ELSEVIER

Reliability of Standard Health Assessment Instruments in a Large, Population-Based Cohort Study

TYLER C. SMITH, MS, BESA SMITH, MPH, ISABEL G. JACOBSON, MPH, THOMAS E. CORBEIL, MCS, AND MARGARET A.K. RYAN, MD, MPH, FOR THE MILLENNIUM COHORT STUDY TEAM*

PURPOSE: The Millennium Cohort Study began in 2001 using mail and Internet questionnaires to gather occupational and environmental exposure, behavioral risk factor, and health outcome data from a large, population-based US military cohort. Standardized instruments, including the Patient Health Questionnaire, the Medical Outcomes Study Short Form-36 for Veterans, and the Posttraumatic Stress Disorder (PTSD) Checklist–Civilian Version, have been validated in various populations. The purpose of this study was to investigate internal consistency of standardized instruments and concordance of responses in a test-retest setting.

METHODS: Cronbach alpha coefficients were used to investigate the internal consistency of standardized instruments among 76,742 participants. Kappa statistics were calculated to measure stability of aggregated responses in a subgroup of 470 participants who voluntarily submitted an additional survey within 6 months of their original submission.

RESULTS: High internal consistency was found for 14 of 16 health components, with lower internal consistency found among two alcohol components. Substantial test-retest stability was observed for stationary variables, while moderate stability was found for more dynamic variables that measured conditions with low prevalence.

CONCLUSIONS: These results substantiate internal consistency and stability of several standard health instruments applied to this large cohort. Such reliability analyses are vital to the integrity of long-term outcome studies.

Ann Epidemiol 2007;17:525–532.

KEY WORDS: Health Surveys, Military Medicine, Reliability, Questionnaires.

INTRODUCTION

Standardized instruments are often used in survey research. Many of these instruments are devised in clinic settings where health assessment is completed by trained health care professionals. However, prohibitive cost and relative ease make participant-assessed outcome measures a more feasible approach to obtain constructs describing functional and mental health outcomes. With these more convenient measures of health increasingly used as primary outcomes in epidemiologic studies, selecting an appropriate assessment tool involves careful review of the many standard survey instruments available. Special consideration of whether the instruments meet the requirements of the proposed application is critical to interpretation of collected data (1). Reliability and validity of these instruments are often tested thoroughly in populations or settings in which the instrument was originally created (2, 3). However, many questionnaires incorporate standardized survey instruments in populations that may be different from those for which the instrument was intended. In these studies, it is important to establish a level of confidence in the information being ascertained prior to declaring the instrument appropriate for the targeted population.

From the Department of Defense Center for Deployment Health Research at the Naval Health Research Center, San Diego, CA.

Address correspondence to: Tyler C. Smith, DoD Center for Deployment Health Research, Naval Health Research Center, P.O. Box 85122, San Diego, CA 92186-5122. Tel.: (619) 553-7593; fax (619) 553-7601. E-mail: Smith@nhrc.navy.mil.

*In addition to the authors, the Millennium Cohort Study Team includes Paul J. Amoroso, MD, MPH (Madigan Army Medical Center, Tacoma, WA); Edward J. Boyko, MD, MPH (Seattle Epidemiologic Research and Information Center, Department of Veterans Affairs Puget Sound Health Care System, Seattle, WA; Gary D. Gackstetter, PhD, DVM, MPH (Department of Preventive Medicine and Biometrics, Uniformed Services University of the Health Sciences, Bethesda, MD and Analytic Services, Inc. [ANSER], Arlington, VA; Gregory C. Gray, MD, MPH (College of Public Health, University of Iowa, Iowa City, IA; Tomoko I. Hooper, MD, MPH, Department of Preventive Medicine and Biometrics, Uniformed Services University of the Health Sciences, Bethesda, MD); James R. Riddle, DVM, MPH, and Timothy S. Wells, PhD, DVM, MPH. (both from Air Force Research Laboratory, Wright Patterson AFB, OH.).

Disclosure: This work represents Report 06-24, supported by the Department of Defense, under work unit No. 60002. The views expressed in this article are those of the authors and do not reflect the official policy or position of the Department of the Navy, Department of the Army, Department of the Air Force, Department of Defense, Department of Veterans Affairs, or the US Government. This research has been conducted in compliance with all applicable federal regulations governing the protection of human subjects in research (Protocol NHRC.2000.007).

Received October 31, 2006; accepted December 19, 2006.

360 Park Avenue South, New York, NY 10010

1047-2797/07/$–see front matter
doi:10.1016/j.annepidem.2006.12.002

AEP Vol. 17, No. 7
July 2007: 525–532

Selected Abbreviations and Acronyms

PCL-C = PTSD Checklist–Civilian Version
PHQ = Primary Care Evaluation of Mental Disorders (PRIME-MD) Patient Health Questionnaire
PTSD = posttraumatic stress disorder
SF-36V = Medical Outcomes Study Short Form 36-item questionnaire for Veterans

The Millennium Cohort, the largest cohort study ever undertaken by the US Department of Defense, was launched in 2001 to gather health outcome information along with occupational and environmental exposures employing a longitudinal approach (4, 5). In the first panel of enrollment, more than 77,000 participants joined the 22-year-long study, filling out either a mailed survey or an identical Web-based survey. The Millennium Cohort Study questionnaire is composed of more than 60 multipart questions comprising more than 400 individual data points, including questions from standardized instruments such as the Medical Outcomes Study Short Form 36-item for Veterans (SF-36V) (6, 7), the Primary Care Evaluation of Mental Disorders (PRIME-MD) Patient Health Questionnaire (PHQ) (2, 8, 9), the Posttraumatic Stress Disorder (PTSD) Checklist–Civilian Version (PCL-C) (3, 10), and the CAGE questionnaire to assess problematic drinking behavior (11), as well as questions that target areas such as medical history, vaccinations, environmental exposures, and occupation. Although the concordance of test-retest responses and internal consistency of the standard instruments have been established (6–10), tests of reliability of these constructs have not been performed in a large, population-based cohort where multiple independent instruments are presented simultaneously. The purpose of this study, therefore, was to establish the reliability as measured by concordance in a test-retest setting and internal consistency of several standardized instruments in a large, population-based military cohort.

MATERIALS AND METHODS

Study Population

The invited Millennium Cohort Study participants were randomly selected from all US military personnel serving in the Army, Navy, Coast Guard, Air Force, and Marine Corps as of October 1, 2000. The population-based sample represented approximately 11% of the 2.3 million men and women in service and, oversampled for those who had been previously deployed, were US Reserve and National Guard personnel, and female service members, to ensure sufficient power to detect differences in smaller subgroups of the population.

This study included 77,047 consenting participants in the first panel of the Millennium Cohort Study. Of these participants, 925 subjects voluntarily submitted either two paper surveys or one Web survey and one paper survey during the 2-year period of enrollment without additional soliciting of participation in a test-retest investigation. Participants were not permitted to log on to the Web survey after completion, prohibiting duplicate Internet submission. All participants gave informed, voluntary consent. Of the 925 participants, 21 had incomplete submission date information for at least one of the surveys they submitted. Of the remaining 904 subjects who voluntarily submitted a second survey, 487 submitted their second survey within 6 months (180 days) of the first submission. This cutoff was used based on the practicality of an a priori design of a proposed test-retest effort of this cohort. The projected retest time of up to 6 months from the first survey completion to their second survey completion was selected because of the known dynamic nature of many of the Millennium Cohort questions. Of this subgroup, 470 of the 487 had complete covariate data and were used in this investigation of concordance.

In addition to using self-reported data obtained from each survey, demographic and military personnel data were linked to each participant and reflected military status as of October 1, 2000. These data included sex, birth year (categorized by groups: pre-1960, 1960–1969, 1970–1979, and 1980 forward), level of education, marital status, pay grade (enlisted or officer), race/ethnicity (white non-Hispanic, black non-Hispanic, and other), service component (active duty or reserve), service branch (Army, Navy/Coast Guard, Air Force, and Marines), and occupation (defined over 10 broad categories). When available, self-reported data were used to supplement missing data from personnel records for marital status, education, and occupation. This reduced those missing data for at least one important demographic characteristic from 1.8% to 0.4% of the general cohort.

Questionnaire Data

The Millennium Cohort Study questionnaire collects information regarding diagnosed medical conditions, reported symptoms, psychosocial assessment, physical status, functional status, alcohol use, tobacco use, occupational status, alternative medicine use, exposures, sleep patterns, deployments, and basic demographic and contact data (5). Standardized instruments include the PHQ (2, 8, 9), used to assess major depressive syndrome, panic syndrome, other anxiety syndrome, bulimia nervosa, alcohol abuse, and binge-eating disorders (overall accuracy = 0.85; sensitivity = 0.75; specificity = 0.90), as well as specific conditions such as major depressive disorder found to be highly sensitive and specific (sensitivity = 0.93; specificity = 0.89) (12) and

panic disorder found to be highly sensitive but moderately specific (sensitivity = 1.00; specificity = 0.63) (13). The SF-36V was also included (6, 7) to define 8 components to assess physical functioning, role limitations caused by physical problems, bodily pain, general health, vitality, social functioning, role limitations caused by emotional problems, and mental health and has been found to have high internal consistency across all 8 domains in a military population (14). These 8 health concepts constitute the component summary measures Physical Component Summary (PCS) and Mental Component Summary (MCS) (15–17) Components of the National Health Survey of Gulf War Era Veterans and Their Families were included to assess wartime exposures (18, 19). The CAGE questionnaire was included for the detection of alcohol problems (11). The PCL-C was included for detection of PTSD symptoms (10, 20, 21). In a population with similar prevalence of PTSD, the PCL-C was shown to be highly specific (specificity = 0.99) with slightly lower sensitivity (0.60), a positive predictive value of 0.75, and a negative predictive value of 0.97 when using the standard criteria of reporting a moderate or above level of at least 1 intrusion symptom, 3 avoidance symptoms, and 2 hyperarousal symptoms (22), with a total score of 50 or more on a scale of 17 to 85 (3, 10, 23, 24).

TABLE 1. Kappa scores for the Millennium Cohort Study questionnaire among 470 participants submitting two questionnaires within 6 months

Question category	Completion rate* No. (%)	Mean kappa†	Interpretation‡
Demographics	460 (97.8)	0.87	Almost perfect
Smoking	330 (70.2)	0.82	Almost perfect
Exposures	458 (97.4)	0.67	Substantial
Women's health	133 (28.3)	0.66	Substantial
PHQ	299 (63.6)	0.55	Moderate
Alcohol (other than PHQ)	463 (98.5)	0.54	Moderate
CAM	462 (98.3)	0.54	Moderate
PCL-C	454 (96.6)	0.54	Moderate
SF-36V	463 (98.5)	0.49	Moderate
Symptoms and conditions	455 (96.8)	0.48	Moderate

CAM = complementary and alternative medicine; PCL-C = Posttraumatic Stress Disorder Patient Checklist, Civilian Version; PHQ = Primary Care Evaluation of Mental Disorders (PRIME-MD) Patient Health Questionnaire; SF-36V = Medical Outcomes Study Short Form 36-item questionnaire for Veterans.
*Lower completion rates are shown for some question categories because participants were instructed to skip portions of the questionnaire that were not applicable, such as males skipping female specific questions.
†When appropriate, weighted kappas were used.
‡Interpretation from Landis and Koch.[29]

Question Groupings

To investigate the nonrandom agreement between the first and second data points of the many questions included in the questionnaire, questions were grouped for ease of reporting and interpretation. Questions representing constructs within the same standardized instrument were grouped together (Table 1). Other like-kind variables were then grouped to make a total of 10 broad categories of question types. The demographic grouping included questions on marital status, education, and whether or not the respondent was a twin. The exposure grouping was composed of questions regarding anthrax vaccination, major life events, military and civilian exposures such as witnessing death and occupational hazards, and military status. Women's health questions were those regarding menstruation, pregnancy, and childbirth. Questions regarding the use of complementary or alternative medicine to treat health problems were grouped together, as were questions on smoking patterns and alcohol consumption (including the CAGE questions but excluding the PHQ questions). Questions from the PHQ, PLC-C, and SF-36V standardized instruments representing various topics discussed above were grouped together for analysis. Symptoms and conditions questions included those about health conditions diagnosed by a doctor or other health professionals and those about persistent symptoms, such as severe headache or cough experienced by the participant during the past year.

Statistical Analyses

Descriptive statistics of the Millennium Cohort Panel 1 members were completed. Additionally, to compare the composition of the first panel with those who submitted two surveys, we used chi-square tests to measure statistical differences for demographic and military characteristics: sex, birth year, education level, marital status, military pay grade, race/ethnicity, service component, branch of service, and occupational category.

The Cronbach alpha coefficient was used to investigate the internal consistency in the response patterns of Millennium Cohort members for standardized scales included in the questionnaire (25). A criterion of 90%-complete data within a scale was implemented for inclusion in analyses, with the mean of the response values substituted for those with 10% or less missing (26). Internal consistency of scale measurements was considered satisfactory with Cronbach alpha of 0.7 or greater (27). Internal consistency was estimated for 5 PHQ scores, 8 SF-36V scores, the CAGE, exposure questions from the National Health Survey of Gulf War Era Veterans and Their Families, and the PCL-C.

Weighted and nonweighted kappa statistics were used to measure the degree of nonrandom agreement between measurements of the same categorical variable retested within 6 months (28). We used 0.8 to 1.0 to distinguish almost perfect agreement, 0.6 to 0.8 to distinguish substantial

AEP Vol. 17, No. 7
July 2007: 525–532

agreement, 0.4 to 0.6 to distinguish moderate agreement, 0.2 to 0.4 to distinguish fair agreement, and 0.0 to 0.2 to distinguish slight or poor agreement (29). The measure of concordance among individual questions was summed and averaged across sections of the survey to determine average agreement by question grouping. In addition, correlation coefficients between the first and second surveys were calculated for the continuous mental and physical component summary scores. Additional analyses of subgroups of the population were completed to investigate differences in reliable reporting among distinct groups of respondents. All data analyses were completed with the use of SAS software (Version 9.1, SAS Institute, Inc., Cary, NC) (30).

RESULTS

Of the 77,047 Millennium Cohort Panel 1 participants, 76,742 (99.6%) had complete demographic and military characteristic data. This population included 73% men, 73% born between 1960 and 1979, 49% without any college experience, 63% married, 70% white non-Hispanic, 77% enlisted personnel, 57% active duty personnel, 48% Army, 20% working as functional support specialists, and 20% combat specialists (Table 2).

Levels of internal consistency among standardized survey scales, as measured by Cronbach alpha coefficient, are reported in Table 3. Completion percentages are reported and do not reflect "skip pattern" questions where participants were instructed to pass over that portion of the questionnaire if they did not exhibit behaviors addressed by the section. Internal consistency measures of mental and physical functioning were well above the 0.7 threshold for being satisfactory except for those scoring alcohol disorders. Scores ranged from 0.58 for the PHQ section on alcohol abuse to a high of 0.94 for the Gulf War veterans' questionnaire pertaining to exposures and the PCL-C. Although the overall completion rate of these instruments was high, respondents were only asked to answer questions that were pertinent to them; for example, those not reporting a panic or anxiety attack in recent weeks were not required to answer questions directed at those types of events, nor were alcohol abstainers required to answer questions about alcohol consumption. Reported data completion rates, therefore, reflect the number of people answering the question grouping and do not take into account the ability of participants to skip a section that was not relevant to them (Table 3).

Participants who submitted two surveys within 6 months were similar in composition to the larger participating population (Table 2). When comparing military and demographic characteristics, only sex and service branch were significantly different, with more female and Army personnel submitting multiple questionnaires. However, although not statistically significant at the alpha = 0.05 level, the differences were approaching statistical significance and suggested those in the test-retest population were older and high-school educated.

Table 1 reports the kappa analyses of the 470 participants submitting two questionnaires within 6 months. Because of the large number of questions on the survey, kappa scores were averaged over several question groupings for meaningful reporting. Weighted kappa values were used in those averages in which the variable in question had more than two possible ordinal values. Almost-perfect agreement was found among demographic and smoking questions (Table 1). Mean kappa scores for questions relating to women's health, as well as questions about exposures, also scored over 0.6, the cutoff point for "substantial agreement." All other categories were found to have mean kappa statistics that were between 0.4 and 0.6, achieving "moderate agreement." Of these, the lowest kappa value was found in questions regarding general health symptoms and conditions (K = 0.48) (Table 1).

Investigation of the correlation between the first and second surveys for the continuous mental and physical component summary scores found a high level of correlation between the test and retest (mental component summary score correlation = 0.77; physical component summary score correlation = 0.74).

Table 4 presents mean kappa scores for broad question groupings stratified by demographic and military characteristics for sex, birth year, education level, marital status, race/ethnicity, military pay grade, service component, and service branch. Men and women differed little in consistency of reporting, although women had a substantially higher agreement for symptom and condition questions than men. Marine Corps personnel were the most consistent service group with regard to reporting of all question groupings. Otherwise, kappa scores varied over characteristic and question groupings, such that no discernible patterns could be noted (Table 4).

DISCUSSION

Standardized instruments are often employed to enhance the value of epidemiologic survey research. Diligence in establishing consistency and comparability to promote confidence in results will become increasingly more important. While the use of established survey instruments may be an enticing addition in pursuit of quality health metrics, suboptimal performance in varying populations may be found instead. In this study, the internal consistency of well-known instruments (PHQ, SF-36V, CAGE, and PCL-C) was assessed in a large military cohort. Additionally, 470 participants who submitted two surveys within a 6-month period

TABLE 2. Demographic characteristics of Millennium Cohort participants and those subjects submitting two questionnaires within a 6-month time period

Characteristics	Millennium Cohort participants (N = 76,742) No. (%)	Millenium Cohort reliability analysis subjects (N = 470) No. (%)	p Value*
Sex			0.03
Male	56,219 (73.3)	324 (68.9)	
Female	20,523 (26.7)	146 (31.1)	
Birth year			0.06
Pre-1960	16,584 (21.6)	121 (25.7)	
1960–1969	29,044 (37.9)	155 (33.0)	
1970–1979	26,581 (34.6)	162 (34.6)	
1980 forward	4,533 (5.9)	32 (6.8)	
Education level			0.06
Less than high school diploma/equivalent	4,716 (6.1)	37 (7.9)	
High school diploma	32,862 (42.8)	222 (47.2)	
Some college	19,623 (25.6)	99 (21.1)	
Bachelor's degree	12,621 (16.5)	75 (16.0)	
Master's/PhD	6,920 (9.0)	37 (7.9)	
Marital status			0.94
Never married	23,057 (30.0)	142 (30.2)	
Married	48,430 (63.1)	294 (62.6)	
Divorced	5,255 (6.9)	34 (7.2)	
Race/ethnicity			0.18
White, non-Hispanic	53,439 (69.6)	309 (65.7)	
Black, non-Hispanic	10,574 (13.8)	72 (15.3)	
Other	12,729 (16.6)	89 (18.9)	
Military rank			0.38
Enlisted	59,173 (77.1)	375 (79.8)	
Officer	17,569 (22.9)	95 (20.2)	
Service component			0.31
Active duty	43,790 (57.1)	279 (59.4)	
Reserve/National Guard	32,952 (42.9)	191 (40.6)	
Branch of service			<0.01
Army	36,427 (47.5)	283 (60.2)	
Air Force	22,343 (29.1)	91 (19.4)	
Navy/Coast Guard	14,047 (18.3)	75 (16.0)	
Marines	3,925 (5.1)	21 (4.5)	
Occupational category			0.74
Combat specialists	15,382 (20.0)	90 (19.2)	
Electronic equipment repair	6,760 (8.8)	40 (8.5)	
Communications/ intelligence	5,400 (7.0)	32 (6.8)	
Health care	7,946 (10.4)	42 (8.9)	
Other technical	1,971 (2.6)	10 (2.1)	
Functional support	15,366 (20.0)	90 (19.2)	
Electrical/mechanical repair	11,360 (14.8)	85 (18.1)	
Craft workers	2,376 (3.1)	13 (2.8)	
Service and supply handlers	6,669 (8.7)	46 (9.8)	
Trainees and other	3,512 (4.6)	22 (4.7)	

*p Values based on the Pearson chi-square test of association.

allowed for the investigation of concordance of all components in the survey, without necessitating costly additional contact and potential negative impact of retesting a subset of the enrolled participants. Comparison of reliability measures over various domains of the questionnaire may be expected to differ because of differences in what is being measured. However, with some noted disparities, the standardized instruments were shown to have high internal consistency in this population, while the investigation of test-retest reliability found moderate to excellent agreement over many question groupings, even among diverse demographic groups.

High internal consistency of several standard instruments suggests that individual Cohort members reported consistently higher or lower levels on questions that correlated well with responses to the other questions of that construct. The finding of lower internal consistency for alcohol misuse, as measured by the PHQ and CAGE, suggests a less consistent response pattern and may reflect a tendency to vary reporting of responses by individuals who perceive their own behavior to be problematic. Military personnel may be less likely to endorse alcohol-related questions because of real or perceived concern for acknowledging career-limiting behavior. Another explanation for inconsistent responses may be that these instruments in fact measure several attributes rather than one, thus deflating the Cronbach alpha and making them imperfect instruments for this type of population.

The population for investigation of concordance among survey questions was reflective of the Cohort in general. This gave the study team a low-cost alternative to traditional test and retest methods. Individual kappa values were calculated for each question on the Millennium Cohort survey and then grouped according to type for more meaningful reporting of results. Demographic questions (pertaining to marital status, education level, and whether the subject was a twin) yielded the highest mean kappa, as would be expected for stationary variables (Table 1). Kappa values were lower for other categories, all of which contained questions that were much more dynamic than the demographic questions, and may be less constant in nature, even over brief periods. While some questions did pertain to exposures or behaviors over the subject's entire lifetime, several questions asked about experiences of the last 2 weeks, 4 weeks, or recent months. This was particularly true of questions regarding anxiety and psychological conditions. Because a 6-month difference between the first and second survey was acceptable, it is understandable that responses to these questions may have changed, causing a lower kappa value than was seen for the demographic questions. In addition, the tragedies of September 11, 2001 occurred about 2 months after the beginning of enrollment for this study; therefore changed answers to questions about

TABLE 3. Internal consistency of the Millennium Cohort Study questionnaire standardized components

Standardized component	Completion rate* No. (%)	Cronbach alpha coefficient	Interpretation†
PHQ			
Somatoform disorder	72,760 (94.4)	0.76	Satisfactory
Depressive disorders	73,519 (95.4)	0.89	Satisfactory
Panic syndrome	3,368 (4.4)	0.76	Satisfactory
Other anxiety syndrome	25,384 (32.9)	0.75	Satisfactory
Alcohol abuse	55,104 (71.5)	0.58	Not satisfactory
SF-36V			
Physical functioning	76,725 (99.6)	0.92	Satisfactory
Role–physical	76,515 (99.3)	0.87	Satisfactory
Bodily pain	76,568 (99.4)	0.90	Satisfactory
General health	76,890 (99.8)	0.82	Satisfactory
Vitality	76,672 (99.5)	0.85	Satisfactory
Social functioning	76,692 (99.5)	0.85	Satisfactory
Role–emotional	76,530 (99.3)	0.86	Satisfactory
Mental health	76,674 (99.5)	0.80	Satisfactory
CAGE	55,104 (71.5)	0.64	Not satisfactory
Gulf War Veterans questionnaire	71,512 (92.8)	0.94	Satisfactory
PCL-C	71,513 (92.8)	0.94	Satisfactory

PCL-C = Posttraumatic Stress Disorder Patient Checklist, Civilian Version; PHQ = Primary Care Evaluation of Mental Disorders (PRIME-MD) Patient Health Questionnaire; SF-36V = Medical Outcomes Study Short Form 36-item questionnaire for Veterans.
*Lower completion rates are shown for panic syndrome, other anxiety syndrome, alcohol abuse, and the CAGE questions because participants were instructed to skip portions of the questionnaire that were not applicable.
†Interpretation from Nunnally.[27]

anxiety and psychological questions would not be surprising, since it has been reported that military members demonstrated healthier scores on the mental health sections of the Millennium Cohort questionnaire shortly after September 11 when compared with those who answered the same questions prior to September 11 (31). Although the same subpopulation was not used for these test-retest analyses, there were 54 individuals of the 470 for the test-retest analyses who submitted their first survey before and their second survey after September 11, 2001. The positive changes in mental health of US service members after September 11 may underestimate the true concordance in this population, thus resulting in lower kappa values. However, this effect would likely be small.

There are limitations of these analyses that should be noted. First, data for the concordance investigation were from a self-selected, nonrandom subset of a large military cohort who voluntarily submitted a second survey within 6 months of the first survey. Participants submitting a second, unsolicited questionnaire were similar to the Cohort in marital status, race/ethnicity, rank, military component, and occupation but differed statistically by sex and service branch. Age and education differed slightly but without being statistically significant. Those participants who had experienced adverse physical and mental health outcomes after their first survey submission may have been more inclined to submit an additional survey in order to report the illness. Service members who were deployed to Afghanistan or other combat zones may have had less of an opportunity to submit a second survey. Also, the kappa statistic is dependent on the true prevalence of the variable being examined. In addition to reflecting changed responses over time, low prevalence of positive answers to some questions may have artificially lowered kappa scores. As prevalence deviates from 0.5, maximum kappa values decrease (32). This was especially evident in questions asked on the Millennium Cohort survey about general health; of 58 questions asked about physician-diagnosed conditions and other symptom-specific health problems, 31 had a prevalence of less than 3%, limiting the kappa statistic's ability to measure nonrandom agreement. There may also be a potential for bias caused by imputing values that comprise a scale with the mean of that scale. However, the Millennium Cohort participant question completion rate has been reported to be predominantly between 95% and 100%, with a low of 84% (5). This strong completion rate may lessen the potential for this type of bias. Lastly, although the standard instruments included in the Millennium Cohort questionnaire have undergone testing in clinical settings and have been found to correlate well with a physician's assessment, the use of a standardized instrument for self-reported data as a surrogate for physician diagnosis is imperfect.

There were strengths, however, of these analyses. The robust size of the Cohort yielded more precise estimates of internal consistency for the standardized instruments included in the questionnaire. Additionally, with a population of 470 repeat participants who were similar to the composition of the Cohort and who had complete covariate data, a variety of estimates of concordance were possible. This large sample of military members also allowed for a breakdown of analysis by specific demographics, the extent of which has not been reported elsewhere. Finally, because this was not a traditional test-retest investigation, there was no need to truncate the survey in any way for retesting; the entire survey was completed twice.

The Millennium Cohort Study was launched in 2001 to assess the effect of military service on long-term mental and physical health outcomes. In these analyses, our findings suggest that the PHQ, SF-36V, and the PCL-C are reliable instruments that can be used to properly assess the mental and physical health of military members. The overall stability of responses to questions should be reassuring to other

TABLE 4. Mean kappa* scores for demographic subgroups of the Millennium Cohort Study

Characteristic	Demographic	Women's health	Exposure	PHQ	SF-36V	Symptoms and conditions	PCL-C	CAM	Alcohol†	Smoking
Sex										
Male	0.85	—‡	0.62	0.54	0.49	0.43	0.55	0.54	0.55	0.82
Female	0.89	0.66	0.59	0.52	0.50	0.64	0.54	0.54	0.44	0.83
Birth year										
Pre-1960	0.89	0.73	0.59	0.57	0.51	0.53	0.58	0.63	0.62	0.85
1960–1969	0.71	0.68	0.60	0.53	0.48	0.59	0.55	0.48	0.47	0.76
1970–1979	0.91	0.61	0.65	0.47	0.45	0.43	0.48	0.40	0.50	0.82
1980 forward	0.78	0.76	0.54	0.62	0.49	0.64	0.66	0.49	0.65	0.85
Education level										
Less than high school diploma/equivalent	0.91	0.74	0.66	0.49	0.38	0.59	0.45	0.46	0.54	0.83
High school diploma	0.84	0.67	0.57	0.51	0.47	0.47	0.52	0.54	0.52	0.80
Some college	0.93	0.54	0.70	0.50	0.53	0.47	0.61	0.71	0.61	0.83
Bachelor's degree	0.96	0.73	0.65	0.60	0.57	0.71	0.66	0.50	0.54	0.89
Master's/PhD	0.63	0.79	0.58	0.50	0.36	0.69	0.29	0.47	0.65	0.78
Marital status										
Never married	0.86	0.61	0.67	0.49	0.47	0.56	0.55	0.38	0.53	0.83
Married	0.80	0.70	0.60	0.56	0.49	0.48	0.55	0.59	0.56	0.82
Divorced	0.98	0.72	0.52	0.48	0.56	0.52	0.49	0.59	0.36	0.82
Race/ethnicity										
White, non-Hispanic	0.81	0.65	0.60	0.55	0.49	0.47	0.53	0.50	0.52	0.82
Black, non-Hispanic	0.96	0.61	0.55	0.48	0.49	0.57	0.55	0.48	0.54	0.80
Other	0.85	0.72	0.65	0.53	0.51	0.56	0.58	0.67	0.60	0.84
Military rank										
Enlisted	0.86	0.66	0.67	0.54	0.49	0.47	0.54	0.52	0.53	0.81
Officer	0.66	0.72	0.58	0.58	0.49	0.62	0.46	0.64	0.52	0.92
Service component										
Active duty	0.77	0.65	0.59	0.52	0.48	0.48	0.54	0.52	0.52	0.83
Reserve/National Guard	0.91	0.66	0.69	0.54	0.49	0.47	0.54	0.54	0.55	0.82
Branch of service										
Army	0.85	0.64	0.66	0.54	0.48	0.47	0.53	0.54	0.53	0.81
Air Force	0.98	0.64	0.58	0.55	0.57	0.51	0.59	0.57	0.58	0.87
Navy/Coast Guard	0.90	0.79	0.59	0.44	0.49	0.60	0.50	0.44	0.50	0.78
Marines	1.00	—‡	0.69	0.63	0.63	0.64	0.66	0.72	0.58	0.97
Occupational category										
Combat specialists	0.80	—‡	0.57	0.56	0.44	0.46	0.59	0.57	0.55	0.80
Electronic equipment repair	0.91	0.78	0.48	0.48	0.46	0.49	0.49	0.34	0.64	0.82
Communications/intelligence	0.80	0.59	0.59	0.43	0.45	0.63	0.52	0.57	0.71	0.91
Health care	0.87	0.59	0.61	0.58	0.50	0.61	0.52	0.65	0.68	0.82
Other technical	1.00	—‡	0.72	0.70	0.64	0.72	0.75	0.21	0.90	0.99
Functional support	0.95	0.66	0.58	0.51	0.46	0.48	0.53	0.60	0.62	0.82
Electrical/mechanical repair	0.95	0.81	0.62	0.51	0.56	0.60	0.53	0.58	0.48	0.85
Craft workers	0.95	—‡	0.79	0.51	0.47	0.65	0.46	0.81	0.81	0.94
Service and supply handlers	0.97	0.68	0.68	0.57	0.55	0.66	0.52	0.55	0.49	0.85
Trainees and other	0.94	0.69	0.68	0.61	0.60	0.63	0.64	0.33	0.57	0.90

CAM = complementary and alternative medicine; PCL-C = Posttraumatic Stress Disorder Patient Checklist, Civilian Version; PHQ = Primary Care Evaluation of Mental Disorders (PRIME-MD) Patient Health Questionnaire; SF-36V = Medical Outcomes Study Short Form 36-item questionnaire for Veterans.
*When appropriate, weighted kappas were used.
†Alcohol questions in this category are not from the PHQ.
‡These categories did not contain enough observations for kappa valuess to be calculated.

military health researchers. The Millennium Cohort questionnaire will continue to be used to gather data on military members until 2022. Continued reliability assessments, such as those reported herein, will help to establish the utility of the many health and exposure assessment instruments included in the questionnaire.

We thank Scott L. Seggerman from the Management Information Division, Defense Manpower Data Center, Seaside, CA. Additionally, we thank Laura Chu, MPH, Lacy Farnell, Gia Gumbs, MPH, Cynthia Leard, MPH, Travis Leleu, Robert Reed, MS, Steven Spiegel, Christina Spooner, MS, Damika Webb, Keri Welch, MA, James Whitmer, and Sylvia Young, MD, MPH, from the Department of Defense Center for Deployment Health Research, Naval Health Research Center, San Diego, CA. We

also thank Dr. Nicole Bell and Laura Senier, from the Army Research Institute of Environmental Medicine, Natick, MA. We appreciate the support of the Henry M. Jackson Foundation for the Advancement of Military Medicine, Rockville, MD.

REFERENCES

1. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. Health Technol Assess. 1998;2 i–iv, 1–74.
2. Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ Primary Care Study. Primary care evaluation of mental disorders. JAMA. 1999;282:1737–1744.
3. Blanchard EB, Jones-Alexander J, Buckley TC, Forneris CA. Psychometric properties of the PTSD Checklist (PCL). Behav Res Ther. 1996;34:669–673.
4. Gray GC, Chesbrough KB, Ryan MAK, Amoroso PJ, Boyko EJ, Gackstetter GD, et al. The Millennium Cohort Study: a 21-year prospective cohort study of 140,000 military personnel. Mil Med. 2002;167:483–488.
5. Ryan MAK, Smith TC, Smith B, Amoroso PJ, Boyko E, Gray GC, et al. Millennium cohort: enrollment begins a 21-year contribution to understanding the impact of military service. J Clin Epidemiol. 2007;60:181–191.
6. Ware JE, Kosinski M, Gandek B. SF-36 Health Survey: manual and interpretation guide. Lincoln (RI): Quality Metric Incorporated; 2000.
7. Ware JE, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36). I. Conceptual framework and item selection. Med Care. 1992;30:473–483.
8. Spitzer RL, Williams JB, Kroenke K, Linzer M, de Gruy FV, Hahn SR, et al. Utility of a new procedure for diagnosing mental disorders in primary care. The PRIME-MD 1000 Study. JAMA. 1994;272:1749–1756.
9. Spitzer RL, Williams JB, Kroenke K, Hornyak R, McMurray J. Validity and utility of the PRIME-MD Patient Health Questionnaire in assessment of 3000 obstetric-gynecologic patients: the PRIME-MD Patient Health Questionnaire Obstetrics-Gynecology Study. Am J Obstet Gynecol. 2000;183:759–769.
10. Weathers FW, Litz BT, Herman DS, Huska JA, Keane TM. The PTSD Checklist (PCL): reliability, validity, and diagnostic utility. Paper presented at the Annual Meeting of International Society for Traumatic Stress Studies. San Antonio, TX. Available at: http://www.pdhealth.mil/library/downloads/PCL_sychometrics.doc. Last accessed July 10, 2006.
11. Ewing JA. Detecting alcoholism. The CAGE questionnaire. JAMA. 1984;252:1905–1907.
12. Fann JR, Bombardier CH, Dikmen S, Esselman P, Warms CA, Pelzer E, et al. Validity of the Patient Health Questionnaire-9 in assessing depression following traumatic brain injury. J Head Trauma Rehabil. 2005;20:501–511.
13. Means-Christensen AJ, Arnau RC, Tonidandel AM, Bramson R, Meagher MW. An efficient method of identifying major depression and panic disorder in primary care. J Behav Med. 2005;28:565–572.
14. Jones D, Kazis L, Lee A, Rogers W, Skinner K, Cassar L, et al. Health status assessments using the Veterans SF-12 and SF-36: methods for evaluating outcomes in the Veterans Health Administration. J Ambul Care Manage. 2001;24:68–86.
15. Ware JE, Kosinski M, Keller SD. SF-36 Physical and Mental Health Summary Scales: A user's manual. Boston (MA): Health Assessment Laboratory; 1994.
16. Ware JE Jr. SF-36 Health Survey update. Spine. 2000;25:3130–3139.
17. Perlin J, Kazis LE, Skinner K, Ren XS, Lee A, Rogers WH, et al. Health status and outcomes of veterans: Physical and mental component summary scores Veterans SF-36 1999 Large Health Survey of Veteran Enrollees. Executive Report. Washington (DC): Department of Veterans Affairs, Veterans Health Administration, Office of Quality and Performance; 2000.
18. Kang HK, Mahan CM, Lee KY, Magee CA, Murphy FM. Illnesses among United States veterans of the Gulf War: a population-based survey of 30,000 veterans. J Occup Environ Med. 2000;42:491–501.
19. Gray GC, Reed RJ, Kaiser KS, Smith TC, Gastanaga VM. The Seabee Health Study: self-reported multi-symptom conditions are common and strongly associated among Gulf War veterans. Am J Epidemiol. 2002;155:1033–1044.
20. Weathers FW, Huska JA, Keane T. The PTSD Checklist Military Version (PCL-M). Boston (MA): National Center for PTSD; October 1991.
21. Lang AJ, Laffaye C, Satz LE, Dresselhaus TR, Stein MB. Sensitivity and specificity of the PTSD checklist in detecting PTSD in female veterans in primary care. J Trauma Stress. 2003;16:257–264.
22. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4th ed. DSM-IV. Washington (DC): American Psychiatric Association; 1994.
23. Hoge CW, Castro CA, Messer SC, McGurk D, Cotting DI, Koffman RL. Combat duty in Iraq and Afghanistan, mental health problems, and barriers to care. N Engl J Med. 2004;351:13–22.
24. Wright KM, Huffman AH, Adler AB, Castro CA. Psychological screening program overview. Mil Med. 2002;167:853–861.
25. Cronbach L. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16:297–334.
26. Afifi AA, Elashoff RM. Missing observations in multivariate statistics. Part 1. Review of the literature. J Am Stat Assoc. 1966;61:595–604.
27. Nunnally JC. Psychometric theory. 2nd ed. New York: McGraw-Hill; 1978.
28. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20:37–46.
29. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–174.
30. SAS Institute Inc. SAS/STAT® 9.1 Users Guide. Cary (NC): SAS Institute Inc.; 2004.
31. Smith TC, Smith B, Corbeil TE, Riddle JR, Ryan MA, for the Millennium Cohort Study Team. Self-reported mental health among US military personnel, prior and subsequent to the terrorist attacks of September 11, 2001. J Occup Environ Med. 2004;46:775–782.
32. Thompson WD, Walter SD. A reappraisal of the kappa coefficient. J Clin Epidemiol. 1988;41:949–958.