who were told that they had gestational diabetes during pregnancy may truthfully answer yes to this item, even though their diabetes was resolved at the conclusion of their pregnancy. This specific issue has evidently not been explored in the civilian literature.

HYPERTENSION

The Army's HRA asks one yes/no question about whether the respondent is taking medication to control hypertension. This item could have been taken either from the RIWC or the CDC/Carter Center HRA, as similar items appear on both instruments. The RIWC also includes an item asking whether the respondent has been told within the last 5 years that their blood pressure was either high or borderline high; the CDC/Carter Center HRA and the Army's HRA do not include this additional item. The 1991 version of the BRFSS asks a series of four questions establishing first whether or not the respondent has high blood pressure, and then asks, "Is any medicine currently prescribed for your high blood pressure?"

There have been few studies of the reliability and validity of self-reported utilization of antihypertensive medications, and most such studies have assessed this behavior in conjunction with other variables of interest, such as self-reports of diagnosis of hypertension and compliance with medication protocols. As reviewed in Chapter 2, Smith et al. found that respondent's self-reports of hypertensive status is typically poor enough to compromise the calculation of accurate risk scores (103, 104). Studies of the BRFSS have demonstrated good reliability of self-reports (24, 102, 106), but two validation studies we reviewed have indicated that respondents cannot accurately report whether their hypertension is adequately controlled (17).

It is unfortunate that the Army's HRA contains only one item asking about use of medication to control hypertension. Even though studies show that self-reported hypertension may not be of sufficient quality to assist in HRA risk score calculations, if the Army's HRA had included other items about diagnosis of hypertension or about compliance with medication regimens, survey responses could have been used to compare Army respondents to population or group norms, or to objectives on health promotion agendas (e.g., Healthy People 2010), or to guide the development of interventions. Without an item asking about diagnosis of hypertension, for example, we cannot know how many soldiers are unaware whether they have the condition (and thus plan for screening initiatives). Carefully designed questions about compliance with antihypertensive measures could have informed the design of interventions and possibly identified individuals who needed assistance with compliance.

There is an opportunity to use Army HRA data to validate this item, by comparing the self-report item about taking antihypertensives to the HRA item that documents blood pressure. A validation study of this sort would have some important caveats and limitations, however. First, although the standard operating procedure for administering the HRA called for measuring the respondent's blood pressure, there are doubts about whether it was universally measured or occasionally estimated based on respondent self-report. One of the first HRA project officers noted in examining HRAs taken during the first year of the program that the distribution of blood pressures was stepped at increments of 5 mm Hg, suggesting that it may have been self-reported rather than measured with a sphygmomanometer.⁴ The best approach to a validation study using

⁴ MAJ (ret) Ken Bush, personal communication, July 10, 2002.

Army HRA data may be to evaluate only HRAs from respondents who took it as part of the Over-40 Cardiovascular Screening Program. The Over-40 screening included a clinical exam by a physician, and we may perhaps have greater confidence that the blood pressure readings recorded on those surveys are accurately measured and not based on participant self-report. Second, the reading that appears on the HRA is but a single measure of blood pressure; a true diagnosis of hypertension requires multiple readings over a period of time (e.g., the average of two measures taken 5 minutes apart over two office visits (64)). A single measure of blood pressure may vary from a person's typical or true blood pressure for a variety of reasons. Finally, the clinical definition of hypertension has changed several times in the past 20 years. For example, in 1999, the therapeutic threshold was lowered to 140/90mm Hg (64). A validation study that examined HRA data gathered over many years would need to account for these changes in treatment practices. A soldier who took the HRA in the late 1980s and who had a blood pressure of 140/90mm Hg may not have had a prescription for antihypertensives because he or she didn't meet the clinical definition of hypertension in operation at that time.

TOBACCO

The HRA asks four questions about cigarette smoking and three questions about other forms of tobacco use (e.g., cigars, pipes, smokeless tobacco). It does not appear that the Army has ever evaluated the reliability and validity of these smoking items, but the items concerning cigarette smoking are similar to those on HRAs used in the civilian world, which have been evaluated in other contexts.

Questions 53, 54, and 55 inquire about cigars, pipes, and smokeless tobacco, with the respondent entering the number of these used per day (1-10 cigars, 1-10 pipes, and a 2-digit response field for smokeless tobacco). We have not found any studies evaluating reliability or validity of these items. Historically, cigar smoking had declined in popularity in the United States over the last half of the twentieth century, and although there has been a recent surge in the prevalence of cigar smoking, many national health surveys do not ask specific questions about cigar consumption (6). Recent work by Sanchez and Bray to document prevalence of smoking behavior among members of the armed forces confirms a marked increase in the prevalence of past-year cigar/pipe smoking in the past 5 years, in spite of a decline in past-month cigarette smoking over the past 2 decades (93). Indeed, prevalence of past-year cigar/pipe smoking in the armed forces exceeded prevalence of past-month cigarette smoking in 1998 for the first time (32.6% vs. 29.9%, respectively).

Item 56 asks cigarette smokers to state their smoking status (e.g., current, ex-smoker, or never smoker) and is similar to an item on the CDC HRA and the CDC/Carter Center's HRA. The Army's HRA, the CDC's HRA, and the CDC/Carter Center's HRA all have items asking former smokers how long it has been since they stopped smoking, in years. The RIWC combines the smoking status and quit status items into one question that asks whether the person currently smokes (with possible responses of yes; no, quit in the last 6 months; no, quit more than 6 months ago; and no, I never smoked). The BRFSS does not ask people to categorize themselves as

current, former, or ex-smokers, but instead asks if they have smoked 100 cigarettes in their lifetime; if they reply yes, they are asked follow-up questions about their current smoking status and habits. All of these HRAs ask about the number of cigarettes smoked per day, and all except the RIWC give a 2-digit response field. The RIWC gives five response levels (don't smoke, less than half-pack per day, half pack per day to one pack per day, one to two packs per day, and two or more packs per day). Thus, the HRAs reviewed all capture information on the respondent's smoking status, an estimate of how long it has been since they stopped smoking, and an estimate of the number of cigarettes smoked per day.

Reliability and Validity

The 1993 study by Stein et al., described elsewhere in this report, evaluated test-retest reliability of the cigarette items in a sample of respondents to the BRFSS in Massachusetts (106). They assessed the reliability of self-reports of current smoking with a κ statistic of 0.83, and calculated a Pearson's r of 0.73 for the item about the number of cigarettes smoked per day. These findings are strikingly similar to those found by Shea et al. in a triethnic population in New York State (102). Shea et al. found a κ statistic of 0.85 on the current smoking item and a Pearson's r of 0.78 with respect to the number of cigarettes smoked per day. Brownson et al. documented a κ statistic of 1.00 in a sample of respondents to the Missouri BRFSS (93% white) (24). They also evaluated the test-retest reliability for the number of cigarettes smoked per day, and calculated a Spearman rank correlation coefficient of 0.85. It should be noted, however, that Shea and Stein both undertook subanalyses to evaluate reliability of reporting among racial and ethnic subgroups and found substantial variations (see Table 5). The differences observed among racial and ethnic subgroups may be due in part to the small number of responses available for analyses.

Table 5. Summary of Studies Evaluating Test-Retest Reliability of Self-Reports of Current Smoking Status and Number of Cigarettes Smoked Per Day

Study	Overall Sample	White Non-Hispanics	Black Non-Hispanics	Hispanics
Current Smoking Item				
Stein et al. (1993)	0.83 ^a	0.90 ^a	0.79 ^a	0.85 ^a
Shea et al. (1991)	0.85 ^a	0.94 ^a	0.90 ^a	0.61 ^a
Number of Cigarettes Smoked Per Day				
Stein et al. (1993)	0.73 ^b	0.63 ^b (N=18)	0.83 ^b (N=13)	0.70 ^b (N=10)
Shea et al. (1991)	0.78 ^b	0.89 ^b (N=9)	0.54 ^b (N=15)	0.95 ^b (N=5)

^a κ statistic

^b Pearson's r

Anda et al. compared self-reports of smoking behavior gathered by telephone and in-person interviews in the state of Michigan in order to determine whether the two methods produce different estimates of the prevalence of these health behaviors (4). The two methods produced very similar estimates of the prevalence of smoking, with the telephone interview being 2% smaller among men and 1.3% smaller among women as compared with the in-person interview. Arday et al. compared the prevalence of self-reported smoking data gathered on the BRFSS with the Current Population Survey (CPS) (5). Conducted by the Census Bureau, the CPS includes the same smoking

items that are on the BRFSS, but includes households without telephones. Arday et al. evaluated the prevalence rates of smoking behaviors that were produced from these two surveys by state for 1985, 1989, and 1992/1993 in order to determine whether there were systematic differences between them, or whether the estimates of smoking behavior varied from state to state or over time. The BRFSS produced estimates of smoking prevalence that were lower than those produced by the CPS; although these differences were not large (approximately 2%), they were statistically significant. Most of the differences between these two surveys were the result of lower estimates of smoking prevalence among men (as compared to women), and among Blacks (as compared to whites or Hispanics).

A similar study compared estimates of self-reported smoking behavior between the BRFSS and data from the Stanford Five-City Project Survey (FCPS) (63). The FCPS collects self-reported information on smoking status and number of cigarettes smoked per day, and validates this information via a saliva thiocyanate pipeline test and a test for exhaled carbon monoxide. Self-reported data from the two surveys produced similar estimates of current smoking and the mean number of cigarettes smoked per day, with no statistically significant differences between the two surveys for the overall sample or for any of the individual communities that were analyzed separately.

Luepker et al. conducted a test to validate self-reported smoking behavior among young adults with a saliva cotinine test (72). Subjects were identified by a telephone survey, and then recruited for an in-home interview (a saliva specimen was collected at this interview). Subjects were classified as nonsmokers, smokers, short-term quitters, and long-term quitters. The authors found that the telephone survey underestimated smoking by approximately 3%-4%, and overestimated nonsmoking. This variation was driven, in part, by people who reported not smoking on the telephone interview but admitted smoking during the home interview and, in part, by people who reported different quit statuses in the telephone and in-person interviews (e.g., identified themselves as a long-term quitter on the telephone survey, but as a short-term quitter at the in-person interview). Luepker et al. were unable to draw any firm conclusions about prevalence or duration of smoking cessation, as the self-reported quit data were unstable, possibly because of relapse or through inaccurate self-reporting. The small size of their sample (N=359) probably prohibited conducting any subanalyses to determine whether accuracy of reporting varied by age, gender, or race.

As described in other sections of this report, Bowlin et al. conducted a study to evaluate the reliability and validity of self-reported cardiovascular risk factors gathered on the BRFSS (16, 17). Subjects were recruited from three rural counties in New York State and, after completing the telephone survey, were invited in for a clinic exam. Upon presentation at the clinic, they were reinterviewed, and a number of physiologic tests were performed to validate their self-reports, including a test to evaluate exhaled carbon monoxide (CO). These different methods produced different estimates of the prevalence of current smoking status, with the CO test consistently producing higher estimates than self-reported data (16). The self-reported estimate of smoking among men was 22% vs. 28% confirmed by CO test; the self-reported prevalence of current smoking status among women was 26% vs. 30% confirmed by CO test. These

discrepancies were greatest among men aged 30-39 (30% self-report vs. 39% CO test) and among women aged 20-29 (16% self-report vs. 23% CO test). The current smoking status item exhibited very good reliability ($\kappa = 0.92$), and there was also a very high level of agreement between interviews on self-reported number of cigarettes smoked per day (intraclass correlation coefficient = 0.80) (17). With respect to validity, the current smoking status item performed fairly well for the overall sample, although it seemed slightly less sensitive among men than among women (78% vs. 86%), and there were slight variations among age-specific groups, with the lowest sensitivities being documented among elderly men (sensitivity = 50%) and among younger women (aged 20-29, sensitivity = 0.67) (17). Specificity was high (> 90%) for all gender and age specific subgroups.

As described previously in this report, Bowlin et al. further sought to determine whether combining repeated measures of a self-reported health habit improved the validity of the measurement by adjusting for random error (17). The analysts experimented with three different methods of combining multiple measures of self-reporting of the risk factor in the telephone and clinic interviews. The strict combination defined the risk factor as present when both the telephone and the clinic interview were positive and absent in other combinations; the loose combination defined the risk factor as present when either the telephone or clinic interview was positive and absent only when both interviews were negative; and the concordant combination only used answers that were the same in both the telephone and clinic interview, whether positive or negative, and discarded discordant pairs. The sensitivity of all of these methods was high for self-reported smoking status (> 80%). The telephone interview alone had a sensitivity of 82%, compared to the clinic interview, which had a sensitivity of 87%. Specificity was $\geq 95\%$ for all of the methods tested. Self-reported smoking status and the objective test yielded differing estimates of the prevalence of smoking behavior for all of the different combinations of telephone and clinic interviews. The self-reported prevalence of smoking behavior was consistently lower than that obtained by the CO test, often by 3%-4%. The investigators had hypothesized that combining measures of self-reported behavior may have increased the sensitivity or specificity of these items, or may have produced self-reported estimates of behavior that were closer to those obtained by objective tests, but the gains in validity with respect to smoking status were marginal. They conclude that although the items demonstrated fairly high reliability, the items exhibit only fair validity, especially insofar as they produce under-reports of true smoking status.

Robbins et al. established the criterion validity of the smoking items on the Army's HRA in a prospective cohort study of 87,991 soldiers, by demonstrating that current smokers incurred more hospitalizations and more lost workdays for a wide variety of health problems (90).

Implications for the Army's HRA Data

Although the Stanford FCPS study and the Arday study do not validate the exact items that are on the Army's HRA, the fact that similar items are producing reliable estimates in a variety of settings suggests that the Army's HRA items may also produce

reliable results. The high degree of correlation found by Shea and Stein in their overall samples is also suggestive that these items may have good reliability, although their findings that reliability of self-reporting may vary among racial and ethnic subgroups is cause for concern, especially because the Army is more ethnically diverse than the U.S. population at large. Although cigarette smoking prevalence has declined among military servicemembers in recent years (23), smoking is still more prevalent among Army soldiers than it is in civilian populations, and exceeds the Healthy People 2000 goal for smoking cessation (123). The large number of smokers and the ethnically diverse population of the Army would make this a good setting for a study of the reliability of self-reported smoking data; such results could inform tobacco control research and prevention initiatives in both the military and civilian sectors.

On the other hand, the validation studies reviewed in this report demonstrate that although self-reports of smoking behavior may yield reliable or consistent results, they probably yield underestimates of actual smoking status or the number of cigarettes smoked per day. The studies reviewed above suggest that this under-reporting may be on the order of 2%-4%. The work by Bowlin et al., however, documented fluctuations in reliability and validity among various age- and gender-specific subgroups; it does not appear that anyone has evaluated validity of self-reported smoking status among racial or ethnic subgroups. For these reasons, researchers should exercise caution when using self-reported smoking data from the Army's HRA or other sources, and should consider the possible impact this level of misclassification might have on their results.

In the absence of any published studies evaluating the reliability and validity of the HRA items concerning cigars, pipes, and smokeless tobacco, it is difficult to say anything about the quality of information elicited by these items. The results by Sanchez and Bray about the increasing prevalence of past-year cigar and pipe use in the armed forces is, however, cause for concern. Sanchez and Bray were hindered in their analysis because their survey asked about cigar and pipe use in a single question, and they were thus not able to parse out the differences in use of these two forms of tobacco smoking. In order to conduct effective surveillance and research on this health issue, and to support the design and implementation of effective interventions, there is a clear need for a well validated instrument that inquires about different methods of tobacco delivery as well as patterns of use (6).

PERIODIC HEALTH EXAMS

The HRA asks about two preventive health practices that apply to both men and women: screening for colorectal cancer and periodic dental care. The 1991 BRFSS included an optional module on colorectal cancer screening, which comprised a series of nine questions on rectal exams, tests for occult blood, and colonoscopy. One of the questions in this series is very similar to the HRA question: a yes/no question on whether the respondent had ever had a digital rectal exam. Nothing is known about reliability or validity of self-reported periodic dental exams.

Brownson et al. evaluated the test-retest reliability of the item concerning the digital rectal exam in a sample of respondents to the BRFSS in the state of Missouri

(24). They documented a κ statistic of 0.59 for this item, showing that this item has fair to good reliability.

Two studies have evaluated validity of reports of digital rectal examinations. Gordon et al. compared self-reported data from a random sample of subscribers in the Kaiser Foundation Health Plan (aged 40-74 years) with medical records for these participants (57). Reports from more than two-thirds (69.8%) of the patients who reported having had such an exam within the past year were corroborated by finding documentation in the medical record. The sensitivity in this sample was quite high (97.4%), although the specificity was quite low (22.1%). This indicates that people who have truly had the exam within the specified time interval will likely report it accurately, but that a very high proportion of people will report this history incorrectly, probably through a tendency to underestimate the time that has elapsed since their last exam. Montano et al. compared rates of digital rectal exams as documented in physician self-reports of typical screening practices, patient self-reports of recent screening practices, and in medical record data (81). They calculated correlation coefficients between the three methods of report and found very good agreement between patient survey data and chart audit; among female patients, the correlation coefficient was 0.84, and among male patients, it was 0.71. However, correlations between physician self-reports and chart audit data and physician self-reports and patient survey data were not as strong, possibly indicating that physicians may overestimate their compliance with recommendations concerning routine screening initiatives. Finally, as we will review later in this report with respect to the items concerning women's cancer screening practices, there are limitations and biases in using medical record data to corroborate patient self-report, especially with respect to screening practices that do not generate a report from a third source (e.g., cytology or radiology reports, as you would obtain from a Pap smear or mammogram). Screening practices that are performed in the physician's office, such as the digital rectal exam, may not always be documented in the patient's chart.

A side note to this item is that Army regulations require male soldiers to undergo a digital rectal exam as part of the periodic physical exam over the age of 40, meaning that beginning at age 40 and up until age 60, they should have one every 5 years, and annually thereafter (43). In addition, regulations require that certain initial physical exams include a digital rectal exam (e.g., class I flight physicals) (43). It is worth noting that this is a developing field of medical practice, and not all medical organizations recommend periodic screenings. The National Cancer Institute, for example, notes that digital rectal examination has failed to show a decrease in mortality, and neither they nor the CDC currently make any recommendations about routine screening (87, 100). Efforts to evaluate the reliability and validity of this item in Army populations could focus on soldiers over 40 who took an HRA.

WOMEN'S HEALTH

The HRA includes eight items on women's health, asking about reproductive history and preventive health practices. All of these items except the item about breast self-exam were on the CDC/Carter Center's HRA, although it does not appear that the

Carter Center undertook any studies to assess their reliability or validity. The items asking about breast self-exam and hysterectomy were on the CDC's HRA. Many of the women's health items on the Army's HRA are similar to items on the BRFSS, except that the BRFSS questions solicit more detailed information on each. For example, the BRFSS asks the following questions: (1) have you had a mammogram (yes/no); (2) how long has it been since your last mammogram; (3) was your last mammogram routine or because of a problem or previous cancer; (4) whose idea was it for you to have mammogram. The BRFSS does not offer a definition of the term hysterectomy. The item reads simply, "have you had a hysterectomy?" The BRFSS similarly asks for more detail about Pap smears: (1) have you heard of the Pap smear; (2) have you had one; (3) when was your last Pap smear. The BRFSS does not ask about breast self-exam, but asks (1) have you had a breast exam by MD or medical assistant?; (2) how long has it been since last breast exam; (3) was your last breast exam routine, because of a problem, or previous cancer. Examining the psychometric properties of these BRFSS items may give us some evidence as to the quality of the information gathered by these HRA items.

Reliability of Cancer Screening Practices

Two studies have evaluated the test-retest reliability of the BRFSS items on Pap smears, mammograms, and clinical breast exams (Table 6). Brownson et al. evaluated the test-retest reliability of the items on mammography and Pap smears in a group of 222 BRFSS respondents from Missouri (24). Stein et al. evaluated the test-retest reliability of the women's health module of the BRFSS in a sample of 270 women from Massachusetts (107). Stein et al. inquired about the prevalence and recency of screening practices, the reason why the screening test was performed (e.g., routine exam or because of problem), the prevalence of hysterectomy and pregnancy, and conducted subanalyses to determine whether accuracy of reporting varied across racial and ethnic subgroups.

Nearly all of the women in both surveys reported having had a Pap smear (97% of the women in the Massachusetts survey and 95% of the women in the Missouri survey), and the proportion of women who gave concordant responses at the first and second survey was also very high (κ s of 0.75 and 0.68, respectively). Participant recall of the length of time since the last Pap smear was consistent in both studies, with 90% of the Massachusetts cohort and 89% of the Missouri cohort reporting the same interval at time 1 and time 2 ($\kappa = 0.64$ and 0.76 , respectively).

Stein et al. found that slightly less than half of the women in their study reported ever having had a mammogram, but documented that more than 93% gave the same response to this question on the two surveys ($\kappa = 0.86$), with 80% of women reporting the same time interval since last mammogram at time 1 and time 2 ($\kappa = 0.50$). Brownson et al. found that 75% of the women in the Missouri sample reported having had a mammogram in the past year. Their findings with regard to reliability were very similar to those in the Massachusetts study, with 95% giving consistent responses at time 1 and time 2 ($\kappa = 0.87$); a slightly smaller percentage (90%) gave consistent

responses on the item querying whether they had had a mammogram in the past year ($\kappa = 0.79$).

Stein et al. also evaluated test-retest reliability of the items on recency of clinical breast exams. Eighty-six percent of the Massachusetts women gave consistent responses on this item ($\kappa = 0.41$) and 86% reported the same interval since the last such exam ($\kappa = 0.51$). Their results seemed to indicate that nonwhite women tended to report clinical breast exam information less consistently than white women, although not all of the tests across subgroups reached statistical significance, and the high degree of concordance across so many of the items limited in their ability to test this avenue of inquiry thoroughly.

In general, these two studies show that survey items inquiring about women's compliance with recommended cancer screening practices generally elicit reliable and consistent responses. Although the kappas on the clinical breast exam items are lower than those for the Pap smear and mammography items, they are still within the 0.40 threshold of desirability.

Table 6. Summary of Studies Evaluating Test-Retest Reliability of Self-Reports of Pap Smears, Mammograms, and Clinical Breast Exam

Study	Item	Pap Smears			Mammograms			Clinical Breast Exam		
		N	% Agreement	κ	N	% Agreement	κ	N	% Agreement	κ
Brownson (1994)	Ever had	135	96	0.68	81	95	0.87			
	Past year	123	89	0.76	39	90	0.79			
Stein (1996)	Ever/never	270	97	0.75	270	93	0.86	270	86.7	0.41
	Time interval	247	87.9	0.64	115	80.9	0.50	216	86.1	0.51
	Reason for test	249	89.2	0.42	113	90.3	0.51	216	95.4	0.52

Reliability of Self-Reported Age at Menarche

A recent project to evaluate the relationship between exposure to organic solvents and development of breast cancer among active duty Army women included a subanalysis to examine test-retest reliability of self-report of age at menarche, using data from the Army HRA (89). Among the 9,925 women who took the HRA more than once, 60% reported no difference (n=6,019). Among the 4,451 women who did report a different age at menarche, reports varied by only one year.⁵

Bean et al. assessed validity of self-reported age at menarche (8). A sample of 160 women who were participants in a longitudinal study, the Menstrual and Reproductive History study, were given a questionnaire eliciting information about various aspects of their menstrual history, including age at menarche. Responses were then compared with interview data that had been gathered at enrollment into the study. Although the length of recall for these women ranged from 17 to 53 years, 59% of women accurately recalled their age at menarche and 90% were accurate within one year. Although the authors determined that recall of other variables concerning menstrual history (e.g., length or variability of cycle) was unreliable, they concluded that most women could accurately recall major milestones such as age at menarche.

To our knowledge, there have not been any studies of reliability or validity of recall of age at first birth, but the results presented by Bean et al. with regard to accuracy of recall of other reproductive milestones seem to indicate that women can accurately recall these events.

Validity of Cancer Screening Practices

Our review of the literature discovered 15 studies evaluating validity of self-reports of several women's health screening practices (see Table 7).

⁵ CAPT C. Rennix, written communication, July 16, 2003.

Table 7. Studies Evaluating Validity of Self-Reports of Women's Health Screening Practices

Author	Year	Location	Screening Test	Population and Study Design
Walter	1988	Canada	Pap smear	- Comparison of interview data and medical records (abstracted by physicians). - Case-control study of 181 women with cervical cancer aged 20-69, and 905 healthy controls. - Case-control study of 250 women with cervical dysplasia and 500 healthy controls.
Sawyer	1989	North Carolina	Pap smear	- Comparison of interview data and medical records (abstracted by office secretaries or nurses). 149 black women aged 16 to 75 in rural areas of three North Carolina counties.
Michielutte	1991	North Carolina	Pap smear	- Comparison of interview data and physician report of procedure at the current visit. 318 women aged 18 and older attending a county public health clinic for sexually transmitted diseases, August 1989-January 1990.
Bowman	1991	Australia	Pap smear	- Comparison of telephone survey data and pathology laboratory records. 234 women aged 18-70 contacted in a random household survey.
King	1990	U.S. health plan	Mammogram	- Comparison of telephone survey data and HMO radiology database and physician records. 199 women aged 50-74 and over enrolled in an HMO.
Degnan	1992	North Carolina	Mammogram	- Comparison of telephone survey data and regional medical center databases. 456 women aged 50-74.
McKenna	1992	Canada	Pap smear	- Comparison of interview data and medical records (abstracted by study personnel). 125 urban black women with cervical cancer diagnosed in 1986-1987, identified through the Illinois tumor registry; study examines accuracy of self-reports of Pap smears within 3 years of diagnosis, but excluding the year of diagnosis.
Fruchter	1992	New York	Pap smear	- Comparison of interview data and cytology laboratory records. 263 women (primarily Black and Latina) in medical clinics of a public hospital.
Gordon	1993	Northern California	Pap smear Mammogram Clinical breast exam	- Comparison of mail survey data (75% response rate) of six different cancer screening practices with medical record audit data (abstracted by study personnel). - Subjects were aged 40-74 and members of Kaiser Permanente Medical Care Program for 5 years prior to date of survey. - Pap smear N = 352 - Mammogram N = 386 - Clinical breast exam N = 371
Montano	1995	Washington	Pap smear Mammogram Clinical breast exam	Comparison of screening rates obtained from physician survey, patient self-reports, and chart audit, among community-based family practitioners in Washington State. Physicians surveyed (N=450; 74% response rate). Patient survey (N=11,005). Chart audit (N=3,281 patient charts).
Suarez	1995	El Paso, Texas	Pap smear Mammogram	- Comparison of interview data and medical records (multiple facilities, abstracted by study personnel). 450 low-income Mexican-American women aged 40 and over (82% response rate); Pap smear N = 215; mammogram N = 215.
Zapka	1996	Massachusetts	Mammogram	- Comparison of mail survey and telephone interview data and physician and radiologist records. 392 ethnically diverse women aged 50-74 seen by primary care physicians (in private offices and at a public teaching hospital).

Author	Year	Location	Screening Test	Population and Study Design
Bowman	1997	Australia	Pap smear	- Comparison of telephone interview data (81% response rate; N=5,706) and cytology laboratory records (data abstracted by cytology laboratory personnel and study personnel). - Study randomly selected 224 women (aged 18-70) who reported a Pap smear and 231 women who reported no Pap smear in the past 3 years for analysis.
McGovern	1998	Minneapolis	Pap smear Mammogram	- Comparison of interview data and cytology/radiology database records (abstracted by study personnel). 477 women aged 40-92 attending non-primary care clinics (e.g., surgery, orthopedics) at a public hospital.
Lawrence	1999	San Antonio, Texas	Mammogram	- Comparison of telephone survey data and financial, radiology, and clinic records from two healthcare systems (civilian and military). 93 military women (54% response rate) and 139 civilian women (33% response rate) aged 50-74 years.

Validity of Self-Reports of Pap Smear History. Overall, the quality of self-reports of Pap smear status is only fair to good, with specificities ranging from 89% to 97% in the studies reviewed (see Table 8). The specificities documented herein are somewhat disappointing, however, ranging from 35% to 64%. These specificities translate into false positives ranging from 36% to 65%, possibly indicating that women are incorrectly recalling the date of their last screening (and as a consequence, may not be getting screened according to the recommended schedule). Comparing the overall number of Pap smears in self-reports and medical records revealed that women report having, on average, one Pap smear every 2.0 years, whereas the medical records documented only one Pap smear every 3.9 years (110).

The studies reviewed offer several possible explanations for poor agreement between patient self-reports and medical record data.

The first and most common source of error between self-reports and medical record data is introduced when the patient inaccurately recalls the date of their last Pap smear. This phenomenon is known as “telescoping” and has been documented in nearly every one of the studies we have reviewed (19, 54, 57, 75, 76, 96, 115). Fruchter et al. found that 78% of the women at an ambulatory care clinic gave self-reported dates of a last Pap smear that were correct within 1 year of the pathology report (54). Of the 22% whose self-reports varied by more than 1 year from the date on the pathology report, 16% gave dates that were more recent than the report, showing a significant tendency to underestimate the length of time since their last screening. Bowman et al. attempted to quantify the effect of telescoping errors on accuracy of self-reports and found that specificity and positive predictive values improved when comparing self-reports against longer intervals of laboratory records (19). For example, for women who said they had had a Pap smear within the past year, they found a specificity of 64%, but when they searched laboratory records for the previous year and 3 months, year and 6 months, 2 years, 3 years, and 4 years, they documented modest incremental increases in specificity (65.6%, 66.1%, 67.1%, 69.2%, and 70.2%, respectively). Bowman et al. refer to this as “leeway,” and suggest that using a window of several months to a year on either side of the self-reported date may improve ability to confirm whether a woman had the test as reported. On the other hand, Gordon et al. found that most of the discrepancies between self-reported Pap smear history and medical record data involved differences of more than 12 months (57), suggesting that although increasing the “leeway” between self-reports and medical record data will improve the agreement between the two, at some point it will dilute the utility of patient self-reports as a clinical decision-making rule in deciding whether or not to administer the screening test.

Second, accuracy of patient self-reports may vary by patient status or history of cervical cancer or abnormalities, although the data are inconsistent in this regard. In a case-control study of accuracy of self-reports among women who did and did not have a history of cervical cancer, Walter et al. used a two-sided test for symmetry to evaluate tendency of patient to systematically report higher or lower values than the physician (115). In the cancer study, they found healthy controls were significantly more likely

than cancer cases to report their last Pap smear as more recent than it truly was. In the dysplasia study, both cases and controls tended to telescope the recall of the most recent symptom-free Pap smear, but when data from all smears were analyzed, dysplasia cases tended to overestimate the amount of time that had passed since their last symptom-free Pap smear, whereas the healthy controls continued to underestimate this interval. There are two possible explanations for this finding: first, women who have had cervical cancer or dysplasia may be more likely to recall details concerning their diagnoses more accurately, or second, women who have had a history of these conditions may be more likely to have Pap smears more frequently than other women, and may thus be more familiar with the procedure. Walter's findings differ from those of Suarez et al., who found that in a population of Mexican-American women, those who had Pap smears for some type of health problem were only slightly more likely to recall the interval accurately, as compared to women who had Pap smears for screening purposes (110). And in marked contrast to both Walter's findings and Suarez's, McKenna et al. found in a population of urban black women with confirmed diagnoses of cervical cancer that the women reported abnormal Pap smears within 3 years prior to diagnosis with much less accuracy than they reported any and all Pap smears in 3 years prior to diagnosis (κ 0.34 for all Pap smears vs. κ 0.08 for abnormal Pap smears) (76).

Third, social desirability may influence whether patients are accurate in their self-reports. Sawyer et al. explored perceived barriers to getting routine Pap smears, and found that women who perceived logistical barriers to getting a Pap smear or who found pelvic examinations unpleasant or embarrassing were more likely to recall the date of their last Pap smear inaccurately (96). In two of the studies we reviewed, the authors speculated that women may have reported complying with screening recommendations simply because they know they ought to have these tests performed routinely (18, 110). Montano et al. demonstrated that physicians themselves might also be susceptible to social desirability biases. In the only study that surveyed physicians about screening practices, they found that although there was a high correlation between chart audit data and patient self-reports (0.79), correlations between chart audit and physician survey and patient survey and physician survey were much lower (0.37 and 0.29, respectively), possibly indicating physicians may overestimate their compliance with recommendations concerning routine screening initiatives (81).

The studies reviewed have enumerated several possible determinants of inaccurate reporting, in an effort to refine the clinical screening guidelines.

First, the studies reviewed have not identified any clear demographic differences among women who do and do not report screening histories accurately. Low educational attainment, for example, has not been found to impact accuracy of reporting (77, 96). McKenna et al. found that urban black women with diagnosed cervical cancer and who were younger than 40 were 2.8 times more likely to correctly report Pap smear history within 3 years of diagnosis (76). The authors also noted, however, that younger women were more likely to report having had more than one Pap smear within the 3 years prior to diagnosis, and hypothesized that the greater accuracy of their reporting may be a byproduct of their degree of familiarity with the procedure.

Second, several studies have documented variations in accuracy of self-reports in different clinical care environments and with different types of medical providers. Suarez et al. found that accuracy of self-reports of Pap smear history was significantly greater among Hispanic women who had obtained care at a public health clinic, when compared to women who had had Pap smears performed in hospitals or in doctor's offices ($P = 0.0005$) (110). Suarez et al. hypothesized that the women who obtained care at a public health clinic may have been more likely to have been seen by nurse practitioners, who may spend more time with their patients or may spend more time educating their patients about the various screening practices that are being performed. In a similar vein, Sawyer et al. found that accuracy of self-reports varied by the type of health practitioner seen, with women who saw nurse practitioners being more likely to report screening status accurately than women who saw internists or family practitioners (although the differences did not reach statistical significance) (96). The authors attribute these inaccuracies to confusion on the part of the woman over whether or not a Pap smear was done at the time of a pelvic examination (not always a safe assumption), and caution survey researchers to distinguish carefully between the two in asking women about these screening practices. Indeed, approximately half of the patients in the study by Michielutte et al. incorrectly reported that they had a Pap smear *at the current visit*, with approximately 90% of these women believing a Pap smear had been performed when it had not (77). In a focus group, Michielutte et al. found that many women believed a Pap smear tested for pregnancy or infection, indicating that there is considerable confusion about the purpose of this procedure. Univariate analyses indicated that self-reporting errors were more common among younger women and never married women, indicating that these women may need to be educated more carefully about the differences in the two procedures and the recommended timing for each.

Finally, Bowman et al. assessed impact of several behavioral and attitudinal influences on accuracy of self-reports, and found only one significant association: the woman's degree of certainty as to whether she was accurately reporting the date of her last Pap smear (19). A subanalysis of the women who were very sure of the date of their last Pap smear documented a positive predictive value of only 66% (only 5% higher than the positive predictive value for the whole sample), rendering the woman's certainty of little practical value in helping clinicians determine whether or not to screen at the present visit.

Table 8. Quantitative Measures of the Validity of Self-Reports of Pap Smear History

Study	Study Group	N	% Agreement	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	Kappa	Symmetry Ratio
Number of Pap smears in past 5 years									
Walter (1988)	Cancer cases	133	58%					0.44	1.33
	Cancer controls	576	64%					0.53	2.89**
	Dysplasia cases	181	42%					0.21	2.62**
	Dysplasia controls	241	38%					0.15	4.00**
Suarez (1995)		215				71%	82%		
Number of Pap smears in past 4 years									
Bowman (1997)		455		96%	38%	63%	89%		
Number of Pap smears in past 3 years									
Sawyer (1989)		98	80%	95%	47%	79%	83%	0.46	
Bowman (1991)		111	78%	93%	55%	77%	82%		
McKenna (1992)		105	65%					0.34	17.5
Fruchter (1992)		138	72%					0.46	
Bowman (1997)		455		96%	42%	61%	92%		
Suarez (1995)		215				67%			
Number of Pap smears in past 2 years									
Gordon (1993)		352	78%	97%	35%			0.38	
Bowman (1997)		455		97%	49%	52%	97%		
Suarez (1995)		215				61%	88%		
Number of Pap smears in past year									
Bowman (1997)		455		89%	64%	40%	95%		
Suarez (1995)		215				46%			
McGovern (1998)		281				66%	86%	0.52	
Recall accuracy of interval since last Pap smear									
Walter (1988)	Cancer cases	88	85%					0.52	0.44
	Cancer controls	318	59%					0.27	0.07**
	Dysplasia cases	181	80%					0.18	3.00*
	Dysplasia controls	241	75%					0.51	0.20**
Recall accuracy of interval since last symptom-free Pap smear									
Walter (1988)	Cancer cases	22	91%					0.70	--
	Cancer controls	218	60%					0.24	0.23**
	Dysplasia cases	135	70%					0.39	0.52*
	Dysplasia controls	212	75%					0.50	0.37**

* p < 0.05; ** p < 0.01

Validity of Self-Reports of Mammography History. One of the earliest studies to assess validity of mammography self-reports found an exceptionally high degree of agreement between women's self-reports and HMO records (67). Not one of the 99 women who reported that they had not had a mammogram in the past year was found to have a mammogram report in the HMO database. Conversely, nearly all (94/100) of the women who reported having had a mammogram had their self-reports confirmed by the positive location of a mammography report in the HMO database. The remaining six women were all found to have had mammograms, but they occurred more than 1 year prior to the survey. Gordon et al. compared women's responses to a question about mammography status within the past 2 years and found a high degree of concordance between self-reports and chart audit data among 386 women subscribed to the Kaiser Foundation Health Plan (83.7%) (57). As in the Pap smear studies reviewed earlier, however, sensitivity was high (98.0%) while specificity was low (50.6%), indicating a high likelihood of inaccuracies on date recall.

In the only validation study to take place in a military healthcare setting, Lawrence et al. compared accuracy of self-reports of mammogram among 232 women in two healthcare systems in the same Texas city: a military hospital and a county hospital (69). They examined financial, radiologic, and clinic records of the two healthcare systems, and identified two groups of women who had and had not undergone mammograms within the previous year, then randomly contacted subsamples of these women by phone to ask about recency of mammography. These researchers used different definitions of sensitivity and specificity in their work than most of the other studies reviewed in this report. Sensitivity (defined in this particular study as the percentage of women who accurately reported not having had a recent mammogram) was the same in the two groups: 65% among the women in the military system and 62% among women in the county health system. Specificity (defined in this study as the percentage of women who accurately reported having had a recent mammogram) differed between women in the military and civilian systems (95% vs. 79%, respectively). The likelihood of inaccurately reporting mammogram history is thus similar in this study to the other studies we've reviewed (approximately 35%-38%). Furthermore, the authors cautioned that this was a small study, that a large proportion of women identified in the pool of eligible subjects could not be contacted, and that the reader should exercise caution in attempting to generalize these results.

As is true for self-reports of Pap smear history, most studies of self-reports of mammogram history found that women tended to telescope the date (Table 9) (38, 75, 110, 125). Degnan et al. found that women inaccurately recalled the date of their last mammogram by about 3 months (38). Zapka et al. surveyed a multiethnic population of women who were all known to have had mammograms and found that although all subjects confirmed having had a mammogram, only 31% correctly reported the exact date (125). They noted, however, that when less strict criteria were used (i.e., self-reports match the clinic record by +/- 3 months), the percentage of women correctly reporting mammogram history rose to 54%, and under even less stringent criteria (i.e., self-reports match the clinic record by +/- 12 months), it increased to 83% (125). This is reminiscent of Bowman's analysis of how much "leeway" one should allow women in

incorrectly reporting the date of their last Pap smear, and has similar implications, that is, that narrow windows of leeway (e.g., approximately 3 months) may provide small incremental improvements in reconciling self-reports and physician records of these screening tests. Zapka et al. (125) and McGovern et al. (75) both concluded that accuracy of recall was significantly related to time interval since the last mammogram. However, McGovern et al. noted that after adjusting for this, they found no significant differences in accuracy of reporting by race, education, or income (75).

The study by Montano et al. reviewed previously with regard to physician screening practices concerning Pap smear screening also compared accuracy of reporting of mammography (81). As was true for Pap smears, the highest degree of correlation was found between chart audit data and patient self-reports (0.74), with correlations between chart audit and physician survey and patient self-reports and physician survey being much lower (0.31 and 0.36, respectively).

Table 9. Quantitative Measures of the Validity of Self-Reports of Mammography

Study	N	% Agreement	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	Kappa
Number of mammograms in past 5 years							
Suarez (1995)	215				79%	98%	
Number of mammograms in past 3 years							
Suarez (1995)	215				77%		
Number of mammograms in past 2 years							
Gordon (1993)	386	84%	98%	51%			0.61
Suarez (1995)	215				75%	98%	
Number of mammograms in the past year							
Suarez (1995)	215				49%		
McGovern (1998)	456				72%	91%	0.63

Validity of Self-Reports of Family History of Breast Cancer. Family history of breast cancer is recognized as a significant risk factor for the disease, but surprisingly few studies have examined the reliability and validity of information on familial cancer gathered via self-reports. Kerber and Slattery compared the family histories of cancer from the cases and controls in the Diet, Activity, and Reproduction in Colon Cancer (DARCC) study with a Utah cancer registry (66). The authors acknowledge that the cancer registry may not have been complete, but they expected it would have confirmed cases reported by the study participants. The sensitivity of reporting for breast cancer was the highest of any of the familial cancers they examined (83%). The κ statistic also showed a moderate degree of agreement between the interview and the cancer registry ($\kappa = 0.63$), but it did seem that the cases reported information more reliably than controls (κ s = 0.73 and 0.58, respectively). Kerber and Slattery furthermore noted that younger persons seemed better able to report familial history of cancer more accurately.

Implications for the Army's HRA Data

We found mixed results in our review of studies concerning the reliability and validity of the eight items on the Army's HRA that pertain to women's health. Although the civilian studies by Stein and Brownson show a moderate degree of reliability on the measures of whether the respondent had ever had one of the recommended screening tests, the form of the item on the Army's HRA prompts for length of time since the *last* such test, and the validation studies reviewed herein demonstrate that women tend to recall this more detailed information less accurately, through the so-called telescoping effect.

Having said that, a certain degree of caution is warranted in interpreting studies that compare self-reports against clinic or laboratory records, as medical records are often incomplete and cannot truly be considered a gold standard. Relying on medical records as the gold standard may result in an underestimate of concordance (57). McKenna et al., for example, commented on their frustrations in medical record review, as they found it difficult, if not impossible, to match dates of Pap smear cytology reports to documented evidence of a pelvic examination having been performed in the clinical exam (76). This may be especially true for the evaluation of cancer screening practices that do not result in a laboratory report in the medical record (e.g., clinical breast exams or digital rectal exams, which are typically documented only in the progress notes section of the chart and do not result in a verifiable third party report such as a cytology lab or radiology clinic). A simple explanation for discordance between the medical records and the woman's self-report may lie in the possibility that women whose medical records were used in these studies may have sought these screening practices elsewhere. Few of the studies we reviewed made exhaustive efforts to locate Pap smear and mammography records from other providers or clinics.

In general, however, the low specificity of self-reported cervical cancer and breast screening rates suggest that it is possible to identify only approximately half of the women who are at risk for being underscreened through self-reports. This calls the utility of the HRA as a screening tool into question. The positive and negative predictive

values found in the studies reviewed illustrate the practical consequences that poor recall may have on a screening program. With respect to Pap smears, for example, if the only women eligible for Pap smears were those who could recall having had one within the past 3 years, anywhere from 3% to 18% of women would be overscreened; that is, they would have received a test when it was not truly necessary. Conversely, anywhere from 21%-60% of women would not have been screened when they truly needed a test, because their tendency to telescope the date had led them to report that they had had a test within the 3 years, even if it had been earlier. Likewise, recall of mammography history could result in overscreening rates of 2%-9%, and underscreening rates of 21%-51%. In their study of Mexican-American women, Suarez et al. found that women reported one Pap smear every 2.5 years (whereas the medical records showed only one Pap smear every 3.9 years) and one mammogram every 5.6 years (whereas the medical records showed one mammogram every 9.6 years), demonstrating the impact that poor self-reports can have on estimates of compliance with recommended screening practices (110). These validation studies cast some doubt on the utility of these items for epidemiologic research, and suggest that HRA data on compliance with recommended cancer screening practices should be used with caution.

The findings on accuracy of self-reports of family history of cancer are encouraging, but further studies are warranted to evaluate the quality of these data before using them in epidemiologic research. Likewise, the results of the study by Bean et al. are encouraging, and suggest that the HRA items about age at menarche, age at first birth, and age at hysterectomy, respectively, may yield results that are valid enough for use in epidemiologic research. It would be possible, moreover, to link Army HRA records to Army hospitalization records, and thus conduct a validation study of self-reported age at first birth. The Army has a large enough population of women of childbearing age to allow for analysis across racial and ethnic subgroups, which would be a useful addition to this body of literature. A similar analysis could be done to validate the item concerning hysterectomy, although there would probably be a smaller number of cases.

MEN'S HEALTH

The HRA asks two questions pertinent to men's health. One of these questions, "How long it has been since your last prostate rectal exam?" appeared on the CDC/Carter Center's HRA. With the exception of the Brownson study described above, we were not able to locate any studies evaluating the reliability or validity of reporting about prostate or rectal exams. Interestingly, Brownson et al. also documented test-retest reliability for the item asking about prostate specific antigen (PSA) test, and documented a very low κ score for respondents having heard of this test (0.21), and a fair to good κ statistic for respondents having had the PSA test within the past year (0.60). The authors note that, in general, assessments of reliability of self-reports for male cancer screening tests are lower than reliability of self-reports for female cancer screening tests.

The source of the item about frequency of testicular self-exam is not known; we were not able to locate studies assessing the reliability or validity of this item.

SUMMARY

For more than a decade, the Army offered the HRA to its soldiers. Although intended primarily as an educational tool in a health promotion campaign, the program collected a vast quantity of data on health habits of Army soldiers. It is important to thoroughly understand the strengths and limitations of these data before using them for surveillance or research, however. Inaccuracies in these data can hamper surveillance and research efforts in numerous ways. If the instrument yields underestimates of the true prevalence of risky behaviors, health promotion programs targeting those behaviors may be underfunded or otherwise misdirected. If the instrument yields unstable estimates of certain behaviors, program planners may become frustrated in their efforts to bring about behavior change. The HRA database represents the best single source of data on health habits for epidemiologic research, but misclassification of exposure could bias estimates of effect, threatening the validity of surveillance and research endeavors.

Table 10 reviews what is known about the reliability and validity of the items on the HRA by topical area. As reviewed in this report, the greatest utility of the HRA is probably in surveillance and research efforts that analyze responses to individual items in order to assess the prevalence of certain health habits and behaviors within the Army. There is considerable evidence in the literature indicating that most of the items perform fairly well, and may be useful in surveillance and research. In some cases, the literature also suggests that the items may be useful in combination with other data on health habits (e.g., the seat belt item may be useful in combination with other items in assessing risk-taking propensity). In other cases, however, there is serious doubt as to whether certain items produce reliable and valid responses; such items from the HRA may not be of sufficient quality for epidemiologic research without corroboration from other sources or adjustment for potential misclassification.

In a review of the literature about the veracity of self-reported alcohol use, Midanik noted that validation of self-reported data is “still not seen as a completely legitimate research direction (79).” Her lament is ironic, given that the field of alcohol research is, indeed, one of the few areas where the validity of self-reported data has received much substantive attention from researchers. As reviewed in this report, self-reported data of many other health habits have received far less attention. Many of the studies in this report were hampered by small sample sizes, making it difficult to parse out variations in the quality of self-reported data among various demographic subgroups, for example. The Army’s HRA database could be combined with other Army data sources to evaluate the reliability and validity of self-reported health habit data within the military population—a population that is not only often understudied, but also has a greater percentage of members from minority racial and ethnic backgrounds than the U.S. population at large. Efforts to evaluate the reliability and validity of data collected by the Army’s HRA can inform not only health promotion efforts within the military, but can inform research efforts in the civilian world as well.

Table 10. Summary of Reliability and Validity of the Self-Reported Health Habit Data Gathered by the Army's HRA

HRA Item	Estimated Utility in Epidemiologic Research	Notes
Exercise	Fair	- Studies in civilian populations show only modest test-retest reliability and poor to fair criterion validity; moreover, the study populations differed substantially from the Army, which is younger and more ethnically diverse. - Because of occupational requirement for physical fitness, the Army is fairly homogeneous with respect to aerobic exercise and strength training habits. While HRA exercise questions may be useful in surveillance, they probably do not capture a sufficient level of detail about exercise habits that would be necessary in epidemiologic research.
Diet	Unknown	No studies have been located to assess either reliability or validity of these HRA items.
Stress	Unknown	No studies have been located to assess either reliability or validity of these HRA items.
Motor Vehicle Safety		
Vehicle Miles Traveled Annually	Poor	Civilian studies suggest that younger drivers tend to underestimate annual vehicle miles traveled. As the Army comprises mostly younger males, this HRA item should probably not be used as a literal measure of driving exposure.
Typical Mode of Travel	Unknown	No studies have been located to assess either reliability or validity of this HRA item.
Seat Belt Use	Good	Probably produces overestimates of actual use, and should therefore be used with caution. However, in combination with other HRA items, it is useful as an indicator of risk-taking propensity.
Adherence to Speed Limit	Unknown	No studies have been located to assess either reliability or validity of this HRA item.
Drinking and Driving	Unknown	No studies have evaluated the reliability and validity of the Army's HRA items. The version of this item that was implemented in the October 1990 version of the HRA form (i.e., the version of the form in use for most of the program's tenure) is double-barreled, and does not permit separate analysis of drinking and driving and riding with a drunken driver. However, this item may be useful in combination with other HRA items to assess risk-taking propensity.
Alcohol		
Consumption	Good	Civilian literature suggests that estimates of consumption are probably under-reported. This item is also limited because it truncates possible responses at a maximum of 99 drinks per week, does not have a separate estimate of frequency, and does not assess binge drinking. However, analyses of the Army's HRA database indicates that this item elicits a wide range of responses, suggesting that even if the actual quantity is under-reported, it probably captures variation in consumption accurately enough for research.
Alcohol-Related Problems	Good	The CAGE has been well validated, and the combination of the CAGE with two additional questions about risky drinking has been shown to accurately identify hazardous drinkers. The CAGE has been shown to perform well in predicting adverse health outcomes associated with alcohol consumption, suggesting that it has good criterion validity.
Diabetes	Fair	- Civilian studies have demonstrated good reliability and validity of this item among whites, but it seems to perform less well among racial or ethnic minorities. - This item asks whether the respondent has ever had diabetes, not whether they currently have it, and may produce a high rate of false positives (e.g., women who had gestational diabetes).

HRA Item	Estimated Utility in Epidemiologic Research	Notes
Hypertension	Poor	- Civilian studies have demonstrated poor reliability and validity of many types of self-reported information on hypertension (e.g., physician diagnosis, compliance with pharmaceutical regimen). - The Army's HRA includes only one item about whether the person takes medication for hypertension and does not gather information about physician diagnosis of hypertension. As the questionnaire does not collect denominator data, it is impossible to calculate rates of hypertension or compliance with medication regimens among soldiers on the basis of HRA data alone.
Tobacco		
Tobacco Use (Other Than Cigarettes)	Unknown	No studies have been found to evaluate reliability and validity of self-reported use of cigars, pipes, or smokeless tobacco.
Smoking Status (Cigarette Smokers)	Fair	Studies evaluating self-reports of smoking status exhibit good reliability, but probably yield estimates of smoking prevalence that are under-reported by approximately 2%-4%. These findings with respect to reliability and validity have been shown to vary with age, gender, and race/ethnicity. Researchers should use caution in using these self-reported data and should consider the possible impact that misclassification of this magnitude may have on their overall results.
Cigarette Consumption	Fair	This item seems to exhibit good reliability, although reliability has been shown to vary with age, gender, and race. It has not been rigorously evaluated for validity. Caution may be warranted before adopting this measure as a literal measure of exposure to tobacco smoke.
Periodic Health Exams		
Rectal Exam	Unknown	This item shows modest test-retest reliability. More data are needed to assess reliability among soldiers, especially among racial or ethnic subgroups and to assess validity of reporting.
Dental Visits	Unknown	No studies have been located to assess either reliability or validity of these HRA items.
Women's Health		
Age at Menarche/Age at First Birth	Good	One civilian study shows that women are able to accurately recall major dates in their reproductive histories.
Mammography and Pap smears	Fair	Two civilian studies demonstrate good reliability of self-reporting, but numerous validation studies show that although women report accurately whether they have ever had one of these screening tests, they are likely to underestimate the length of time since their last such test (telescoping). These data should be used with caution, or perhaps in combination with other sources of data to allow for assessment and correction of misclassification if necessary.
Familial History of Breast Cancer	Fair	A civilian study comparing self-reports to medical records attested favorably to the validity of self-reported data; however, self-reported data on familial cancers may vary with educational attainment, race, and age, and further study is needed.
Hysterectomy	Good	One civilian study reported high correlation on test-retest reliability of self-reports of hysterectomy; validity data reviewed with respect to validity of recall of age at menarche may suggest that women could accurately report having had a hysterectomy.
Breast Self-Exam/Clinical Exam	Poor	Civilian validation studies have shown poor agreement between medical record data and self-reports, and have demonstrated that self-reports are susceptible to inaccuracies through telescoping of the date.
Men's Health		
Prostate Rectal Exam	Unknown	In one study, the item shows modest test-retest reliability. More data are needed to assess reliability among soldiers, especially among racial or ethnic subgroups and to assess validity of reporting.
Testicular Self-Exam	Unknown	No studies have been located to assess either reliability or validity of this HRA item.

In summary, because of problems such as underreporting or telescoping, the information elicited by many of the items on the HRA may not be useful as literal reports of health behaviors. Many of the items can, however, be used in combination with other items on the HRA or other sources of data to develop an understanding of patterns of risky behavior; these risk-taking propensities may then be used to inform epidemiologic research or the development of health promotion programs directed toward the Army population generally. The literature is sparse on variations in reliability and validity of reporting among racial or ethnic minorities, and it is unclear whether or how HRA responses may be useful in developing more targeted interventions.

CHAPTER 4: LESSONS LEARNED, CONCLUSIONS, AND RECOMMENDATIONS

More than 15 years have passed since the Army's health promotion program was launched and the HRA questionnaire was implemented. The Army used this questionnaire for more than a decade to collect data on health behaviors of their soldiers. There is much to be learned from this experience, and many lessons that can be applied to the development of future questionnaires or health behavior surveys. A full understanding of these lessons that pertain to questionnaire development, however, also requires an understanding of some of the challenges encountered and lessons learned in the larger context of the Army's health promotion program.

LESSONS LEARNED FROM THE ARMY'S HEALTH PROMOTION PROGRAM

Over a two-decade period, from approximately 1980 until the turn of the century, the Army made impressive progress in instituting a cultural appreciation for health promotion, and expanded the definition of health and health promotion from a narrow focus on medical treatment and disease prevention to a broader understanding of total well-being. This period also ushered in a significant expansion in the scope of health promotion activities, from programs designed exclusively for soldiers to programs targeting the total Army population (e.g., active duty soldiers, reservists, civilian employees, retirees, and dependents). These changes were the direct result of an enormous effort on the part of a fairly small number of individuals. In the process of developing this report, we interviewed many people who were involved in the early phases of the health promotion program. The individuals with whom we spoke demonstrated their continued enthusiasm and commitment to the principles and mission of the Army health promotion program. Although these efforts were successful in the long run insofar as they rendered attitudinal changes about health and wellness possible, it is clear from examining the history of the health promotion program that some internal processes and external forces were at work to limit or hinder the overall success of these specific efforts.

Proponency and Ideology

As noted previously, the Army regulation that governed health promotion activities directed that responsibility be shared between two agencies: the Office of the Surgeon General (OTSG) and the Deputy Chief of Staff of Personnel (ODCSPER). Historically, these two agencies had been parties to a rivalry for control over various medical and personnel issues pertaining to health promotion, such as body fat and physical fitness standards and nutritional guidelines (109). Many of the individuals we spoke with described the development of the health promotion program as a "turfed" battle between these two agencies, specifically oriented around the philosophical underpinnings of the health promotion program.

The ODCSPER contingent favored a model based on an ideology of corporate wellness, whereas the OTSG believed that health promotion activities should foster

personal wellness and readiness. The corporate wellness model places emphasis on the health of units or divisions, and thus rates the health of an individual based on how he or she compares to the rest of his or her peers. The HRA reporting software that was developed for the Army would produce summary data reports that would “allow the unit commander to compare the health risk status of his or her unit to the Army overall and to unit performance over the last year (112).” These reports were intended for use by the commander and the local health promotion council to “(size) up relative risks at a glance, for identifying target risks for improvement, and for setting or evaluating goals for health risk reduction (112).” Advocates of the personal wellness model, however, believed that the greatest strength of the HRA was in its use as an educational tool to teach individuals about healthy habits and health risks and to motivate individual behavior change toward a healthier lifestyle. They feared that the emphasis on comparing the health of the unit might lead some commanders to use these so-called unit health report cards inappropriately, as the only metric of how successful their base’s health promotion program was. There was also some concern that commanders may place inappropriate emphasis on attaining a particular unit score with respect to any given health behavior, and that this may have led to some pressure (whether indirect or explicit) to sway the responses of soldiers who took the HRA.

This difference between corporate wellness and personal wellness orientations was not the only ideological disagreement present in the landscape of Army health promotion. As reviewed in Chapter 1, there was a similarly heated debate involving parties within the Preventive Medicine Division at OTSG regarding the selection of an HRA survey tool. Although the camp that championed the RIWC disparaged the risk estimation methodology favored by the committee designing the Over-40 Cardiovascular Screening Program, the RIWC was not without its limitations. After several months of experience with the RIWC-based HRA, the Army concluded that it was limited because it was not epidemiologically driven, it did not allow for comparisons between the Army and the U.S. population at large, and the version implemented by the Army had no identifier field, making it impossible to track the health behaviors of a person throughout his or her Army career. This was not a limitation of the RIWC per se, but the health risk appraisal selection committee had elected initially not to have an identifier on the form, out of concerns that soldiers may be reluctant to give honest answers to sensitive questions about risky health behaviors (109). The ideological battle over the choice of the HRA instrument ultimately ended in a victory for the camp that favored the CDC’s version, but it is not entirely clear that this was the right choice. Although the tool was deemed optimal for rating cardiovascular risk in a population of respondents over age 40, it may not have been suitable for the entire Army population, for a variety of reasons.

As reviewed in Chapter 1, the HRA methodology compares a person’s risk behaviors and health habits to those of people of similar age and sex, and quantifies the impact these habits have on the respondent’s prospects for health and longevity. It is supposed that the personalized nature and the quantitative presentation of the findings may lend greater impact or urgency to the communication of some health messages, and that the rank-ordered presentation of risks allows people to focus on which behavioral changes might have the greatest positive impact on their overall health

(122). There are many criticisms to this approach. First, it fails to take into account the client's readiness to change their behavior. Second, by themselves, education and knowledge are insufficient to spark behavior change, especially if the client does not have access to resources that will help him or her improve health, or if the environment does not support healthier behaviors. Third, if the messages are presented in too dire or threatening a way, they may be ignored. Fourth, if the messages are presented judgmentally, they may be interpreted as an attempt to "blame the victim" for his or her own health problems. Fifth, at the organizational level, it may place too much emphasis on conformity with the group. Finally, this approach tends to overemphasize the deterrence of negative health behaviors rather than focusing on or encouraging people to adopt more positive ones (35, 122). While each of these objections and concerns may have applied to the Army's health promotion program, a more serious stumbling block may have arisen in the decision to base the Army's HRA on the CDC's instrument, insofar as it was not optimized for a young, healthy population. The epidemiologic data on which the CDC's instrument is based (indeed, upon which many HRAs are based) is derived from a primarily middle-aged, middle-class, white, adult population (3) and, as reviewed in Chapter 2, it is not clear whether the risk algorithms can produce accurate risk estimates for younger adults. Moreover, because physical fitness is a job requirement, the majority of soldiers maintain higher levels of physical fitness than the civilian population; as physical inactivity is a major risk factor for many chronic diseases, it may not be accurate to draw comparisons about morbidity and mortality risks between soldiers and their civilian counterparts. Although the HRA tool selection committee may have thought they were doing the right thing by seeking a tool that was epidemiologically driven, the efforts to generalize what is known about the impact of health behaviors on civilian health to a military population may have suffered through a failure to take these factors into account. These problems with the algorithm may have compromised the utility of the HRA as a tool to motivate and sustain lasting behavior change among soldiers of all ages. For example, because the HRA operationalizes health risks in terms of "risk of dying within the next 10 years," the concrete impact of these risky behaviors can seem remote to a young, healthy person (122). Even among older soldiers, however, the HRA faced similar problems; although the risk of dying within the next 10 years may be more proximal among this group than it is for younger soldiers, many older soldiers are more physically fit than their civilian counterparts of the same age and sex upon whom the algorithms are based. HRA reports that highlight improvements the person could make in terms of reduction might quantify the impact of their health behaviors on their overall mortality in terms of lengthening or shortening their life by a matter of hours, thus undermining the impact of the HRA as a motivational tool.

Implementation Issues

There are two types of implementation issues that may have hindered the successful delivery of the health promotion program: inattention to high-quality outcomes research to justify the continuation of the program and insure the quality of information being obtained; and the combination of inadequate funding and decentralized administration.

Research Findings and Program Justification. In addition to the ideological differences between OTSG and ODCSPER, there were stylistic differences in how the two agencies chose to implement various components of the health promotion program. As noted in Chapter 1, the Army had embarked on various health promotion initiatives at the time DoDD 1010.10 came into effect. The most high-profile example of this was a study done at the Pentagon in the early 1980s that offered stress management and cardiovascular fitness training to military and civilian employees of the ODCSPER (109). The principal investigator on this study was assigned to the OTSG Preventive Medicine Division and oversaw the execution of the study protocol, program delivery, and analysis of all results. Military and civilian employees of the Army Staff Headquarters were randomly assigned to one of four groups (physical conditioning and Type A behavior modification; physical conditioning only; Type A behavior modification only; and a control group). Subjects were offered nutrition education and smoking cessation counseling, and civilian employees were given special permission to use work time to participate in physical fitness conditioning classes. By regulation, military servicemembers are allowed to use duty time for physical fitness conditioning (42). Among other things, findings demonstrated reductions in coronary risk behaviors, improvements in physical fitness outcomes, and improvements in outcomes such as energy level, morale, and mental alertness (109).

A follow-up study, dubbed the ARSTAF (Army Staff Headquarters) Corporate Fitness study, was launched throughout the Pentagon in 1984, and was specifically designed to demonstrate cost-benefit and cost-effectiveness results that may be associated with health promotion activities (109). Health promotion advocates believed that unless they were able to demonstrate that health promotion activities could be translated into measurable cost savings, they would be unable to implement some desirable aspects of the program. For example, civilian employees were not allowed to use duty time to participate in fitness-related activities, but it was believed that if such programs could demonstrate cost savings (e.g., reduced health costs or less use of sick time), the Office of Personnel Management might be persuaded to endorse civilian participation. At its inception, the ARSTAF program was again under the direction of the OTSG, but midway through its implementation, oversight responsibilities were shifted to the ODCSPER. Even before the ARSTAF protocol was completed or the cost-effectiveness analyses were performed to examine the efficacy of the program, the ODCSPER began making plans to implement this health promotion program Army-wide, and began work on the so-called Exportable Package. The health promotion program the Army would implement in 1987 under AR 600-63 thus had its roots in the ARSTAF Corporate Fitness Program. While the ODCSPER was proceeding with development of the Exportable Package, however, leadership and oversight of the research component of the ARSTAF program was becoming mired in personality conflicts and competition for control (109). This unfortunate turn of events had implications for the development and implementation of all health promotion activities. ODCSPER easily incorporated the development of the Exportable Package into its mission, understanding it to be part of a tangible, visible mission to develop and disseminate health promotion materials. However, they were unwilling to wait for the political jockeying that was swirling around the research effort to play itself out. In this

climate, the development of the Exportable Package proceeded without the benefit of a well-designed and meaningful study results to justify its implementation (109).

Inadequate Program Funding and Decentralized Program Administration.

As noted in Chapter 1, the regulation that established Army health promotion activities placed responsibility with the installation commander, but did not provide funding for health promotion activities, beyond the hiring of a community health nurse to administer the HRA and the provision of a card reader and computer to analyze the data (109). These funding discrepancies resulted in wide variations in the substance and quality of health promotion efforts at Army installations worldwide. A December 1989 analysis of health promotion activities Army-wide found substantial discrepancies among the four major Army commands, with 100% of Training and Doctrine Command (TRADOC) installations having a full-time Fit-to-Win coordinator, as compared to 68% of Forces Command (FORSCOM) and 60% of Army Materiel Command (AMC) installations (109). The thorough adoption of the health promotion effort within the TRADOC command was likely the result of personal efforts by General Maxwell Thurman. Thurman had been in charge of ODCSPER in the early 1980s and was the driving force behind the corporate wellness study and the ARSTAF study that eventually evolved into the Army-wide health promotion program. Thurman left the Pentagon and became Commanding General for TRADOC in 1987. His charisma and commitment to the principles of the health promotion initiatives he fostered at the Pentagon were manifested in a well-supported health promotion initiative throughout TRADOC. Thurman's peers who led the other major Army commands did not embrace health promotion initiatives with the same charisma and commitment. Army-wide, this lack of a "strong command philosophy" manifested itself in the more modest implementation of health promotion activities (109).

The ODCSPER committees in charge of implementing health promotion activities did what they could to anticipate and mitigate the effects of these discrepancies, but the degree of support they could realistically provide was necessarily limited. For example, the committee that designed the Exportable Package produced a kit (the so-called "ammo box") containing a series of pamphlets and printed educational materials that addressed a number of health promotion topics (e.g., nutrition, smoking, stress management). Beyond this, however, there was no centrally administered financial support for programmatic interventions, nor was there any local mandate for installation commanders to provide them. In this environment, the number and quality of health promotion programs offered varied widely from installation to installation (35, 109, 122). An anecdotal analysis of the lessons learned from the Army's health promotion program suggests that,

at an installation where the personalities of the commander, the health promotion council, and the health promotion coordinator were enthusiastic, the ammo box played a small role in the success of the program. If the personalities were unenthusiastic, the ammo box became a larger part of the program, but the program was unlikely to be successful if the human factor was missing. Overall, it seemed to be the opinion that

too much emphasis was placed on the ammo box and its ability to *be* the health promotion program (italics in original) (109).

In principle, HRA respondents were supposed to receive a personalized report that reviewed their health risks. The community health nurse was supposed to review it with them in a one-on-one counseling session, and then refer them for interventions (e.g., smoking cessation, nutrition counseling), if needed. In some cases, the one-on-one counseling session with the nurse may have been the only educational or intervention support provided, unless there were health promotion programs or support initiatives locally available (122). It does not appear that the Army ever attempted to track how many soldiers were being referred for counseling or interventions.

It is not difficult to appreciate how these funding deficiencies and the decentralized nature of the program's administration may have impacted the success of health promotion activities, but it is important to understand that they also very likely had a negative impact on the quality of the HRA data that were collected. To some extent, the overall success of the HRA administration hinged on the stature accorded to the community health nurse, and unfortunately, this stature varied from installation to installation. Although the OTSG's committee had selected the questionnaire based, in part, on a criterion of low labor intensity, the HRA questionnaire did require physiologic metrics and anthropometric values such as the respondent's blood pressure, cholesterol, height, weight, and resting electrocardiogram. However, because administration of the health promotion program was allowed to vary locally, at installations that provided only lackluster support for the program, the nurse may not have been able to draw upon resources to assist in the proper collection of HRA data. The HRA database has high proportions of missing values with respect to some of these physiologic measures, and the distribution of others is suspect. For example, in the early days of the program, one of the first HRA project officers noted that the distribution of blood pressures was stepped at increments of 5 mm Hg, suggesting that it may have been self-reported rather than measured with a sphygmomanometer.⁶ The validation studies we reviewed in Chapter 2 point to the detrimental effect of self-reported data on the validity of the risk estimation scores. Without valid and reliable data on these variables, it is impossible to calculate valid and reliable risk scores. This means that the HRA's utility as a screening device, either for general health promotion purposes, or as the screening device for the Over-40 program, may have been compromised.

External Pressures

Apart from these issues surrounding proponentcy and implementation, there were external factors at work during the 1990s that drew attention and resources away from the Army's health promotion efforts. First were the joint phenomena of downsizing and increased tempo of operations. In 1988, the Secretary of Defense commissioned a bipartisan Commission on Base Realignment and Closure (BRAC). Since then, 125 major and 225 minor military facilities have been closed, and an additional 145 facilities

⁶ MAJ (ret) Ken Bush, personal communication, July 10, 2002.

have been “realigned.” Between 1989 and 1997, the DoD also reduced the size of the total active duty military force by 32%(7). This same time frame saw a dramatic increase in the number of military deployments. The four decades from 1950 to 1989 were characterized by less than a dozen major military deployments, whereas the 1990s saw more than 40 deployments (25). The nature of deployment missions also expanded significantly during this time, to include humanitarian assistance, counter-narcotic operations, and peacekeeping missions around the world. In short, the Army was trying to launch a comprehensive health promotion program in a time when everyone in the military was being asked to do more with fewer resources.

The continuing evolution of health promotion as a theoretical discipline also exerted influences on the implementation of the Army’s health promotion program and the role of the HRA within it. In 1991, Wilson and Howe reviewed the trajectory of sentiment in the professional literature of the 1980s with regard to health promotion efforts:

In the early part of the decade, the number of articles cited is quite small. . . . The years 1986-1988 are highwater marks for the wellness literature. . . . Some hints of frustration in the advocates for wellness and health promotion begin to appear in the mid to late 1980s. . . . Locus of control and compliance are important themes. . . . When viewed from a distance, collectively, the authors convey an image of great hope and promise for reducing morbidity due to lifestyle behaviors. . . . One then senses that the clients participating in the wellness programs experience difficulty in maintaining the lifestyle changes. . . . The literature suggests that the enchantment with the concept of wellness is now being tempered by reality and the complex issue of assisting clients in changing their lifestyles and behaviors.

In the absence of a research program that produced unequivocal proof that health promotion efforts resulted in cost savings, the Army may have had difficulty defending the orientation of its health promotion program around personal wellness.

LESSONS LEARNED IN THE DEVELOPMENT OF THE HRA QUESTIONNAIRE

The lessons the Army learned in its experience with the health promotion program include gaining expertise in the design, development, and implementation of health habit survey instruments, and valuable capabilities in analyzing the data gathered with such tools. This section gives a brief overview of some things the Army could have improved in the development of the instrument, and articulates some lessons they might apply to the development of future survey instruments.

- One of the keys to designing a good survey instrument is beginning with a clearly defined set of objectives. It is necessary to avoid the temptation to add extra questions on topics that are not related to specific project objectives (52). This temptation may be even more difficult to resist if the survey instrument is being designed by a committee, as each party brings to the table their own interests

and may lobby for the inclusion of additional survey items. As stated earlier, the Army's HRA was based on risk calculation methodology from the CDC's HRA and from epidemiologic studies such as the Framingham Heart Study. There are many items on the survey, however, that do not figure into the calculation of the risk scores, and this may have blurred the objectives of the questionnaire.

- A related issue surrounding survey objectives concerns the administration of the HRA in multiple contexts. In the final analysis, it may not have been the wisest decision to use the same data collection instrument for the HRA as for the Over-40 program. If these two programs had been kept separate and if the proponents had developed and maintained separate instruments for them, both programs might have been better off. The use of a single instrument in multiple contexts also highlights some issues in biases of self-reported health data. Soldiers who are responding to questions about health habits in the context of preparing for a routine physical exam may be more forthright about some of their habits, especially if they believe that the physician may use this information to guide decisions about their care. In contrast, a soldier who is reporting to a new post or duty assignment may be less than candid about revealing some health habits.
- Development of survey questionnaires should be more rigorously documented. If existing questions are used in the construction of a new survey instrument, the decision to include them should be made in careful consideration of the flaws of the original question and how well they are likely to perform in the target population (37). The military may be somewhat limited in adapting items from existing sources, and may be inclined to either borrow exclusively from public domain sources or to write new questions, even on topics that have been well studied in the survey literature. Taking the time to document the decision-making process, on such issues as when and whether to borrow items, or to use public domain items, or even to write new questions is a useful exercise in making sure the instrument stays true to its stated purpose and objectives and in arriving at the best questions to gather the information desired. Fortunately, this lesson seems to have been adopted by at least two teams currently launching new military survey projects: the Millennium Cohort Study team and the team developing the HEAR. The Millennium Cohort Study is a prospective study of the impact of deployment on soldier health, and is being conducted primarily through postal surveys (58). The authors of the questionnaire have relied heavily on existing survey scales (e.g., SF-36, Patient Health Questionnaire). The authors of the HEAR have likewise taken care to identify which items have been adopted from other sources.
- Because military personnel change jobs frequently, and because many of the individuals involved in the creation of the health promotion program were officers in mid-career, it has been enormously challenging to learn even the names of many of the key players in the early days of the program, much less their current whereabouts. This dissipation of the institutional knowledge on such a high profile, Army-wide project has rendered it difficult for researchers who want to

use HRA data. If more scrupulous attention had been paid to documentation during the early days of the program, it would be easier to obtain HRA data and to fully understand more of the idiosyncrasies expressed in the HRA database.

- New survey questionnaires should be rigorously piloted and pretested, and the results of these pilot experiments should be published or documented in reports. It is important that survey research experts, as well as content experts, be involved in the development phase. Pretesting should occur several times over the development of the questionnaire, with the goal of clarifying any questions that are confusing, refining the flow and order of questions, and correcting any problems with the logic of skip patterns (37). Questionnaires should be pretested in a sample that resembles the targeted population and among a large enough group of subjects to permit subgroup analyses, if relevant (e.g., by race/ethnicity and gender). If pretesting indicates that refinements are needed to the questionnaire, these decisions should be documented carefully. Although the HRA questionnaire was pretested and piloted on six U.S. bases, there are no reports documenting the results of these evaluations, or what changes, if any, were made to the questionnaire in response to the findings.
- In the pilot phase, questionnaires should be formally evaluated with respect to the reliability and validity of the responses they garner. As outlined in Chapter 2 of this report, there are many different facets of reliability and validity and different means of assessing each. Here again, findings should guide the refinement of the survey instrument and should be documented scrupulously.

As noted elsewhere in this report, the HRA was not intended as a research tool, per se, but has yielded a great wealth of information that is potentially useful in surveillance and research. This database could have been even more useful if the creators had exercised greater planning and foresight in the design and management of the original questionnaire. Furthermore, although the HRA may be a “dying” instrument, and has been supplanted by the HEAR, the lessons learned in this painstaking and thorough review of the HRA questionnaire can be used to better inform the development of future self reporting tools, whether intended for research (e.g., the Millennium Cohort Study), baseline health assessment (e.g., Recruit Assessment Program), or health care planning and health promotion (e.g., HEAR).

REFERENCES

1. Adams, W. L., K. L. Barry, and M. F. Fleming. Screening for problem drinking in older primary care patients. *Jama*. 276(24): 1964-1967, 1996.
2. Aithal, G. P., H. Thornes, A. D. Dwarakanath, and A. R. Tanner. Measurement of carbohydrate-deficient transferrin (CDT) in a general medical clinic: is this test useful in assessing alcohol consumption. *Alcohol Alcohol*. 33(3): 304-309, 1998.
3. Alexander, G. Health risk appraisal. In: *Handbook of Health Assessment Tools*, edited by G. C. Hyner, K. W. Peterson, J. W. Travis, J. E. Dewey, J. J. Foerster, and E. M. Framer. Pittsburgh, PA: Society of Prospective Medicine, 1999.
4. Anda, R. F., P. L. Remington, D. L. Dodson, P. J. DeGuire, M. R. Forman, and R. A. Gunn. Patterns of self-reported drinking and driving in Michigan. *Am J Prev Med*. 3(5): 271-275, 1987.
5. Arday, D. R., S. L. Tomar, D. E. Nelson, R. K. Merritt, M. W. Schooley, and P. Mowery. State smoking prevalence estimates: a comparison of the Behavioral Risk Factor Surveillance System and current population surveys. *Am J Public Health*. 87(10): 1665-1669, 1997.
6. Baker, F., S. R. Ainsworth, J. T. Dye, C. Crammer, M. J. Thun, D. Hoffmann, J. L. Repace, J. E. Henningfield, J. Slade, J. Pinney, T. Shanks, D. M. Burns, G. N. Connolly, and D. R. Shopland. Health risks associated with cigar smoking. *Jama*. 284(6): 735-740, 2000.
7. *Base Realignment and Closure (BRAC)*. Global Security. Available at: <http://198.65.138.161/military/facility/brac.htm>. Last updated: August 8, 2002. Date accessed: August 21, 2002.
8. Bean, J. A., J. D. Leeper, R. B. Wallace, B. M. Sherman, and H. Jagger. Variations in the reporting of menstrual histories. *Am J Epidemiol*. 109(2): 181-185, 1979.
9. Beery, W., V. J. Schoenbach, and E. H. Wagner. *Health Risk Appraisal: Methods and Programs, with Annotated Bibliography*. Rockville, MD: National Center for Health Services Research and Health Care Technology Assessment. PHS 86-3396, January 1986.
10. Begg, D. J., and J. D. Langley. Seat-belt use and related behaviors among young adults. *J Safety Res*. 31(3): 211-220, 2000.
11. Bell, N. S., P. J. Amoroso, M. M. Yore, G. S. Smith, and B. H. Jones. Self-reported risk-taking behaviors and hospitalization for motor vehicle injury among active duty Army personnel. *Am J Prev Med*. 18(3 Suppl): 85-95, 2000.
12. Bell, N. S., J. O. Williams, L. Senier, P. J. Amoroso, and S. R. Strowman. *The U.S. Army's Health Risk Appraisal (HRA) Survey Part II: Generalizability, Sample Selection, and Bias*. Natick, MA: U.S. Army Research Institute of Environmental Medicine. Technical Report T02/18, June 2002.
13. Bell, N. S., J. O. Williams, L. Senier, S. R. Strowman, and P. J. Amoroso. The reliability and validity of the self-reported drinking measures in the Army's Health Risk Appraisal survey. *Alcohol Clin Exp Res*. 27(5): 826-834, 2003.
14. Bisson, J. I., and A. Milford-Ward. A comparison of carbohydrate deficient transferrin with other markers of alcohol misuse in male soldiers under the age of thirty. *Alcohol Alcohol*. 29(3): 315-321, 1994.
15. Bohrnstedt, G. W. Measurement. In: *Handbook of Survey Research*, edited by P. H. Rossi, J. D. Wright, and A. B. Anderson. San Diego, CA: Academic Press, 1983.
16. Bowlin, S. J., B. D. Morrill, A. N. Nafziger, P. L. Jenkins, C. Lewis, and T. A. Pearson. Validity of cardiovascular disease risk factors assessed by telephone survey: the Behavioral Risk Factor Survey. *J Clin Epidemiol*. 46(6): 561-571, 1993.

17. Bowlin, S. J., B. D. Morrill, A. N. Nafziger, C. Lewis, and T. A. Pearson. Reliability and changes in validity of self-reported cardiovascular disease risk factors using dual response: the behavioral risk factor survey. *J Clin Epidemiol.* 49(5): 511-517, 1996.
18. Bowman, J. A., S. Redman, J. A. Dickinson, R. Gibberd, and R. W. Sanson-Fisher. The accuracy of Pap smear utilization self-report: a methodological consideration in cervical screening research. *Health Serv Res.* 26(1): 97-107, 1991.
19. Bowman, J. A., R. Sanson-Fisher, and S. Redman. The accuracy of self-reported Pap smear utilisation. *Soc Sci Med.* 44(7): 969-976, 1997.
20. Bradburn, N. M. Response effects. In: *Handbook of Survey Research*, edited by P. H. Rossi, J. D. Wright, and A. B. Anderson. San Diego, CA: Academic Press, 1983.
21. Branch, L. G., and D. J. Rabiner. Rediscovering the patient's role in receiving health promotion services. *Med Care.* 38(1): 70-77, 2000.
22. Branch, L. G., D. J. Rabiner, P. Patterson, and R. J. Sullivan, Jr. Prevention services received by veterans visiting VHA facilities. *Prev Med.* 27(4): 604-610, 1998.
23. Bray, R. M., R. P. Sanches, M. L. Ornstein, D. Lentine, A. A. Vincus, T. U. Baird, J. A. Walker, S. C. Wheelless, L. L. Guess, L. A. Kroutil, and V. G. Iannacchione. *1998 Department of Defense Survey of Health Related Behaviors Among Military Personnel*. Research Triangle Park, NC: Research Triangle Institute. 1998.
24. Brownson, R. C., J. Jackson-Thompson, J. C. Wilkerson, and F. Kiani. Reliability of information on chronic disease risk factors collected in the Missouri Behavioral Risk Factor Surveillance System. *Epidemiology.* 5(5): 545-549, 1994.
25. Buckner, J. The Changing Nature of Stress as it Relates to "Army Vision 2010," Including the Peacekeeping Mission and Roles of U.S. Army Reserve and National Guard. Presentation at: Leadership and Operational Stress Conference June 20, 2000; Fort McNair, Washington, D.C.
26. Bush, B., S. Shaw, P. Cleary, T. L. Delbanco, and M. D. Aronson. Screening for alcohol abuse using the CAGE questionnaire. *Am J Med.* 82(2): 231-235, 1987.
27. Campbell, N. R., M. J. Ashley, S. G. Carruthers, Y. Lacourciere, and D. W. McKay. Lifestyle modifications to prevent and control hypertension. 3. Recommendations on alcohol consumption. Canadian Hypertension Society, Canadian Coalition for High Blood Pressure Prevention and Control, Laboratory Centre for Disease Control at Health Canada, Heart and Stroke Foundation of Canada. *Cmaj.* 160(9 Suppl): S13-20, 1999.
28. Cherpitel, C. J. Analysis of cut points for screening instruments for alcohol problems in the emergency room. *J Stud Alcohol.* 56(6): 695-700, 1995.
29. Cherpitel, C. J. Comparison of screening instruments for alcohol problems between black and white emergency room patients from two regions of the country. *Alcohol Clin Exp Res.* 21(8): 1391-1397, 1997.
30. Cherpitel, C. J. Differences in performance of screening instruments for problem drinking among blacks, whites and Hispanics in an emergency room population. *J Stud Alcohol.* 59(4): 420-426, 1998.
31. Cherpitel, C. J. Screening for alcohol problems in the emergency department. *Ann Emerg Med.* 26(2): 158-166, 1995.
32. Cherpitel, C. J., and W. B. Clark. Ethnic differences in performance of screening instruments for identifying harmful drinking and alcohol dependence in the emergency room. *Alcohol Clin Exp Res.* 19(3): 628-634, 1995.
33. Cleary, P. J. A checklist for life event research. *J Psychosom Res.* 24(3-4): 199-207, 1980.
34. Cohen, S., T. Kamarck, and R. Mermelstein. A global measure of perceived stress. *J Health Soc Behav.* 24(4): 385-396, 1983.

35. Collins, B. S., and S. H. Custis. Health promotion in a shrinking military: the call for structural integration and a conceptual systems approach. *Mil Med.* 158(6): 386-391, 1993.
36. Comparison of observed and self-reported seat belt use rates --- United States. *MMWR Morbid Mortal Weekly Rep.* 37(36): 549-551, 1988.
37. Converse, J. M., and S. Presser. *Survey Questions: Handcrafting the Standardized Questionnaire*. Newbury Park, CA: SAGE Publications, Inc., 1986.
38. Degnan, D., R. Harris, J. Ranney, D. Quade, J. A. Earp, and J. Gonzalez. Measuring the use of mammography: two methods compared. *Am J Public Health.* 82(10): 1386-1388, 1992.
39. Department of Defense. *Health Promotion*. Washington, D.C. Directive 1010.10, March 11, 1986.
40. Department of Defense. *Physical Fitness and Body Fat Program*. Washington, D.C. Directive 1308.1, June 29, 1981.
41. Department of the Army. *Army Health Promotion*. Washington, D.C. AR 600-63, November 17, 1987.
42. Department of the Army. *The Army Physical Fitness Program*. Washington, D.C. 1989.
43. Department of the Army. *Standards of Medical Fitness*. Washington, D.C. AR 40-501, 1998.
44. Derogatis, L. R., and H. L. Coons. Self-report measures of stress. In: *Handbook of Stress: Theoretical and Clinical Aspects*, (2nd ed.), edited by L. Goldberger and S. Breznitz. New York, NY: The Free Press, 1993.
45. Edington, D. W., L. Yen, and A. Braunstein. The reliability and validity of HRAs. In: *Handbook of Health Assessment Tools*, edited by G. C. Hyner, K. W. Peterson, J. W. Travis, J. E. Dewey, J. J. Foerster, and E. M. Framer. Pittsburgh, PA: Society of Prospective Medicine, 1999.
46. Ewing, J. A. Detecting alcoholism. The CAGE questionnaire. *Jama.* 252(14): 1905-1907, 1984.
47. Fertig, J. B., and J. P. Allen. Health behavior correlates of hazardous drinking by Army personnel. *Mil Med.* 161(6): 352-355, 1996.
48. Fertig, J. B., J. P. Allen, and G. M. Cross. CAGE as a predictor of hazardous alcohol consumption in U.S. Army personnel. *Alcohol Clin Exp Res.* 17(6): 1184-1187, 1993.
49. Feunekes, G. I., P. van 't Veer, W. A. van Staveren, and F. J. Kok. Alcohol intake assessment: the sober facts. *Am J Epidemiol.* 150(1): 105-112, 1999.
50. Fitzpatrick, D. T., and S. Shannon. Health-risk behaviors of Army aircrew. *J Occup Med.* 34(8): 810-814, 1992.
51. Foss, R. D., D. J. Beirness, and K. Sprattler. Seat belt use among drinking drivers in Minnesota. *Am J Public Health.* 84(11): 1732-1737, 1994.
52. Fowler, F. J. *Survey Research Methods*. Newbury Park, CA: Sage Publications, Inc., 1993.
53. *Frequently Asked Questions*. National Institute on Alcohol Abuse and Alcoholism. Available at: <http://www.niaaa.nih.gov/faq/q-a.htm#question13>. Last updated: January 1998. Date accessed: September 25, 2001.
54. Fruchter, R. G., K. Rones, T. Roth, C. A. Webber, L. Camilien, and J. G. Boyce. Pap smear histories in a medical clinic: accuracy of patients' self-reports. *N Y State J Med.* 92(10): 421-424, 1992.
55. Gazmararian, J. A., B. Foxman, L. T. Yen, H. Morgenstern, and D. W. Edington. Comparing the predictive accuracy of health risk appraisal: the Centers for Disease Control versus Carter Center program. *Am J Public Health.* 81(10): 1296-1301, 1991.
56. Goldberg, E. L., and G. W. Comstock. Life events and subsequent illness. *Am J Epidemiol.* 104(2): 146-158, 1976.

57. Gordon, N. P., R. A. Hiatt, and D. I. Lampert. Concordance of self-reported data and medical record audit for six cancer screening procedures. *J Natl Cancer Inst.* 85(7): 566-570, 1993.
58. Gray, G. C., K. B. Chesbrough, M. A. Ryan, P. Amoroso, E. J. Boyko, G. D. Gackstetter, T. I. Hooper, and J. R. Riddle. The millennium Cohort Study: a 21-year prospective cohort study of 140,000 military personnel. *Mil Med.* 167(6): 483-488, 2002.
59. Gross, W. C., and R. E. Billingham. Alcohol consumption and sexual victimization among college women. *Psychological Reports.* 82(1): 80-82, 1998.
60. Heather, N., and I. Robertson. *Problem Drinking*. Oxford: Oxford University Press, 1997.
61. Heck, E. J. Developing a screening questionnaire for problem drinking in college students. *J Am Coll Health.* 39(5): 227-231, 1991.
62. Heck, E. J., and M. D. Williams. Using the CAGE to screen for drinking-related problems in college students. *J Stud Alcohol.* 56(3): 282-286, 1995.
63. Jackson, C., D. E. Jatulis, and S. P. Fortmann. The Behavioral Risk Factor Survey and the Stanford Five-City Project Survey: a comparison of cardiovascular risk behavior estimates. *Am J Public Health.* 82(3): 412-416, 1992.
64. *JNC 7 Express: The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure*. Bethesda, MD: National Heart, Lung, and Blood Institute. NIH Publication No. 03-5233, May 2003.
65. Jonah, B. A., R. Thiessen, and E. Au-Yeung. Sensation seeking, risky driving and behavioral adaptation. *Accid Anal Prev.* 33(5): 679-684, 2001.
66. Kerber, R. A., and M. L. Slattery. Comparison of self-reported and database-linked family history of cancer data in a case-control study. *Am J Epidemiol.* 146(3): 244-248, 1997.
67. King, E. S., B. K. Rimer, B. Trock, A. Balshem, and P. Engstrom. How valid are mammography self-reports? *Am J Public Health.* 80(11): 1386-1388, 1990.
68. King, M. At risk drinking among general practice attenders: validation of the CAGE questionnaire. *Psychol Med.* 16(1): 213-217, 1986.
69. Lawrence, V. A., C. De Moor, and M. E. Glenn. Systematic differences in validity of self-reported mammography behavior: A problem for intergroup comparisons? *Prev Med.* 29(6 Pt 1): 577-580, 1999.
70. Lee, D. J., and R. S. DeFrank. Interrelationships among self-reported alcohol intake, physiological indices and alcoholism screening measures. *J Stud Alcohol.* 49(6): 532-537, 1988.
71. Litwin, M. S. *How to Measure Survey Reliability and Validity*. Thousand Oaks, CA: Sage Publications, Inc., 1995.
72. Luepker, R. V., U. E. Pallonen, D. M. Murray, and P. L. Pirie. Validity of telephone surveys in assessing cigarette smoking in young adults. *Am J Public Health.* 79(2): 202-204, 1989.
73. Luttrell, S., V. Watkin, G. Livingston, Z. Walker, P. D'Ath, P. Patel, S. Shergill, A. Dain, C. Bielawska, and C. Katona. Screening for alcohol misuse in older people. *Int J Geriatr Psychiatry.* 12(12): 1151-1154, 1997.
74. Mayfield, D., G. McLeod, and P. Hall. The CAGE questionnaire: validation of a new alcoholism screening instrument. *Am J Psychiatry.* 131(10): 1121-1123, 1974.
75. McGovern, P. G., N. Lurie, K. L. Margolis, and J. S. Slater. Accuracy of self-report of mammography and Pap smear in a low-income urban population. *Am J Prev Med.* 14(3): 201-208, 1998.
76. McKenna, M. T., M. Speers, K. Mallin, and R. Warnecke. Agreement between patient self-reports and medical records for Pap smear histories. *Am J Prev Med.* 8(5): 287-291, 1992.

77. Michielutte, R., M. B. Dignan, H. B. Wells, J. Bahnson, M. Smith, R. Wooten, and L. N. Hale. Errors in reporting cervical screening among public health clinic patients. *J Clin Epidemiol.* 44(4-5): 403-408, 1991.
78. Midanik, L. The validity of self-reported alcohol consumption and alcohol problems: a literature review. *Br J Addict.* 77(4): 357-382, 1982.
79. Midanik, L. T. Validity of self-reported alcohol use: a literature review and assessment. *Br J Addict.* 83(9): 1019-1030, 1988.
80. Midanik, L. T., and R. Room. The epidemiology of alcohol consumption. *Alcohol Health Res World.* 16(3): 183-190, 1992.
81. Montano, D. E., and W. R. Phillips. Cancer screening by primary care physicians: a comparison of rates obtained from physician self-report, patient survey, and chart audit. *Am J Public Health.* 85(6): 795-800, 1995.
82. Pagano, M., and K. Gauvreau. *Principles of Biostatistics.* Belmont, CA: Duxbury Press, 1993.
83. Parada, M. A., L. D. Cohn, E. Gonzalez, T. Byrd, and M. Cortes. The validity of self-reported seatbelt use: Hispanic and non-Hispanic drivers in El Paso. *Accid Anal Prev.* 33(1): 139-143, 2001.
84. Pileire, B., J. Bredent-Bangou, and M. Valentino. Comparison of questionnaire and biochemical markers to detect alcohol abuse in a West Indian population. *Alcohol Alcohol.* 26(3): 353-359, 1991.
85. Polich, J. M., and B. R. Orvis. *Alcohol Problems: Patterns and Prevalence in the U.S. Air Force.* Santa Monica, CA: The RAND Corporation. Technical report (DTIC ADA074740), June 1979.
86. Preusser, D. F., A. F. Williams, and A. K. Lund. Characteristics of belted and unbelted drivers. *Accid Anal Prev.* 23(6): 475-482, 1991.
87. *Prostate Cancer: The Public Health Perspective.* Centers for Disease Control and Prevention. Available at: <http://www.cdc.gov/cancer/prostate/prospdf/proaag01.pdf>. Date accessed: August 20, 2001.
88. Rao, V., B. Bataglia, C. Blevins, W. Grove, C. Hayn, L. LaMarca, S. Nagel, and J. Roberts. *Quantitative Analysis of the Army Health Promotion Program Health Risk Appraisal Database, Summary Report: 1991-1995.* Frederick, MD: Science Applications International Corporation. 1997.
89. Rennix, C. P., M. M. Yore, and P. Amoroso. *Organic Chemical Exposure and the Risk of Breast Cancer among Active Duty Women in the U.S. Army, 1980-1996.* Natick, MA: U.S. Army Research Institute of Environmental Medicine. Technical Report T02-05, October 2001.
90. Robbins, A. S., V. P. Fonseca, S. Y. Chao, G. A. Coil, N. S. Bell, and P. J. Amoroso. Short-term effects of cigarette smoking on hospitalisation and associated lost workdays in a young healthy population. *Tob Control.* 9(4): 389-396, 2000.
91. Robertson, L. S. The validity of self-reported behavioral risk factors: seatbelt and alcohol use. *J Trauma.* 32(1): 58-59, 1992.
92. Rothman, K. J., and S. Greenland. *Modern Epidemiology.* (2nd ed.). Philadelphia, PA: Lippincott, Williams & Wilkins, 1998.
93. Sanchez, R. P., and R. M. Bray. Cigar and pipe smoking in the U.S. military: prevalence, trends, and correlates. *Mil Med.* 166(10): 903-908, 2001.
94. SAS help file. Version 8.2. Cary, NC: SAS Institute, Inc., 2001.
95. Saunders, J. B., O. G. Aasland, T. F. Babor, J. R. de la Fuente, and M. Grant. Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO Collaborative Project on Early Detection of Persons with Harmful Alcohol Consumption--II. *Addiction.* 88(6): 791-804, 1993.

96. Sawyer, J. A., J. A. Earp, R. H. Fletcher, F. F. Daye, and T. M. Wynn. Accuracy of women's self-report of their last Pap smear. *Am J Public Health*. 79(8): 1036-1037, 1989.
97. Schaeffer, N. C. Asking questions about threatening topics: a selective overview. In: *The Science of Self-Report*, edited by A. A. Stone, J. S. Turkkan, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, and V. S. Cain. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.
98. Schoenbach, V. J. Appraising health risk appraisal. *Am J Public Health*. 77(4): 409-411, 1987.
99. Schroeder, D. H., and P. T. Costa, Jr. Influence of life event stress on physical illness: substantive effects or methodological flaws? *J Pers Soc Psychol*. 46(4): 853-863, 1984.
100. *Screening for Prostate Cancer*. National Cancer Institute. Available at: http://cancernet.nci.nih.gov/cgi-bin/srchcgi.exe?DBID=pdq&TYPE=search&SFMT=pdq_statement/1/0/0&Z208=208_04727H. Last updated: August 2001. Date accessed: August 20, 2001.
101. Serdula, M. K., A. H. Mokdad, T. Byers, and P. Z. Siegel. Assessing alcohol consumption: beverage-specific versus grouped-beverage questions. *J Stud Alcohol*. 60(1): 99-102, 1999.
102. Shea, S., A. D. Stein, R. Lantigua, and C. E. Basch. Reliability of the behavioral risk factor survey in a triethnic population. *Am J Epidemiol*. 133(5): 489-500, 1991.
103. Smith, K. W., S. M. McKinlay, and J. B. McKinlay. The reliability of health risk appraisals: a field trial of four instruments. *Am J Public Health*. 79(12): 1603-1607, 1989.
104. Smith, K. W., S. M. McKinlay, and J. B. McKinlay. The validity of health risk appraisals for coronary heart disease: results from a randomized field trial. *Am J Public Health*. 81(4): 466-470, 1991.
105. Smith, P. F., P. L. Remington, D. F. Williamson, and R. F. Anda. A comparison of alcohol sales data with survey data on self-reported alcohol use in 21 states. *Am J Public Health*. 80(3): 309-312, 1990.
106. Stein, A. D., R. I. Lederman, and S. Shea. The Behavioral Risk Factor Surveillance System questionnaire: its reliability in a statewide sample. *Am J Public Health*. 83(12): 1768-1772, 1993.
107. Stein, A. D., R. I. Lederman, and S. Shea. Reproducibility of the women's module of the Behavioral Risk Factor Surveillance System questionnaire. *Ann Epidemiol*. 6(1): 47-52, 1996.
108. Steinweg, D. L., and H. Worth. Alcoholism: the keys to the CAGE. *Am J Med*. 94(5): 520-523, 1993.
109. Stevenson, M. O. *A decade of change: a historical and anecdotal record of cardiovascular health promotion in the United States Army, 1989-1989* [Doctoral dissertation]. Washington, D.C.: The American University, 1990.
110. Suarez, L., D. A. Goldman, and N. S. Weiss. Validity of Pap smear and mammogram self-reports in a low-income Hispanic population. *Am J Prev Med*. 11(2): 94-98, 1995.
111. Thompson, D. C., F. P. Rivara, R. S. Thompson, P. M. Salzberg, M. E. Wolf, and D. C. Pearson. Use of behavioral risk factor surveys to predict alcohol-related motor vehicle events. *Am J Prev Med*. 9(4): 224-230, 1993.
112. U.S. Army Health Risk Appraisal System (HRA) End User Manual, Version 5.01. AISM 18-HE1-RWF-ZZZ-EM. San Antonio, TX: U.S. Army Health Care Systems Support Activity, 1994.
113. Urbano-Marquez, A., R. Estruch, J. Fernandez-Sola, J. M. Nicolas, J. C. Pare, and E. Rubin. The greater risk of alcoholic cardiomyopathy and myopathy in women compared with men. *Jama*. 274(2): 149-154, 1995.
114. Vegega, M. E., and M. D. Klitzner. Drinking and driving among youth: a study of situational risk factors. *Health Educ Q*. 16(3): 373-388, 1989.

115. Walter, S. D., E. A. Clarke, J. Hatcher, and L. W. Stitt. A comparison of physician and patient reports of Pap smear histories. *J Clin Epidemiol.* 41(4): 401-410, 1988.
116. Wechsler, H., and S. B. Austin. Binge drinking: the five/four measure. *J Stud Alcohol.* 59(1): 122-124, 1998.
117. Wechsler, H., A. Davenport, G. Dowdall, B. Moeykens, and S. Castillo. Health and behavioral consequences of binge drinking in college. A national survey of students at 140 campuses. *Jama.* 272(21): 1672-1677, 1994.
118. Wechsler, H., G. W. Dowdall, A. Davenport, and E. B. Rimm. A gender-specific measure of binge drinking among college students. *Am J Public Health.* 85(7): 982-985, 1995.
119. Wechsler, H., and N. Isaac. Binge drinkers at Massachusetts colleges. Prevalence, drinking style, time trends, and associated problems. *Jama.* 267(21): 2929-2931, 1992.
120. Wetterling, T., R. D. Kanitz, H. J. Rumpf, U. Hapke, and D. Fischer. Comparison of CAGE and MAST with the alcohol markers CDT, gamma-GT, ALAT, ASAT and MCV. *Alcohol Alcohol.* 33(4): 424-430, 1998.
121. Williams, J. O., N. S. Bell, and P. J. Amoroso. Drinking and other risk-taking behaviors of enlisted male soldiers in the U.S. Army. *Work.* 18(2): 141-150, 2002.
122. Wilson, W. L., and R. S. Howe. The promotion of wellness. In: *Handbook of Military Psychology*, edited by R. Gal and A. D. Mangelsdorff. New York, NY: John Wiley & Sons, 1991.
123. Yore, M. M., N. S. Bell, L. Senier, and P. J. Amoroso. Progress toward attainment of the Healthy People 2000 objectives in the U.S. Army. Measured by health risk appraisal results. *Am J Prev Med.* 19(2): 87-93, 2000.
124. Zador, P. L. Alcohol-related relative risk of fatal driver injuries in relation to driver age and sex. *J Stud Alcohol.* 52(4): 302-310, 1991.
125. Zapka, J. G., C. Bigelow, T. Hurley, L. D. Ford, J. Egelhofer, W. M. Cloud, and E. Sachsse. Mammography use among sociodemographically diverse women: the accuracy of self-report. *Am J Public Health.* 86(7): 1016-1021, 1996.

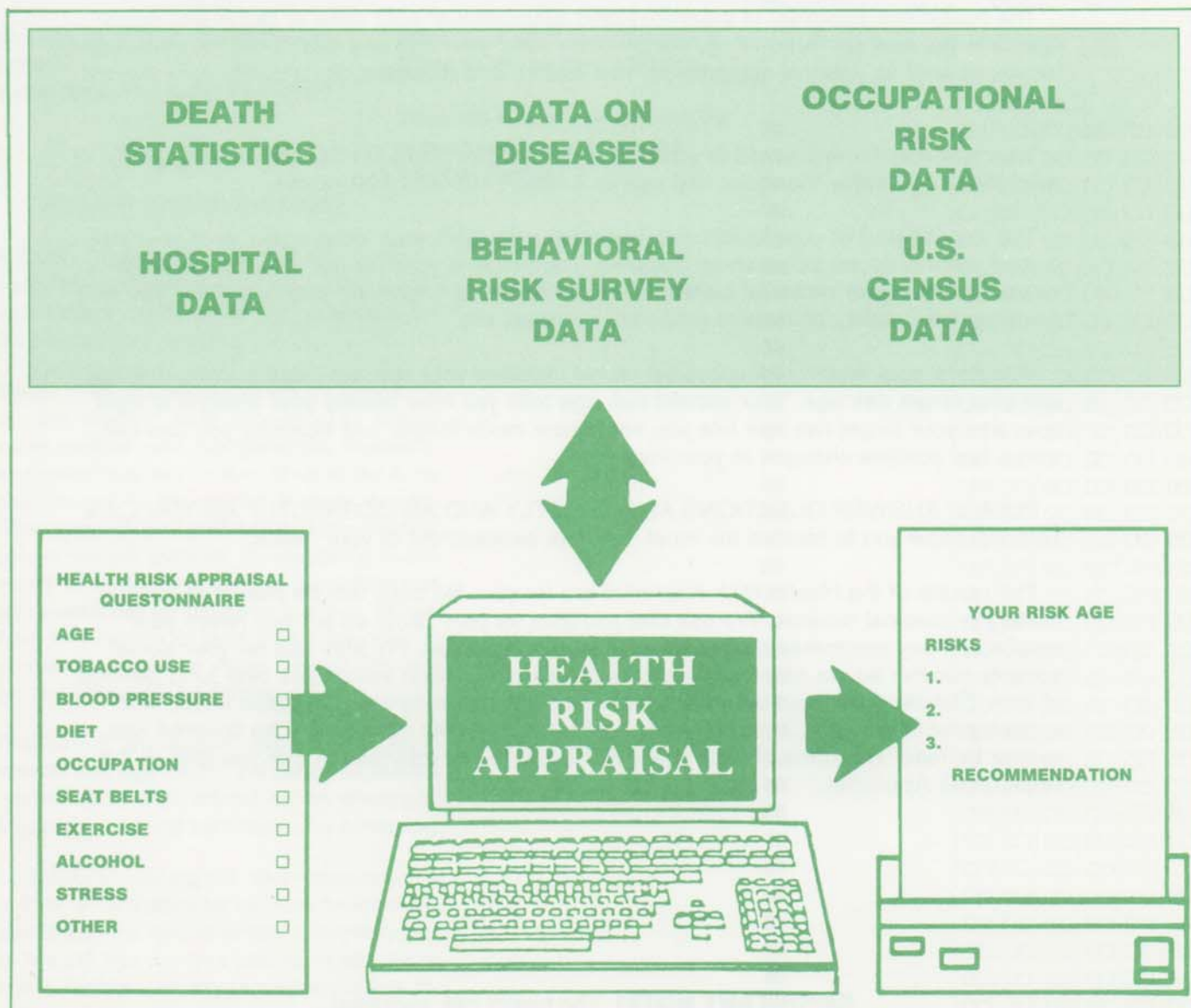
APPENDIX A: THE HRA QUESTIONNAIRE

THE ARMY HEALTH PROMOTION PROGRAM

Fit to Win



HEALTH RISK APPRAISAL



For use of this form, see AR40-501 and AR600-63; the proponent agency is TSG

UNITED STATES ARMY

FIT TO WIN



The **HEALTH RISK APPRAISAL** is an activity of
THE ARMY HEALTH PROMOTION PROGRAM

How does the Health Risk Appraisal work?

The health risk appraisal is a personalized estimation of your risks of death and major illness in the next ten years. First, the program uses your age and health-related personal habits, as well as national statistics on risk factors and diseases, to calculate your current risks.

Your risk may be expressed in terms of **RISK AGE** or **HEALTH SCORE**. Ideally, you want a risk age lower than your real age or a health score of 100 points.

The second part of your health risk appraisal calculates your risks again, as if your risk factors were reduced as much as possible. The result is your "target" risk age or health score. It shows your potential benefit, in health terms, of improving your lifestyle-if you quit smoking, wear safety belts, take moderate exercise, etc.

Therefore, your health risk appraisal report includes your real age, your current risk age and your target risk age. Your current risk age tells you how healthy your lifestyle is right now, and your target risk age lets you know how much longer and healthier you can live with a few positive changes in your lifestyle.

PLEASE ANSWER QUESTIONS AS HONESTLY AND AS CORRECTLY AS YOU CAN. This will allow you to receive the most accurate assessment of your health.

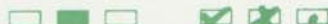
The results of the Health Risk Appraisal are for you. No copy will be placed in your military or medical records. We ask that you give us your name so we can return your results and any recommendations for follow-up care to you. We also ask for your social security number so we can statistically track trends in health awareness over long periods of time. Statistical information may be collected from an armywide database which will contain your information, but your name and social security number will be covered and cannot be read. The rules of the Privacy Act apply to any information that you give in the Health Risk Appraisal.

IMPORTANT NOTE! The health risk appraisal is no substitute for a physical examination or check-up. It will not give you a diagnosis nor will it tell you how long you will actually live. However, the health risk appraisal will help you understand and recognize your risk factors.

INSTRUCTIONS

Please use a No. 2 Pencil only to complete this survey. Make dark, black marks that fill the response boxes completely.

EXAMPLE: Correct Incorrect



Health Risk Appraisal (HRA)
for use of this form, see
AR40-501 and AR600-63;
the proponent is TSG

For MILITARY ONLY: Complete Questions 1-4.

1. What is your branch of service?

1. ☐ U.S. Army ☐ U.S. Marines
☐ U.S. Navy ☐ U.S. Coast Guard
☐ U.S. Air Force ☐ Other

2. What is your military status?

2. ☐ Regular Army ☐ USAR
☐ USAR/AGR ☐ ARNG
☐ ARNG/AGR ☐ Other

3. What is your current rank?

3.

ENLISTED		OFFICER		WARR. OFFIC.
☐ E-1	☐ E-6	☐ O-1	☐ O-6	☐ WO-1
☐ E-2	☐ E-7	☐ O-2	☐ O-7	☐ WO-2
☐ E-3	☐ E-8	☐ O-3	☐ O-8	☐ WO-3
☐ E-4	☐ E-9	☐ O-4	☐ O-9	☐ WO-4
☐ E-5		☐ O-5	☐ O-10	

4. What is your Unit Identification Code?
(Enter Specific Unit Identifier)

Print your Unit Identification Code in these blank boxes.

Then fill in the corresponding response box below each number/letter.

PRIVACY ACT STATEMENT

AUTHORITY: 29 CFR Chapter XVII, Occupational Safety and Health Standards; 5 U.S.C., section 150; Executive Orders 11612 and 11807 authorize the collection of this information.

PURPOSE: The primary use of this information is by the unit medical care providers to assure competent medical care. Additional disclosures of this information may be: To the Office of the Army Surgeon General in aggregated form to develop Army/Command fitness profiles; to Army medical researchers for the purpose of correlating health precursors to health problems or to commercial medical researchers for the same purpose. Where data from this system of records are provided to agencies external to the Army, Social Security Number and Name will be deleted.

ROUTINE USES: Information may be disclosed to departments and agencies of the Executive Branch in performance of their official duties relating to health risk appraisal and cardiovascular screening.

DISCLOSURE: Furnishing the information required on this form is mandatory for all Department of the Army active duty and reserve component military personnel. We ask that you give your name so we can return your results and any recommendations for follow-up care to you. We also ask for your social security number so we can statistically track trends in health awareness over long periods of time.

4.

UNIT CODE					
A	A	A	A	A	A
B	B	B	B	B	B
C	C	C	C	C	C
D	D	D	D	D	D
E	E	E	E	E	E
F	F	F	F	F	F
G	G	G	G	G	G
H	H	H	H	H	H
I	I	I	I	I	I
J	J	J	J	J	J
K	K	K	K	K	K
L	L	L	L	L	L
M	M	M	M	M	M
N	N	N	N	N	N
O	O	O	O	O	O
P	P	P	P	P	P
Q	Q	Q	Q	Q	Q
R	R	R	R	R	R
S	S	S	S	S	S
T	T	T	T	T	T
U	U	U	U	U	U
V	V	V	V	V	V
W	W	W	W	W	W
X	X	X	X	X	X
Y	Y	Y	Y	Y	Y
Z	Z	Z	Z	Z	Z
0	0	0	0	0	0
1	1	1	1	1	1
2	2	2	2	2	2
3	3	3	3	3	3
4	4	4	4	4	4
5	5	5	5	5	5
6	6	6	6	6	6
7	7	7	7	7	7
8	8	8	8	8	8
9	9	9	9	9	9

5. ☐ Spouse (husband or wife of active duty or Military Retiree)
☐ Retiree
☐ Son or daughter of Active Duty or Military Retiree
☐ DOD Employee
☐ Non-DOD Employee
☐ Other

5. For CIVILIANS ONLY: Complete Questions 5-6.
Mark ALL categories applicable to you.

6. ☐ WG ☐ GS ☐ SES ☐ GM
☐ 1 ☐ 6 ☐ 11 ☐ 16
☐ 2 ☐ 7 ☐ 12 ☐ 17
☐ 3 ☐ 8 ☐ 13 ☐ 18
☐ 4 ☐ 9 ☐ 14
☐ 5 ☐ 10 ☐ 15

6. If you are a Civilian Government Employee, enter your category and current pay grade.

7. LAST NAME

FI

A	A	A	A	A	A	A	A	A	A	A
B	B	B	B	B	B	B	B	B	B	B
C	C	C	C	C	C	C	C	C	C	C
D	D	D	D	D	D	D	D	D	D	D
E	E	E	E	E	E	E	E	E	E	E
F	F	F	F	F	F	F	F	F	F	F
G	G	G	G	G	G	G	G	G	G	G
H	H	H	H	H	H	H	H	H	H	H
I	I	I	I	I	I	I	I	I	I	I
J	J	J	J	J	J	J	J	J	J	J
K	K	K	K	K	K	K	K	K	K	K
L	L	L	L	L	L	L	L	L	L	L
M	M	M	M	M	M	M	M	M	M	M
N	N	N	N	N	N	N	N	N	N	N
O	O	O	O	O	O	O	O	O	O	O
P	P	P	P	P	P	P	P	P	P	P
Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q
R	R	R	R	R	R	R	R	R	R	R
S	S	S	S	S	S	S	S	S	S	S
T	T	T	T	T	T	T	T	T	T	T
U	U	U	U	U	U	U	U	U	U	U
V	V	V	V	V	V	V	V	V	V	V
W	W	W	W	W	W	W	W	W	W	W
X	X	X	X	X	X	X	X	X	X	X
Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z

FOR ALL INDIVIDUALS

7. Your Name.

Print the first ten letters of your last name and your first initial in these blank boxes.

Then fill in the corresponding response box below each letter.

8. ☐ AD or RM
☐ Spouse of AD or RM
☐ 1st ☐ 2nd ☐ 3rd ☐ 4th ☐ 5th Child
☐ Not Applicable

8. ARE YOU: (Mark ALL applicable categories)

Active Duty or Retired Military
 Spouse of Active Duty or Retired Military
 1st, 2nd, 3rd, 4th, or 5th child of Active Duty or Retired Military
 Not Applicable

9. YOUR SPONSOR'S SOCIAL SECURITY NUMBER OR YOUR SOCIAL SECURITY NUMBER

9. Print your SSN in the blank boxes. Then fill in the corresponding response box below each number.

- * If **ACTIVE DUTY** or **RETIRED** military, enter your SSN
- * If a **FAMILY MEMBER OF** active duty or retired, enter sponsors SSN
- * For **ALL OTHERS**, enter your SSN

0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9	9	9

10. This Health Risk Appraisal is being administered in the following situation:

10. ☐ In-Processing
☐ Periodic Physical Examination
☐ Pre-Physical Fitness Test
☐ Occupational Health Program
☐ Walk-In
☐ Other

11. Racial/Ethnic Background
Mark the most appropriate category.

11. ☐ American Indian or Alaska Native
☐ Asian/Oriental ☐ White, Hispanic
☐ Black, Hispanic ☐ White, Non-Hispanic
☐ Black Non-Hispanic ☐ Other
☐ Pacific Islander

12. Marital Status.
Mark the most appropriate category.

12. ☐ Married ☐ Separated
☐ Never Married ☐ Widowed
☐ Divorced ☐ Other

13. Are you MALE or FEMALE?

13. ☐ Male ☐ Female

14. Your Age

15. Your Height

16. Your Weight

BEFORE you fill in the response boxes
write age, height, and weight at the
top of the columns.

EXAMPLE:

HEIGHT = 6 feet-0 inches
(Must enter if 0 inches)

HEIGHT	
FEET	INCHES
6	0
4	
5	1
	2
7	3

AGE YEARS	
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
	8
	9

HEIGHT FEET INCHES	
4	0
5	1
6	2
7	3
	4
	5
	6
	7
	8
	9
	10
	11

WEIGHT POUNDS		
0	0	0
1	1	1
2	2	2
3	3	3
4	4	4
5	5	5
6	6	6
7	7	7
8	8	8
9	9	9

17. What is your Body Frame Size?

17. ☐ Small
☐ Medium
☐ Large

18. How often do you do exercises that improve muscle strength, such as pushups, situps, weight lifting, a Nautilus/Universal workout, resistance training, etc...?

18. ☐ 3 or more times a week
☐ 1 or 2 times a week
☐ Rarely or never

19. How often do you do at least 20 minutes of non-stop aerobic activity (vigorous exercise that greatly increases your breathing and heart rate such as running, fast walking, biking, swimming, rowing, etc...)?

19. ☐ 3 or more times a week
☐ 1 or 2 times a week
☐ Rarely or never

20. How often do you eat high fiber foods such as whole grain breads, cereals, bran, raw fruit, or raw vegetables?

20. ☐ At every meal
☐ Daily
☐ 3-5 days a week
☐ Less than 3 days a week
☐ Rarely or never

21. How often do you eat foods high in saturated fats such as beef, hamburger, pork, sausage, butter, whole milk, cheese, etc...?

21. ☐ At every meal
☐ Daily
☐ 3-5 days a week
☐ Less than 3 days a week
☐ Rarely or never

22. Do you usually salt your food before tasting?

22. ☐ Yes ☐ No

23	CAR/TRK/VAN	23	MOTORCYCLE
a.	,000	b.	,000
0	0	0	0
1	1	1	1
2	2	2	2
3	3	3	3
4	4	4	4
5	5	5	5
6	6	6	6
7	7	7	7
8	8	8	8
9	9	9	9

23.a. In the next 12 months how many thousands of miles will you travel by car, truck or van?

23.b. In the next 12 months how many thousands of miles will you travel by motorcycle?

NOTE: U.S. average for cars is 10,000 miles

- 24.** ☐ Walk ☐ Sub/Compact Car ☐ Truck/Van
☐ Bike ☐ Mid or Full Car ☐ Stay at Home
☐ Motorcycle ☐ Bus/Subway/Train

24. On a typical day how do you usually travel? (Mark only one)

25.

0	1
0	1 2 3 4 5 6 7 8 9
0	1 2 3 4 5 6 7 8 9

25. What percent of the time do you usually buckle your safety belt when driving or riding?

EXAMPLE: 50%

0	1
5	0 1 2 3 4 5 6 7 8 9
0	1 2 3 4 5 6 7 8 9

- 26.** ☐ Within 5 MPH of limit ☐ 11-15 MPH Over
☐ 6-10 MPH Over ☐ More than 15 MPH Over
☐ Don't Drive

26. On the average, how close to the speed limit do you usually drive?

27. NO. OF TIMES

28. NO. OF DRINKS

0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9

0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9

27. How many times in the last month did you drive or ride when the driver had perhaps too much alcohol to drink?

28. How many drinks of alcoholic beverages do you have in a typical week?

NOTE:

1 Drink = 1 glass of wine = 1 can of beer = 1 shot of liquor

EXAMPLE: 2 DRINKS

0	2
0	0
1	1
2	2

IF YOU DON'T DRINK SKIP TO QUESTION 35

29. ☐ Yes ☐ No

29. Have you ever felt you should cut down on your drinking?

30. ☐ Yes ☐ No

30. Have people ever annoyed you by criticizing your drinking?

31. ☐ Yes ☐ No

31. Have you ever felt bad or guilty about your drinking?

32. ☐ Yes ☐ No

32. Have you ever had a drink first thing in the morning to steady your nerves or get rid of a hangover (eye opener)?

33. ☐ Yes ☐ No

33. Do your friends ever worry about your drinking?

34. ☐ Yes ☐ No

34. Have you ever had a drinking problem?

35. ☐ Yes ☐ No

35. Have you ever been told that you have diabetes (or sugar diabetes)?

36. ☐ Yes ☐ No

36. Are you now taking medicine for high blood pressure?

- 37.** ☐ Daily or almost daily
☐ 3 to 5 days a week
☐ Less than 3 days a week
☐ Rarely or never

37. How often do you eat two well-balanced meals per day?

- 38.** ☐ Daily or almost daily
☐ 3 to 5 days a week
☐ Less than 3 days a week
☐ Rarely or never

38. How often do you eat foods high in salt or sodium such as cold cuts, bacon, canned soups, potato chips, etc...?

- 39.** ☐ Not Satisfied ☐ Somewhat Satisfied ☐ Mostly Satisfied ☐ Totally Satisfied ☐ Not Applicable

39. I am satisfied with my present job assignment and unit.

- 40.** ☐ Money ☐ Supervisor ☐ No Problem
☐ Social Life ☐ Job
☐ Family ☐ Health

40. What causes the biggest problem in your life?

41. In the last year, how many serious personal losses or difficult problems have you had to handle (example, promotion passover, divorce/separation, legal or disciplinary action, bankruptcy, death of someone close, serious illness/injury of a loved one, etc.)?	41. ☐ Several ☐ Some ☐ Few ☐ None																				
42. In general, how satisfied are you with your life (e.g., work situation, social activity, accomplishing what you set out to do)?	42. ☐ Not Satisfied ☐ Somewhat Satisfied ☐ Mostly Satisfied ☐ Totally Satisfied																				
43. How often are there people available that you can turn to for support in bad moments or illness?	43. ☐ Never ☐ Hardly Ever ☐ Sometimes ☐ Always																				
44. How many hours of sleep do you usually get at night?	44. ☐ 5 Hours or less ☐ 6-8 Hours ☐ 9 Hours or more																				
45. Have you seriously considered suicide within the last two years?	45. ☐ Yes ☐ Yes, within the last year ☐ Yes, within the last 2 months ☐ No																				
46. How often do you have any serious problems dealing with your husband or wife, parents, friends or with your children?	46. ☐ Often ☐ Sometimes ☐ Seldom ☐ Never																				
47. How often did you experience a major pleasant change in the past year? (for example, promotion, marriage, birth, award, etc.)?	47. ☐ Often ☐ Sometimes ☐ Seldom ☐ Never																				
48. How often has life been so overwhelming in the last year that you seriously considered hurting yourself?	48. ☐ Often ☐ Sometimes ☐ Seldom ☐ Never																				
49. In the past year, how often have you experienced repeated or long periods of depression?	49. ☐ Often ☐ Sometimes ☐ Seldom ☐ Never																				
50. In the past year, how often have your worries interfered with your daily life?	50. ☐ Often ☐ Sometimes ☐ Seldom ☐ Never																				
51. How often are you able to find times to relax?	51. ☐ Often ☐ Sometimes ☐ Seldom ☐ Never																				
52. How often do you feel that your present work situation is putting you under too much stress?	52. ☐ Often ☐ Sometimes ☐ Seldom ☐ Never																				
TOBACCO USE HISTORY																					
53. How many cigars do you usually smoke per day?	53. ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10																				
54. How many pipes of tobacco do you usually smoke per day?	54. ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10																				
55. How many times per day do you usually use smokeless tobacco? (Chewing tobacco, snuff, pouches, etc.)	55. ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9																				
EXAMPLE: 20 times 0 1 2 3 4 5 6 7 8 9																					
56. CIGARETTE SMOKING How would you describe your cigarette smoking habits?	56. ☐ Never Smoked (SKIP TO QUESTION 58) ☐ Current Smoker ☐ Ex-Smoker																				
57. STILL SMOKE a. How many cigarettes a day do you smoke?	57. USED TO SMOKE b. How many years has it been since you smoked cigarettes fairly regularly?																				
c. What was the average number of cigarettes you smoked per day during the two years before you quit?	57. a. NUMBER b. YEARS c. AVERAGE <table border="1" style="width: 100%; border-collapse: collapse; text-align: center;"> <tr> <td style="width: 10%;">0</td><td style="width: 10%;">1</td><td style="width: 10%;">2</td><td style="width: 10%;">3</td><td style="width: 10%;">4</td><td style="width: 10%;">5</td><td style="width: 10%;">6</td><td style="width: 10%;">7</td><td style="width: 10%;">8</td><td style="width: 10%;">9</td> </tr> <tr> <td>0</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td> </tr> </table>	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9												
0	1	2	3	4	5	6	7	8	9												
58. About how long has it been since you had a rectal exam?	58. ☐ Less than 1 year ☐ 1 year ☐ 2 years ☐ 3 or more years ☐ Never																				
59. When was the last time you visited the dental clinic for a check-up?	59. ☐ Within the last year ☐ Between one and two years ago ☐ Over two years ago																				

WOMEN ONLY

MEN ONLY			MEN ONLY		
68.	☐ Less than 1 year ☐ 1 year	☐ 2 years ☐ 3 or more years	68.	About how long has it been since you had a prostate (rectal) exam?	
69.	☐ Monthly ☐ Rarely/Never	☐ Every few months	69.	How often do you do a testicular (sex organs) self exam?	

70.	TOTAL CHOL			71.	HDL CHOL			72.	12 HR. FAST			70. Blood Lipids Total Cholesterol (mg/dl)	71. Blood Lipids HDL Cholesterol (mg/dl)	72. Blood Glucose 12 Hr. Fasting (mg %)
	0	0	0		0	0	0		0	0	0			
	1	1	1		1	1	1		1	1	1			
	2	2	2		2	2	2		2	2	2			
	3	3	3		3	3	3		3	3	3			
	4	4	4		4	4	4		4	4	4			
	5	5	5		5	5	5		5	5	5			
	6	6	6		6	6	6		6	6	6			
	7	7	7		7	7	7		7	7	7			
	8	8	8		8	8	8		8	8	8			
	9	9	9		9	9	9		9	9	9			

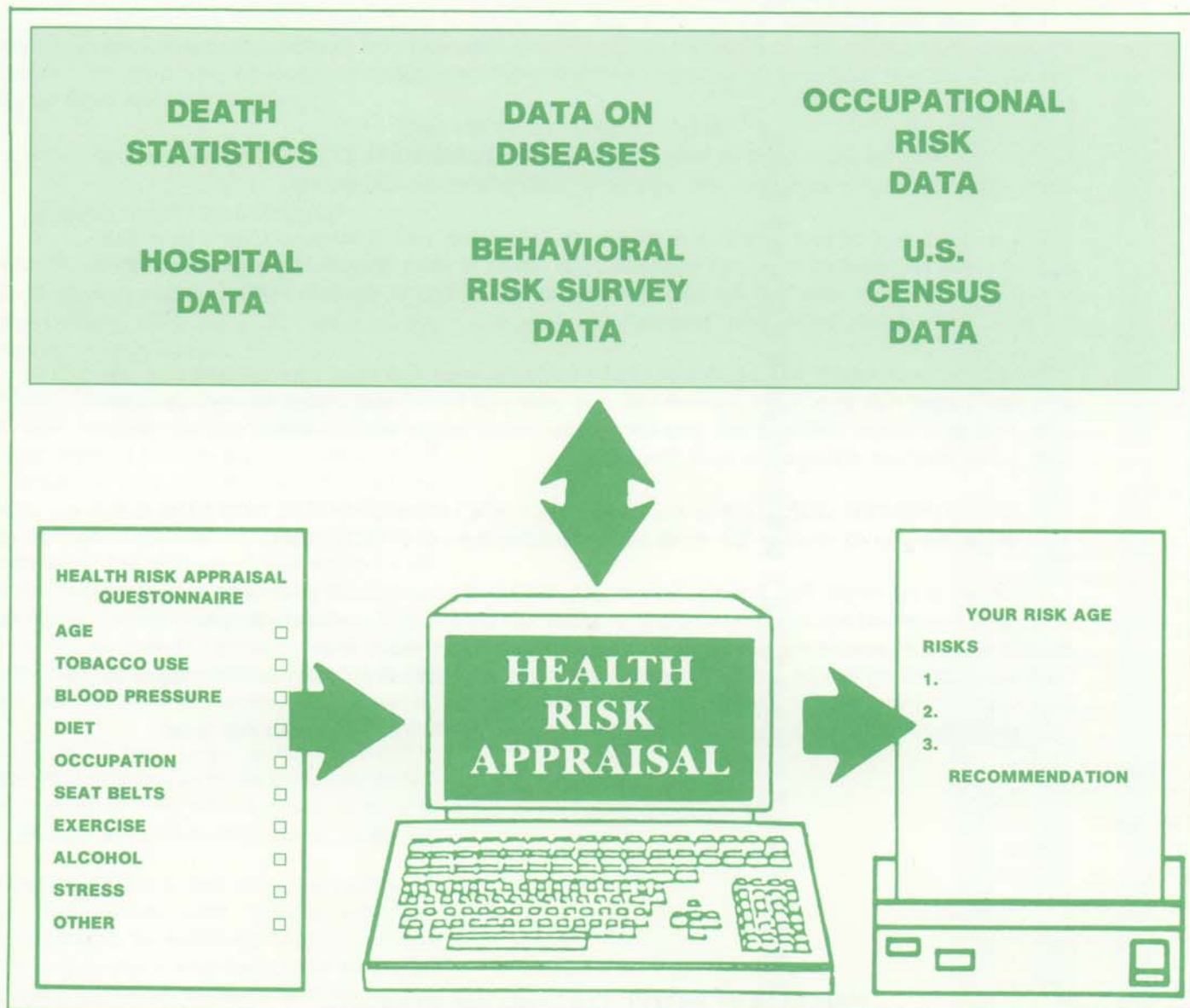
73.	B.P.-SYSTOLIC			74.	B.P.-DIASTOLIC			73. Blood Pressure (Systolic)	74. Blood Pressure (Diastolic)
	0	0	0		0	0	0		
	1	1	1		1	1	1		
	2	2	2		2	2	2		
	3	3	3		3	3	3		
	4	4	4		4	4	4		
	5	5	5		5	5	5		
	6	6	6		6	6	6		
	7	7	7		7	7	7		
	8	8	8		8	8	8		
	9	9	9		9	9	9		

THE HEALTH PROMOTION PROGRAM

Fit to Win



HEALTH RISK APPRAISAL



For use of this form, see AR40-501 and AR600-63; the proponent agency is TSG

FIT TO WIN



TM

The **HEALTH RISK APPRAISAL** is an activity of
THE HEALTH PROMOTION PROGRAM

How does the Health Risk Appraisal work?

The health risk appraisal is a personalized estimation of your risks of death and major illness in the next ten years. First, the program uses your age and health-related personal habits, as well as national statistics on risk factors and diseases, to calculate your current risks.

Your risk may be expressed in terms of **RISK AGE** or **HEALTH SCORE**. Ideally, you want a risk age lower than your real age or a health score of 100 points.

The second part of your health risk appraisal calculates your risks again, as if your risk factors were reduced as much as possible. The result is your "target" risk age or health score. It shows your potential benefit, in health terms, of improving your lifestyle-if you quit smoking, wear safety belts, take moderate exercise, etc.

Therefore, your health risk appraisal report includes your real age, your current risk age and your target risk age. Your current risk age tells you how healthy your lifestyle is right now, and your target risk age lets you know how much longer and healthier you can live with a few positive changes in your lifestyle.

PLEASE ANSWER QUESTIONS AS HONESTLY AND AS CORRECTLY AS YOU CAN. This will allow you to receive the most accurate assessment of your health.

The results of the Health Risk Appraisal are for you. We ask that you give us your name so we can return your results and any recommendations for follow-up care to you. We also ask for your social security number so we can statistically track trends in health awareness over long periods of time. Statistical information may be collected from an wide database which will contain your information, but your name and social security number will be covered and cannot be read. The rules of the Privacy Act apply to any information that you give in the Health Risk Appraisal.

IMPORTANT NOTE! The health risk appraisal is no substitute for a physical examination or check-up. It will not give you a diagnosis nor will it tell you how long you will actually live. However, the health risk appraisal will help you understand and recognize your risk factors.

INSTRUCTIONS

Please use a No. 2 Pencil only to complete this survey. Make dark, black marks that fill the response boxes completely.

EXAMPLE: Correct

Incorrect



Health Risk Appraisal (HRA)
for use of this form, see
AR40-501 and AR600-63;
the proponent is TSG

For MILITARY SERVICE MEMBERS ONLY: Complete Questions 1-4. (All others go to Question 5.)

1. What is your branch of service?

1. ☐ U.S. Army ☐ U.S. Marines
☐ U.S. Navy ☐ U.S. Coast Guard
☐ U.S. Air Force ☐ Other

2. What is your military status?

2. ☐ Active ☐ Reserve
☐ Active Reserve ☐ Guard
☐ Active Guard ☐ Other

3. What is your current rank?

3.

ENLISTED

- ☐ E-1 ☐ E-6
☐ E-2 ☐ E-7
☐ E-3 ☐ E-8
☐ E-4 ☐ E-9
☐ E-5

OFFICER

- ☐ O-1 ☐ O-6
☐ O-2 ☐ O-7
☐ O-3 ☐ O-8
☐ O-4 ☐ O-9
☐ O-5 ☐ O-10

WARR. OFFIC.

- ☐ WO-1
☐ WO-2
☐ WO-3
☐ WO-4

4. What is your Unit Identification Code?
(Enter Specific Unit Identifier)

Print your Unit Identification Code in these blank boxes.

Then fill in the corresponding response box below each number/letter.

4.

UNIT CODE

A	A	A	A	A	A
B	B	B	B	B	B
C	C	C	C	C	C
D	D	D	D	D	D
E	E	E	E	E	E
F	F	F	F	F	F
G	G	G	G	G	G
H	H	H	H	H	H
I	I	I	I	I	I
J	J	J	J	J	J
K	K	K	K	K	K
L	L	L	L	L	L
M	M	M	M	M	M
N	N	N	N	N	N
O	O	O	O	O	O
P	P	P	P	P	P
Q	Q	Q	Q	Q	Q
R	R	R	R	R	R
S	S	S	S	S	S
T	T	T	T	T	T
U	U	U	U	U	U
V	V	V	V	V	V
W	W	W	W	W	W
X	X	X	X	X	X
Y	Y	Y	Y	Y	Y
Z	Z	Z	Z	Z	Z
0	0	0	0	0	0
1	1	1	1	1	1
2	2	2	2	2	2
3	3	3	3	3	3
4	4	4	4	4	4
5	5	5	5	5	5
6	6	6	6	6	6
7	7	7	7	7	7
8	8	8	8	8	8
9	9	9	9	9	9

PRIVACY ACT STATEMENT

AUTHORITY: 29 CFR Chapter XVII, Occupational Safety and Health Standards; 5 U.S.C., section 150; Executive Orders 11612 and 11807 authorize the collection of this information.

PURPOSE: The primary use of this information is by the unit medical care providers to assure competent medical care. Additional disclosures of this information may be: To the Office of the Surgeons General in aggregated form to develop Command fitness profiles; to military medical researchers for the purpose of correlating health precursors to health problems or to commercial medical researchers for the same purpose. Where data from this system of records are provided to agencies external to the military, Social Security Number and Name will be deleted.

ROUTINE USES: Information may be disclosed to departments and agencies of the Executive Branch in performance of their official duties relating to health risk appraisal and cardiovascular screening.

DISCLOSURE: We ask that you give your name so we can return your results and any recommendations for follow-up care to you. We also ask for your social security number so we can statistically track trends in health awareness over long periods of time.

5. ☐ Spouse (husband or wife of active duty or Military Retiree)
☐ Retiree
☐ Son or daughter of Active Duty or Military Retiree
☐ DOD Employee
☐ Non-DOD Employee
☐ Other

5. For CIVILIANS, MILITARY RETIREES, AND FAMILY MEMBERS ONLY: Complete questions 5-6
Mark ALL categories applicable to you.

6. ☐ WG ☐ GS ☐ SES ☐ GM
☐ 1 ☐ 6 ☐ 11 ☐ 16
☐ 2 ☐ 7 ☐ 12 ☐ 17
☐ 3 ☐ 8 ☐ 13 ☐ 18
☐ 4 ☐ 9 ☐ 14
☐ 5 ☐ 10 ☐ 15

6. If you are a Civilian Government Employee, enter your category and current pay grade.

7. LAST NAME

FI

A	A	A	A	A	A	A	A	A	A	A
B	B	B	B	B	B	B	B	B	B	B
C	C	C	C	C	C	C	C	C	C	C
D	D	D	D	D	D	D	D	D	D	D
E	E	E	E	E	E	E	E	E	E	E
F	F	F	F	F	F	F	F	F	F	F
G	G	G	G	G	G	G	G	G	G	G
H	H	H	H	H	H	H	H	H	H	H
I	I	I	I	I	I	I	I	I	I	I
J	J	J	J	J	J	J	J	J	J	J
K	K	K	K	K	K	K	K	K	K	K
L	L	L	L	L	L	L	L	L	L	L
M	M	M	M	M	M	M	M	M	M	M
N	N	N	N	N	N	N	N	N	N	N
O	O	O	O	O	O	O	O	O	O	O
P	P	P	P	P	P	P	P	P	P	P
Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q
R	R	R	R	R	R	R	R	R	R	R
S	S	S	S	S	S	S	S	S	S	S
T	T	T	T	T	T	T	T	T	T	T
U	U	U	U	U	U	U	U	U	U	U
V	V	V	V	V	V	V	V	V	V	V
W	W	W	W	W	W	W	W	W	W	W
X	X	X	X	X	X	X	X	X	X	X
Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z

FOR ALL INDIVIDUALS

7. Your Name.

Print the first ten letters of your last name and your first initial in these blank boxes.

Then fill in the corresponding response box below each letter.

8. ☐ AD or RM
☐ Spouse of AD or RM
☐ 1st ☐ 2nd ☐ 3rd ☐ 4th ☐ 5th Child
☐ Not Applicable

8. ARE YOU: (Mark ALL applicable categories)

Active Duty or Retired Military
 Spouse of Active Duty or Retired Military
 1st, 2nd, 3rd, 4th, or 5th child of Active Duty or Retired Military
 Not Applicable

9. YOUR SPONSOR'S SOCIAL SECURITY NUMBER OR YOUR SOCIAL SECURITY NUMBER

9. Print your SSN in the blank boxes. Then fill in the corresponding response box below each number.

- * If **ACTIVE DUTY** or **RETIRED** military, enter your SSN
- * If a **FAMILY MEMBER** OF active duty or retired, enter sponsors SSN
- * For **ALL OTHERS**, enter your SSN

0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9	9	9

10. This Health Risk Appraisal is being administered in the following situation:

- ☐ In-Processing
☐ Periodic Physical Examination
☐ Pre-Physical Fitness Test
☐ Occupational Health Program
☐ Walk-In
☐ Other

11. Racial/Ethnic Background
Mark the most appropriate category.

- ☐ American Indian or Alaska Native
☐ Asian/Oriental ☐ White, Hispanic
☐ Black, Hispanic ☐ White, Non-Hispanic
☐ Black Non-Hispanic ☐ Other
☐ Pacific Islander

12. Marital Status.
Mark the most appropriate category.

- ☐ Married ☐ Separated
☐ Never Married ☐ Widowed
☐ Divorced ☐ Other

13. Are you MALE or FEMALE?

13. ☐ MALE ☐ FEMALE

14. Your Age

15. Your Height

16. Your Weight

BEFORE you fill in the response boxes
write age, height, and weight at the
top of the columns.

EXAMPLE:

HEIGHT = 6 feet-0 inches
(Must enter if 0 inches)

HEIGHT	
FEET	INCHES
6	0
4	
5	1
	2
7	3

14. AGE	15. HEIGHT	16. WEIGHT
YEARS	FEET INCHES	POUNDS
0 0	4 0	0 0 0
1 1	5 1	1 1 1
2 2	6 2	2 2 2
3 3	7 3	3 3 3
4 4	4	4 4 4
5 5	5	5 5 5
6 6	6	6 6 6
7 7	7	7 7 7
	8	8 8 8
	9	9 9 9
	10	
	11	

17. What is your Body Frame Size?

17.
☐ Small
☐ Medium
☐ Large

18. How often do you do exercises that improve muscle strength, such as pushups, situps, weight lifting, a Nautilus/Universal workout, resistance training, etc...?

18.
☐ 3 or more times a week
☐ 1 or 2 times a week
☐ Rarely or never

19. How often do you do at least 20 minutes of non-stop aerobic activity (vigorous exercise that greatly increases your breathing and heart rate such as running, fast walking, biking, swimming, rowing, etc...)?

19.
☐ 3 or more times a week
☐ 1 or 2 times a week
☐ Rarely or never

20. How often do you eat high fiber foods such as whole grain breads, cereals, bran, raw fruit, or raw vegetables?

20. ☐ At every meal
☐ Daily
☐ 3-5 days a week
☐ Less than 3 days a week
☐ Rarely or never

21. How often do you eat foods high in saturated fats such as beef, hamburger, pork, sausage, butter, whole milk, cheese, etc...?

21. ☐ At every meal
☐ Daily
☐ 3-5 days a week
☐ Less than 3 days a week
☐ Rarely or never

22. Do you usually salt your food before tasting?

22. ☐ YES ☐ NO

23	CAR/TRK/VAN	23	MOTORCYCLE
a.	,000	b.	,000
	0		0
	1		1
	2		2
	3		3
	4		4
	5		5
	6		6
	7		7
	8		8
	9		9

23.a. In the next 12 months how many thousands of miles will you travel by car, truck or van?

23.b. In the next 12 months how many thousands of miles will you travel by motorcycle?

NOTE: U.S. average for cars is 10,000 miles

24. ☐ Walk ☐ Sub/Compact Car ☐ Truck/Van
☐ Bike ☐ Mid or Full Car ☐ Stay at Home
☐ Motorcycle ☐ Bus/Subway/Train

24. On a typical day how do you usually travel?
(Mark only one)

25.

0	1
0	1
0	1

25. What percent of the time do you usually buckle your safety belt when driving or riding?

EXAMPLE: 50%

0	1
5	0
0	1

26. ☐ Within 5 MPH of limit ☐ 11-15 MPH Over
☐ 6-10 MPH Over ☐ More than 15 MPH Over
☐ Don't Drive

26. On the average, how close to the speed limit do you usually drive?

27. NO. OF TIMES

28. NO. OF DRINKS

0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9

0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9

27. How many times in the last month did you drive or ride when the driver had perhaps too much alcohol to drink?

28. How many drinks of alcoholic beverages do you have in a typical week?

NOTE:

1 Drink = 1 glass of wine or wine cooler = 1 can of beer = 1 shot of liquor = 1 mixed drink.

EXAMPLE: 2 DRINKS

0	2
1	1
2	0

29. ☐ Yes ☐ No

29. Have you ever felt you should cut down on your drinking?

30. ☐ Yes ☐ No

30. Have people ever annoyed you by criticizing your drinking?

31. ☐ Yes ☐ No

31. Have you ever felt bad or guilty about your drinking?

32. ☐ Yes ☐ No

32. Have you ever had a drink first thing in the morning to steady your nerves or get rid of a hangover (eye opener)?

33. ☐ Yes ☐ No

33. Do your friends ever worry about your drinking?

34. ☐ Yes ☐ No

34. Have you ever had a drinking problem?

35. ☐ Yes ☐ No

35. Have you ever been told that you have diabetes (or sugar diabetes)?

36. ☐ Yes ☐ No

36. Are you now taking medicine for high blood pressure?

37. ☐ Daily or almost daily
☐ 3 to 5 days a week
☐ Less than 3 days a week
☐ Rarely or never

37. How often do you eat two well-balanced meals per day?

38. ☐ Daily or almost daily
☐ 3 to 5 days a week
☐ Less than 3 days a week
☐ Rarely or never

38. How often do you eat foods high in salt or sodium such as cold cuts, bacon, canned soups, potato chips, etc...?

39. ☐ Not Satisfied ☐ Somewhat Satisfied ☐ Mostly Satisfied ☐ Totally Satisfied ☐ Not Applicable

39. I am satisfied with my present job assignment and unit.

40. ☐ Money ☐ Supervisor ☐ No Problem
☐ Social Life ☐ Job
☐ Family ☐ Health

40. What causes the biggest problem in your life?

41. In the last year, how many serious personal losses or difficult problems have you had to handle (example, promotion passover, divorce/separation, legal or disciplinary action, bankruptcy, death of someone close, serious illness/injury of a loved one, etc.)?	41. ☐ Several ☐ Few ☐ Some ☐ None
42. In general, how satisfied are you with your life (e.g., work situation, social activity, accomplishing what you set out to do)?	42. ☐ Not Satisfied ☐ Somewhat Satisfied ☐ Mostly Satisfied ☐ Totally Satisfied
43. How often are there people available that you can turn to for support in bad moments or illness?	43. ☐ Never ☐ Hardly Ever ☐ Sometimes ☐ Always
44. How many hours of sleep do you usually get at night?	44. ☐ 5 Hours or less ☐ 6-8 Hours ☐ 9 Hours or more
45. Have you seriously considered suicide within the last two years?	45. ☐ Yes ☐ Yes, within the last year ☐ Yes, within the last 2 months ☐ No
46. How often do you have any serious problems dealing with your husband or wife, parents, friends or with your children?	46. ☐ Often ☐ Sometimes ☐ Seldom ☐ Never
47. How often did you experience a major pleasant change in the past year? (for example, promotion, marriage, birth, award, etc.)?	47. ☐ Often ☐ Sometimes ☐ Seldom ☐ Never
48. How often has life been so overwhelming in the last year that you seriously considered hurting yourself?	48. ☐ Often ☐ Sometimes ☐ Seldom ☐ Never
49. In the past year, how often have you experienced repeated or long periods of depression?	49. ☐ Often ☐ Sometimes ☐ Seldom ☐ Never
50. In the past year, how often have your worries interfered with your daily life?	50. ☐ Often ☐ Sometimes ☐ Seldom ☐ Never
51. How often are you able to find times to relax?	51. ☐ Often ☐ Sometimes ☐ Seldom ☐ Never
52. How often do you feel that your present work situation is putting you under too much stress?	52. ☐ Often ☐ Sometimes ☐ Seldom ☐ Never
TOBACCO USE HISTORY	
53. How many cigars do you usually smoke per day?	53. ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
54. How many pipes of tobacco do you usually smoke per day?	54. ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
55. How many times per day do you usually use smokeless tobacco? (Chewing tobacco, snuff, pouches, etc.)	55. ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9
56. CIGARETTE SMOKING How would you describe your cigarette smoking habits?	56. ☐ Never Smoked (SKIP TO QUESTION 58) ☐ Current Smoker ☐ Ex-Smoker
57. STILL SMOKE a. How many cigarettes a day do you smoke?	57. a. ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9
USED TO SMOKE b. How many years has it been since you smoked cigarettes fairly regularly?	57. b. ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9
c. What was the average number of cigarettes you smoked per day during the two years before you quit?	57. c. ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9
58. About how long has it been since you had a rectal exam?	58. ☐ Less than 1 year ☐ 1 year ☐ 2 years ☐ 3 or more years ☐ Never
59. When was the last time you visited the dental clinic for a check-up?	59. ☐ Within the last year ☐ Between one and two years ago ☐ Over two years ago

WOMEN ONLY

60. ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
☐ 11 ☐ 12 ☐ 13 ☐ 14 ☐ 15 ☐ 16 ☐ 17 ☐ 18 ☐ 19 ☐ 20
☐ No Children ☐ 10

61. ☐ 11 ☐ 12 ☐ 13 ☐ 14 ☐ 15 ☐ 16 ☐ 17 ☐ 18 ☐ 19 ☐ 20
☐ 21 ☐ 22 ☐ 23 ☐ 24 ☐ 25 ☐ 26 ☐ 27 ☐ 28 ☐ 29 ☐ 30
☐ 31 ☐ 32 ☐ 33 ☐ 34 ☐ 35 ☐ 36 ☐ 37 ☐ 38 ☐ 39 ☐ 40
☐ 41 ☐ 42 ☐ 43 ☐ 44 ☐ 45 ☐ 46 ☐ 47 ☐ 48 ☐ 49 ☐ 50

62. ☐ Less than 1 year ☐ 3 or more years
☐ 1 year ☐ 2 years ☐ Never

63. ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10

64. ☐ Yes ☐ No ☐ Don't know

65. ☐ Less than 1 year ☐ 2 years ☐ Never
☐ 1 year ☐ 3 or more years

66. ☐ Monthly ☐ Rarely/Never ☐ Every few months

67. ☐ Less than 1 year ☐ 2 years ☐ Never
☐ 1 year ☐ 3 or more years

WOMEN ONLY

60. At what age did you have your first menstrual period?

61. How old were you when your first child was born?

62. How long has it been since your last breast X-ray (Mammogram)?

63. How many women in your natural family (mother and sisters only) have had breast cancer?

64. Have you had a hysterectomy operation? (removal of the uterus)

65. How long has it been since you had a pap smear for cancer?

66. How often do you examine your breasts for lumps?

67. About how long has it been since you had your breasts examined by a physician or nurse?

MEN ONLY

68. ☐ Less than 1 year ☐ 2 years ☐ Never
☐ 1 year ☐ 3 or more years

69. ☐ Monthly ☐ Rarely/Never ☐ Every few months

MEN ONLY

68. About how long has it been since you had a prostate (rectal) exam?

69. How often do you do a testicular (sex organs) self exam?

Questions 70 - 75 should be completed by MEDICAL PERSONNEL ONLY.

70. TOTAL CHOL ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9

71. HDL CHOL ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9

72. 12 HR. FAST ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9

70. Blood Lipids
Total Cholesterol
(mg/dl)

71. Blood Lipids
HDL Cholesterol
(mg/dl)

72. Blood Glucose
12 Hr. Fasting
(mg %)

73. B.P.-SYSTOLIC ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9

74. B.P.-DIASTOLIC ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9

73. Blood Pressure
(Systolic)

74. Blood Pressure
(Diastolic)

75. ☐ NL ☐ ABN w/o LVH
☐ ABN w/LVH ☐ UNKNOWN

75. Most recent electrocardiogram results.

X1. ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10

X2. ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10

X3. ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10

X4. ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10

X5. ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10

X6. ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10

X7. ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10

X8. ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10