International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 22(4): 288–302 (2013)
Published online in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/mpr.1399

Response bias, weighting adjustments, and design effects in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS)

RONALD C. KESSLER,[1] STEVEN G. HEERINGA,[2] LISA J. COLPE,[3] CAROL S. FULLERTON,[4] NANCY GEBLER,[2] IRVING HWANG,[1] JAMES A. NAIFEH,[4] MATTHEW K. NOCK,[5] NANCY A. SAMPSON,[1] MICHAEL SCHOENBAUM,[3] ALAN M. ZASLAVSKY,[1] MURRAY B. STEIN[6,7] & ROBERT J. URSANO[4]

1 Department of Health Care Policy, Harvard Medical School, Boston, MA, USA
2 University of Michigan, Institute for Social Research, Ann Arbor, MI, USA
3 National Institute of Mental Health, Bethesda, MD, USA
4 Center for the Study of Traumatic Stress, Department of Psychiatry, Uniformed Services University School of Medicine, Bethesda, MD, USA
5 Department of Psychology, Harvard University, Cambridge, MA, USA
6 Departments of Psychiatry and Family and Preventive Medicine, University of California San Diego, La Jolla, CA, USA
7 VA San Diego Healthcare System, San Diego, CA, USA

Key words
suicide, mental disorders, US Army, epidemiologic research design, design effects, sample bias, sample weights, survey design efficiency, survey sampling

Correspondence
Ronald C. Kessler, Department of Health Care Policy, Harvard Medical School, Boston, MA, USA.
Telephone (+1) 617-432-3587,
Fax (+1) 617-432-3588
Email: NCS@hcp.med.harvard.edu

Received 10 July 2013;
accepted 15 July 2013

Abstract

The Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS) is a multi-component epidemiological and neurobiological study designed to generate actionable recommendations to reduce US Army suicides and increase knowledge about determinants of suicidality. Three Army STARRS component studies are large-scale surveys: one of new soldiers prior to beginning Basic Combat Training (BCT; $n = 50{,}765$ completed self-administered questionnaires); another of other soldiers exclusive of those in BCT ($n = 35{,}372$); and a third of three Brigade Combat Teams about to deploy to Afghanistan who are being followed multiple times after returning from deployment ($n = 9421$). Although the response rates in these surveys are quite good (72.0–90.8%), questions can be raised about sample biases in estimating prevalence of mental disorders and suicidality, the main outcomes of the surveys based on evidence that people in the general population with mental disorders are under-represented in community surveys. This paper presents the results of analyses designed to determine whether such bias exists in the Army STARRS surveys and, if so, to develop weights to correct for these biases. Data are also presented on sample inefficiencies introduced by weighting and sample clustering and on analyses of the trade-off between bias and efficiency in weight trimming. *Copyright © 2013 John Wiley & Sons, Ltd.*

Introduction

The Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS; http://www.army-starrs.org) is a multi-component epidemiological and neurobiological study of risk and resilience factors for suicidality and its psychopathological correlates among US Army personnel (Ursano *et al.*, under review). One of these components, the Historical Administrative Data Study (HADS) is a study examining associations among information collected on all soldiers (2004–2009) using Army and Department of Defense (DoD) administrative data records to predict suicide outcomes. Two others are retrospective case–control studies of suicide attempts and fatalities. The other main component studies in Army STARRS are three large-scale surveys (Kessler *et al.*, 2013). One of these, the New Soldier Study (NSS), attempted to obtain information from self-administered neurocognitive tests and self-administered questionnaires (SAQs) in a representative sample of over 57,000 of new soldiers reporting for Basic Combat Training (BCT) (Heeringa *et al.*, 2013). The second, the All-Army Study (AAS), attempted to obtain SAQ information from a representative sample of nearly 50,000 soldiers other than those in BCT (Heeringa *et al.*, 2013). The third, the Pre-Post Deployment Study (PPDS), attempted to obtain SAQ information from all 10,380 members of three Brigade Combat Teams scheduled to deploy to Afghanistan shortly after the baseline PPDS was carried out (Heeringa *et al.*, 2013). The NSS and PPDS additionally attempted to collect blood samples from all respondents, while all three studies attempted to obtain informed consent from SAQ respondents to link their Army/ DoD administrative records with their self-report responses.

An important characteristic of the Army STARRS surveys is that identifying information is needed from SAQ respondents to link administrative records with SAQ data. Concerns can be raised about the absence of anonymity in this design, as some military researchers have suggested that lack of anonymity can lead to under-reporting of emotional problems in military surveys (Warner *et al.*, 2008; Warner *et al.*, 2007). A number of large military surveys, like the DoD Survey of Health Related Behaviors Among Active Duty Military Personnel (DoD Health Behavior Surveys; Ryan *et al.*, 2007) and the Mental Health Surveillance Surveys in combat environments carried out by US Army Mental Health Advisory Teams (MHATs; Bliese *et al.*, 2011), are administered anonymously based on this concern in an effort to encourage complete and accurate reporting.

A good deal of methodological research has been carried out on the effects of anonymity in surveys. One line of this research investigates the effects of experimentally manipulating perceived risk of disclosure of survey responses (Couper *et al.*, 2008, 2010). These studies find that only when risk of disclosure is virtually certain and the information in the survey is potentially damaging to the individual does risk of disclosure reduce survey response rates. Emphasizing the confidentiality of responses in identified surveys, however, has been shown consistently to increase survey response rates significantly (Edwards *et al.*, 2009). Based on this evidence, the informed consent sessions preceding the Army STARRS surveys were designed to be quite elaborate (30-minute group-based sessions) and presented detailed information on the tight security measures put in place to guarantee survey response confidentiality.

A second line of experimental research investigates the effects of anonymity on honesty of responding to sensitive questions among people who participate in surveys. The results of this research are mixed, with some studies showing that anonymity increases reports of embarrassing behaviors (Ong and Weiss, 2000; Werch, 1990) and others finding no such effects (Brink, 1995; Campbell and Waters, 1990). It is unclear why this variability exists, but it has been found even in studies examining the same types of behaviors (Begin *et al.*, 1979; Fidler and Kleinknecht, 1977). A broader experimental literature documents effects of "social distance" on reporting of potentially embarrassing behaviors even within anonymous surveys, with highest reported rates in self-administered surveys, lower rates in telephone surveys, and lowest rates in face-to-face surveys (Rogers *et al.*, 1998; Turner *et al.*, 1998).

Non-experimental studies have also been carried out on this issue. For example, a meta-analysis of studies designed to estimate prevalence of major depression in surveys of military samples found that anonymous surveys, all else equal, yielded higher prevalence estimates than confidential surveys that were not anonymous (Gadermann *et al.*, 2012). The most dramatic non-experimental evidence for such an effect came in a study of responses to the Post-Deployment Health Assessment (PDHA) in a sample of infantry soldiers returning from Iraq (Warner *et al.*, 2011). Completion of the PDHA is required of all soldiers returning from deployment. PDHA responses are neither anonymous nor confidential, as each soldier who completes a PDHA is required to have an in-person review of responses with a health care provider and to discuss deployment-related health problems reported in the survey and to allow the health care professional an opportunity to provide referrals for needed treatment (http://www.pdhealth.mil/dcs/dd_form_2796.asp). The effects of this lack of

confidentiality on PDHA responses were examined by administering a completely anonymous survey containing some of the same questions as the PDHA about emotional problems to a group of soldiers shortly after they completed the PDHA. Reported prevalence of depression was over three times as high in the anonymous survey as in the PDHA (7.0% versus 1.9%, $\chi^2_1 = 87.7$, $p < 0.001$), with similar differences found for a number of other reports, such as having symptoms of post-traumatic stress disorder (PTSD) (7.7% versus 3.3%, $\chi^2_1 = 48.9$, $p < 0.001$) and of having thoughts-concerns about losing control or hurting someone (8.6% versus 3.4%, $\chi^2_1 = 63.1$, $p < 0.001$).

A number of factors could be involved in the dramatic under-reporting of emotional problems in the PDHA, as respondents know with certainty that their responses will be reviewed in a meeting with a health professional. The situation is quite different, of course, in the Army STARRS surveys, where respondents are guaranteed that their self-reports will be used only for research purposes, that personally identifying information will never be linked to research data, that the identifying information they provide will be maintained securely by the civilian academic research team carrying out the study, and that this identifying information will never be shared with the Army. It is unclear whether lack of anonymity will affect reports of emotional problems in a situation of this sort.

In an effort to address the possibility of such an effect in the Army STARRS surveys, a strategic decision was made to allow Army STARRS survey respondents to provide completely anonymous survey reports. This was done by asking first for informed consent to complete the survey and then asking separately for identifying information to link survey data to administrative data. Importantly, the survey cooperation rates (i.e. the proportions of soldiers attending the consent sessions that agreed to complete the surveys) were comparable to those achieved in anonymous surveys of similar samples (Heeringa *et al.*, 2013). However, meaningful proportions of SAQ respondents in the three surveys chose not to provide identifying information: 22.9% in the NSS ($n = 11{,}633$), 31.4% in the AAS ($n = 11{,}106$), and 21.2% in the baseline PPDS ($n = 1996$). These respondents would presumably either not have completed the surveys or would have under-reported emotional problems in the surveys if the option for anonymous reporting was not provided.

Access to these anonymous surveys made it possible for us to compare the characteristics of soldiers who completed anonymous versus confidential (i.e. not anonymous) surveys. Furthermore, we had access not only to the Army/DoD administrative records of all respondents who completed confidential (i.e. non-anonymous surveys in which respondents provided identifying information for purposes of linking the SAQ responses to their administrative records) but also to a limited amount of de-identified individual-level administrative record data for all soldiers in the Army. The latter data were provided by the Army for purposes of sample post-stratification. We were able to use these data to make part-whole comparisons aimed at investigating basic differences between survey respondents who consented to administrative data linkage and all other soldiers (i.e. both those who did not complete the survey and those who completed the survey but did consent to provide the identifying data needed to link survey responses to administrative records). These comparisons were used to evaluate response bias in the Army STARRS surveys and to develop weighting adjustments designed to correct for these biases to the extent possible by adjusting for two types of differences: (i) differences between the anonymous survey sample and the de-identified survey sample in variables assessed in the survey; and (ii) differences between the de-identified survey sample and the population in variables available in the Army/DoD administrative records. The results of these analyses are presented in the current report. Data are also presented on sample inefficiencies introduced by weighting and time-space clustering and on analyses of the trade-off between bias and efficiency in weight trimming.

Data adjustments and processing

Sample clustering

The time-space clustering of observations in the NSS, AAS, and PPDS studies could lead to inefficiencies in estimation due to increases in the variances of statistics estimated from the survey data (Heeringa *et al.*, 2010). To obtain correct estimates of variances and associated inferences about the survey population, we used design-based methods of estimation (Wolter, 1985) that required us to define strata and within-stratum sampling error calculation units (SECUs) for each sample to characterize the sample design stratification and the time-space clustering of observations within strata. In the case of the NSS, this was done by beginning with the fact that each week between January 2011 and November 2012 NSS group-administered SAQ data collections were conducted with 200 to 400 new soldiers at each of three Army training installations shortly after they arrived for BCT. Both the implicit stratification of the sample by location and time and the "clustering effects" of weekly administrations to groups of incoming soldiers introduced complex design effects. (The weighting of observations, discussed later in the sub-section on case-level missing data, also contributes to

Int. J. Methods Psychiatr. Res. 22(4): 288–302 (2013). DOI: 10.1002/mpr

design effects.) The NSS "two SECU-per-stratum" sampling error calculation model for design-based variance was formed by first defining pseudo strata based on the training facility location of the survey and bi-weekly windows of time. Each of the weekly time-space clusters of respondents was defined as a separate SECU and two-week pairs of SECUs were combined at a specific BCT installation to define strata to capture the stratification influences on time-space clustering. The two-SECU coding approach, while not necessary, was chosen because of its flexibility in permitting design-based variance estimation under both the Taylor Series Linearization (TSL), Balanced Repeated Replication (BRR) and Jackknife Repeated Replication (JRR) methods. The same sampling error calculation model also permits analysts the option to use Bootstrap methods of inference for the complex sample of NSS observations.

The AAS, in comparison, was selected in quarterly replicates at the unit level stratified by Army command and unit size within command. Large units from substrata within commands (where computer-administered interviewing [CAI] was the data collection mode) were typically treated as pseudo-self representing (SR) units and split into two random SECU groups for variance estimation purposes. Splitting was done at the session level whenever possible and at the individual soldier level for units that were surveyed in a single session. Non-SR smaller units were usually paired with another similar unit within the same command and quarterly time period to create a sampling error stratum for variance estimation. Unit pairing was always carried out not only within command, but also within size stratum and survey mode (i.e. either CAI or paper-and-pencil interviewing [PAPI]) in order to allow data to be analyzed within meaningful subgroups of interest (e.g. United States Army Forces Command [FORSCOM]-only, CAI-only, etc.) while still maintaining the ability to perform design-based variance estimation.

The PPDS sample, finally, consisted of all soldiers in three Brigade Combat Teams scheduled to deploy to Afghanistan (and return) in the 2011–2012 time frame. Two of the three were Infantry Brigade Combat Teams (one light infantry, the other airborne), each consisting of six battalions (two infantry and one each of cavalry, fires, special troops, and support) and the third was a Stryker (mechanized infantry) Brigade Combat Team consisting of three infantry battalions, one artillery battalion, one support battalion, a number of separate companies (network support, military intelligence, engineer, anti-tank, and headquarters), and one cavalry squadron.

PPDS was designed as a "census" of all soldiers in these three Brigade Combat Teams. While the three Brigade Combat Teams in the PPDS were selected purposefully because of their deployment schedule, a design-based approach to PPDS estimation and inference serves to capture the influence of non-response and post-stratification weighting adjustments on the sampling error of statistics estimated from the PPDS data The design-based sampling error calculation model developed for the analysis of these data effectively treats the three Brigade Combat Teams as a sample from a super-population of all possible such units that underwent a similar deployment experience. A two SECU-per-stratum sampling error calculation model for PPDS design-based variance estimation was formed by first randomly creating strata of 50 to 100 soldiers within each of these units and then further randomly creating half-samples of soldiers within each of these strata to define SECUs. The two-SECU-per-stratum coding approach, as noted earlier, is not the only one that could have been used to estimate variances, but was used here because of its flexibility in allowing implementation of design-based variance estimation methods of the sort used in substantive analyses of the Army STARRS data.

Adjusting for item-level non-response

Item-missing data are generally more common in SAQs than interviewer-administered surveys. Army STARRS is no exception to this rule, as indicated by the fact that a meaningful proportion of SAQ respondents failed to complete all SAQ items (Heeringa *et al.*, 2013). In addition, sporadic item-level missing data could be found in a substantial proportion of completed SAQs. A two-step process was used to address this problem. First, SAQs were coded as missing if the data pattern suggested that respondents were giving random responses or if the amount of missing data was so large that imputation was infeasible. Second, item-level missing data were imputed using a three-part process that began with conservative rational imputation for missing items in sections that had selective missing items. For example, in the section on exposure to traumatic experiences, missing values for respondents that endorsed some items but left others blank were recoded as negative responses. The second part of this three-part process involved psychometric scales, where respondents were assigned a total scale score based on partial values using model-based imputation (e.g. estimated true score values on an item response theory [IRT] scale). The third part, finally, involved the use of multiple imputation to assign plausible values to item-missing data based on responses to other questions (Schafer, 1999).

Adjusting for case-level missing data

Recruiting difficult-to-reach cases

One way to deal with case-level missing data is to develop special field procedures aimed at tracking, recruiting, and interviewing hard-to-reach cases. These procedures were not used in the NSS, AAS, or baseline PPDS because of logistical constraints. However, these procedures are being used in the third wave of the PPDS follow-up survey by selecting a probability sub-sample of non-respondents at the end of the standard field period and using special tracing procedures, personalized recruitment procedures, and financial incentives to obtain interview data from as many of these cases as possible. Up-weighting of these cases will be used to adjust for the fact that they are being under-represented in the consolidated analysis dataset. Similar procedures will be used in future planned follow-up surveys of the baseline NSS and AAS samples and further follow-ups of the PPDS sample.

Weighting for case-level non-response

As noted in the Introduction, we were able to adjust for case-level missing data by comparing characteristics of respondents with those of non-respondents. This was done in two ways: by comparing SAQ responses of respondents who did versus did not consent to Army/DoD administrative data linkage; and by comparing profiles of SAQ respondents who consented to linkage with population profiles on the small set of administrative record variables (e.g. age, sex, rank) we were given access to for post-stratification. We developed weights based on both of these comparisons to make weighting adjustments for case-level non-response. Weight 1 (WT1) adjusted for discrepancies in SAQ responses of survey completers with versus without record linkage. Weight 2 (WT2) then adjusted for discrepancies in multivariate profiles of weighted (WT1) survey respondents with administrative record linkage versus the population. Each weight was constructed based on an iterative process of stepwise logistic regression analysis designed to arrive at a stable weighting solution. WT1 was the inverse of the probability of agreement to link administrative data with SAQ data in the sample of SAQ completers based on a prediction equation using SAQ responses as predictors. WT2 was the inverse of the probability of completion of the SAQ based on the comparison of SAQ respondents who agreed to linkage and were weighted (WT1) to represent all SAQ respondents compared to the population based on a prediction equation using administrative record variables as predictors.

Inspection of detailed results for the replicates weighted up to now, which consist of NSS and AAS respondents from Q2–4 2011 and the baseline PPDS, shows that survey respondents who consented to administrative record linkage differ from non-consenters in having experienced more stress in their lifetime and the recent past and in having generally higher self-reported rates of Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) mental disorders. However, these differences are not dramatic even though they are statistically significant. This is illustrated in Table 1, which shows that linkage consenters across the three main Army STARRS surveys were somewhat more likely than non-consenters to report having 30-day DSM-IV mental disorders, a history of trauma exposure, and a history of head injuries, but that these differences are quite modest in substantive terms despite being significant from a statistical point of view.

The fact that consenters do not differ dramatically from non-consenters leads to the ratio of high to low weights based on the best-fitting logistic regression equations (i.e. the ratio of $1/p_1$ divided by $1/p_{99}$, where p_1 is the predicted probability of consent for respondents at the first percentile of this probability in the sample and p_{99} is the predicted probability of consent for respondent at the 99th percentile of this probability in the sample) being relatively low: 4.2–8.4 for the NSS, 4.9–9.4 for the AAS, and 1.7 for the PPDS. In addition, the bodies of the weight distributions are fairly symmetrical. These distributional characteristics typically reduce the impact of weights on variances of coefficient estimates (Kish, 1976; Little and Vartivarian, 2005)

Inspection of detailed results of the logistic regression equations used to produce WT2 shows that NSS respondents who provided administrative data linkage consent are somewhat younger than the population of all soldiers eligible for the survey and somewhat more likely than soldiers in the population to be female, non-Hispanic White, never married, and Protestant, but less likely to have no religion, and somewhat more highly educated than all soldiers in the population. NSS respondents with linked administrative data are also somewhat more likely than the population to be in the Regular Army rather than the US Air National Guard (USANG) or US Army Reserve (USAR). Some of these patterns are shown in Table 2, where we see that sample versus population differences are modest in substantive terms even though statistically significant.

Similar patterns of statistically significant but substantially modest sample versus population difference in socio-demographic characteristics were found in the AAS, including the sample being somewhat younger, less female (as opposed to more female in the NSS), more

Int. J. Methods Psychiatr. Res. 22(4): 288–302 (2013). DOI: 10.1002/mpr

Table 1. Selected comparisons on self-administered survey (SAQ) questions of SAQ respondents who consented and provided valid information for administrative data record linkage versus those who did not consent in the three main Army STARRS survey samples[1]

	New Soldier Study (Q2–4 2011)[2]				All Army Study (Q2–4 2011)[2]				Baseline Pre-Post Deployment Study			
	Consented		Did not consent		Consented		Did not consent		Consented		Did not consent	
	Percent	(SE)	Percent	(SE)	Percent	(SE)	Percent	(SE)	Percent	(SE)	Percent	(SE)
Current DSM-IV/CIDI disorders												
30-Day MDE	5.3	(0.3)	4.6	(0.4)	3.3	(0.2)	4.3	(0.4)	4.5	0.2	4.3	0.5
30-Day GAD	4.4	(0.3)	3.6	(0.3)	1.2	(0.1)	1.0	(0.2)	2.4	0.2	2.8	0.4
30-Day Panic	4.2	(0.3)	3.6	(0.3)	3.4	(0.2)	3.4	(0.3)	3.4	0.2	3.2	0.4
30-Day IED	10.7*	(0.4)	8.0	(0.5)	7.5	(0.2)	6.6	(0.4)	8.7	0.3	8.1	0.6
30-Day PTSD	4.5	(0.3)	3.7	(0.3)	4.2	(0.2)	4.7	(0.4)	3.2	0.2	3.0	0.4
Any of the above	17.6	(0.5)	15.5	(0.7)	13.2	(0.3)	13.1	(0.6)	14.3	0.4	13.1	0.8
χ^2_5	18.1*				17.3*				3.4			
History of traumatic life stress												
None	16.6*	(0.5)	24.7	(0.8)	27.7*	(0.4)	31.0	(0.8)	16.3*	0.4	20.3	0.9
Low	28.8	(0.6)	28.3	(0.8)	19.7	(0.4)	18.5	(0.7)	32.9	0.5	33.1	1.1
Intermediate	26.4*	(0.6)	22.7	(0.8)	24.9*	(0.4)	22.3	(0.7)	25.5	0.5	23.6	1.0
High	28.2*	(0.6)	24.3	(0.8)	27.7	(0.4)	28.2	(0.8)	25.3	0.5	22.9	0.9
χ^2_3	86.1*				19.1*				21.8*			
History of head injury												
None	38.9*	(0.7)	48.1	(0.9)	51.8	(0.5)	53.4	(0.9)	36.8*	0.6	43.5	1.1
Low	9.3	(0.4)	8.0	(0.5)	11.5	(0.3)	11.7	(0.6)	10.2	0.4	8.8	0.6
Intermediate	24.7*	(0.6)	19.8	(0.7)	19.5	(0.4)	18.4	(0.7)	24.7	0.5	22.3	0.9
High	27.2*	(0.6)	24.1	(0.8)	17.3	(0.3)	16.5	(0.7)	28.3	0.5	25.4	1.0
χ^2_3	69.6*				4.0				31.5*			
(*n*)	(11,802)		(3256)		(5528)		(2977)		(7425)		(1996)	

*Significant difference between respondents who consented versus did not consent at the 0.05 level, two-sided test.

[1]The NSS target sample consisted of a number of new soldiers in Reception Battalion in designated cohorts equal to the numbers we could accommodate in the group survey administration settings established on the training bases. Survey Research staff worked with Army point-of-contacts (POCs) to select a representative sample of new soldiers from each cohort to fill those quotas. These new soldiers were ordered to attend the Army STARRS consent session but then provided voluntary informed consent for participation in the NSS. The AAS and PPDS target samples were all soldiers in designated units who were ordered to attend the Army STARRS consent session but then provided voluntary informed consent for participation in the study.

[2]The NSS and AAS studies were piloted in Q1 2011 absent the questions about suicidality and the safety plan associated with those questions (which did not receive Institutional Review Board [IRB] approval until after the Q1 replicates were fielded). Full implementation started in Q2 2011, which is why this was the first replicate included in the weighting. Data for 2012 are not reported here because weighting of the 2012 NSS and AAS data are being carried out separately using an updated population post-stratification dataset for that year.

Note: MDE, major depressive episode; GAD, generalized anxiety disorder; Panic, panic disorder; IED, intermittent explosive disorder, PTSD, post-traumatic stress disorder.

Int. J. Methods Psychiatr. Res. 22(4): 288–302 (2013). DOI: 10.1002/mpr

non-Hispanic White, more currently married (as opposed to more "never married" in the NSS), less highly educated (as opposed to more highly educated in the NSS), and less likely to have any religion than soldiers in the population. Some of these patterns are shown in Table 3, where we see that the differences between sample and population are quite modest in substantive terms even though they are statistically significant. Differences between the AAS sample and the population in Army career characteristics are more substantial, though, with a higher proportion of the sample than the population in the lower enlisted ranks (E2–4), having somewhat less time in service, and being more likely to have been deployed exactly once (as opposed either never or more than once). More detailed analyses found that respondents in the sample are more likely than the population to be in the Medical Command and less likely to be in Area Service Component Commands (North/South America, Europe/Central/Africa, Pacific) and to have quite different distributions than the population on Military Occupational Specialties (MOS). These differences are due to differential sampling of units in the first year of the AAS. In the case of the baseline PPDS, finally, differences between sample and population were found to be very modest in all respects other than that the sample was more likely to have deployed two or more times.

The substantial sample versus population differences in the AAS in Command and MOS led to the ratio of consolidated weights (i.e. WT1 × WT2) based on the best-fitting logistic regression equations being a good deal higher (53.3) than for the NSS (14.2) or the PPDS (3.8). However, as with WT1, the consolidated WT1 × WT2 distributions were found to be smooth and fairly symmetric in all three surveys, with no evidence of bimodality toward the extremes. In addition, as respondents with suicidality and mental disorders are over-represented in the samples, respondents with the highest weights tend to be those who do not have these outcomes. This, as shown in the next sub-section, minimizes the adverse effects on sample efficiency that might otherwise occur as a result of weighting. However, it is possible that results will differ in the remaining sample replicates. As a result, all weighting calculations will be repeated in future Army STARRS study replicates once data collection is completed. Consolidated weights will then be created that allow for changes in optimal weighting procedures over the course of the study.

Weighting for under-represented time periods in the ARFORGEN cycle

As noted in an earlier paper in this issue (Kessler *et al.*, 2013), the initial AAS replicates were restricted to the continental United States and only later expanded to include units in other parts of the world. It was not until rather late in the data collection period, furthermore, that we were able to add soldiers who were currently deployed to Afghanistan by interviewing these soldiers when they were passing through Kuwait either leaving for or returning from their mid-tour leave. Other than for those deployed soldiers, the AAS replicates under-represented activated USANG and USAR units in the continental United States due to the fact that soldiers in such units typically activated for only a short time before deployment, spent only a short time in the continental United Stated after returning from deployment prior deactivating, and were reluctant to participate in the AAS during either of these short time periods. For a similar reason, the AAS under-represented units that were scheduled to deploy in the near future as well as units that recently returned from deployment. As we know that the suicide rate is related to these fine-grained time distinctions, the AAS is biased in that it under-represents certain time periods in the unit deployment cycle.

In order to capture such subtleties of a unit's location in the ARFORGEN (Army Forces Generation) cycle we added replicates late in the AAS field period to include USANG and USAR units that (i) were scheduled either to deploy soon after completing the AAS or that (ii) recently returned from Afghanistan and were scheduled to deactivate soon after completing the AAS. In addition, the baseline PPDS sample provided us with information about Brigade Combat Teams that were going to deploy shortly after completing an Army STARRS survey. Importantly, this baseline PPDS survey contained all (and more than) the information in the AAS. In addition, the T2 PPDS survey provided us with comparable information for the same respondents approximately three months after they returned from their deployment. Once the data from all these final surveys are available for analysis, we will combine them with the larger AAS sample to construct a composite portrait of the entire Army with appropriate weights for the cross-classification of Command (i.e. Training and Doctrine Command [TRADOC], Forces Command [FORSCOM], Medical Command [MEDCOM], etc.), Component (i.e. Regular Army, USAR, and USANG), and phase of the ARFORGEN cycle to reproduce the actual distribution of the total Army across the cells of this cross-classification for the time period under study.

Design effects

Conventional methods of estimating significance, which assume a simple random sample, do not take the

Int. J. Methods Psychiatr. Res. 22(4): 288–302 (2013). DOI: 10.1002/mpr

Table 2. Selected comparisons on administrative data record variables of weighted self-administered survey respondents who consented and provided valid information for administrative record linkage versus the population in the three main Army STARRS survey samples[1]

	New Soldier Study (Q2–4 2011)[2]				All Army Study (Q2–4 2011)[2]				Baseline Pre-Post Deployment Study			
	Population		Sample		Population		Sample		Population		Sample	
	Percent	(SE)	Percent	(SE)	Percent	(SE)	Percent	(SE)	Percent	(SE)	Percent	(SE)
Age												
17–20	59.7	(0.1)	60.9	(1.0)	7.2*	(0.0)	8.2	(0.4)	14.3	(0.3)	14.3	(0.4)
21–24	25.2	(0.1)	24.5	(0.7)	22.7*	(0.0)	27.9	(0.6)	34.0	(0.3)	33.7	(0.6)
25–29	9.9	(0.1)	9.9	(0.4)	26.0*	(0.0)	27.5	(0.6)	27.4	(0.3)	27.7	(0.5)
30+	5.1	(0.0)	4.7	(0.2)	44.1*	(0.0)	36.4	(0.6)	24.2	(0.3)	24.2	(0.5)
χ^2_3		9.2*				150.2*				0.3		
Sex												
Male	82.4	(0.1)	81.6	(0.8)	85.8*	(0.0)	87.4	(0.4)	94.5	(0.2)	94.4	(0.3)
Female	17.6	(0.1)	18.4	(0.8)	14.2*	(0.0)	12.6	(0.4)	5.5	(0.2)	5.6	(0.3)
χ^2_1		3.9*				11.7*				0.0		
Race/ethnicity												
Non-Hispanic White	63.7*	(0.1)	65.7	(0.7)	61.2*	(0.0)	66.5	(0.6)	68.3	(0.3)	67.9	(0.6)
Non-Hispanic Black	18.8	(0.1)	19.4	(0.5)	20.5*	(0.0)	16.0	(0.5)	12.6	(0.2)	12.6	(0.4)
Hispanic	11.9*	(0.1)	10.4	(0.4)	11.2	(0.0)	10.6	(0.4)	12.1	(0.2)	12.3	(0.4)
Other	5.6*	(0.1)	4.4	(0.3)	7.1	(0.0)	6.9	(0.3)	7.0	(0.2)	7.2	(0.3)
χ^2_3		51.7*				81.0*				0.4		
Marital status												
Never married	85.7*	(0.1)	87.3	(0.4)	32.0	(0.0)	31.2	(0.6)	44.8	(0.4)	44.3	(0.6)
Previously married	1.3*	(0.0)	0.2	(0.0)	6.4*	(0.0)	5.6	(0.3)	3.7	(0.1)	4.3	(0.3)
Currently married	13.0	(0.1)	12.5	(0.4)	61.5*	(0.0)	63.1	(0.6)	51.5	(0.4)	51.4	(0.6)
χ^2_2		83.1*				9.1*				4.2		
Education												
Less than high school[3]	22.3*	(0.1)	17.9	(0.8)	13.1*	(0.0)	11.2	(0.4)	14.9	(0.3)	14.7	(0.5)
High school	69.2*	(0.1)	72.0	(0.7)	60.0*	(0.0)	68.1	(0.6)	70.6	(0.3)	70.9	(0.6)
Some college	1.9	(0.0)	2.0	(0.2)	3.8	(0.0)	3.9	(0.3)	2.5	(0.1)	3.0	(0.2)
College graduate	6.6*	(0.1)	8.0	(0.6)	23.1*	(0.0)	16.8	(0.5)	12.0	(0.2)	11.4	(0.4)
χ^2_3		131.4*				167.6*				4.5		
Religion												
Christian: Catholic	13.2	(0.1)	12.6	(0.4)	19.0	(0.0)	19.3	(0.5)	18.6	(0.3)	18.6	(0.5)
Christian: Other	52.8*	(0.1)	55.8	(0.6)	52.3	(0.0)	52.1	(0.7)	49.6	(0.4)	49.1	(0.6)
Other religion	1.5	(0.0)	1.7	(0.2)	2.0*	(0.0)	2.5	(0.2)	1.9	(0.1)	1.9	(0.2)

(*Continues*)

Int. J. Methods Psychiatr. Res. 22(4): 288–302 (2013). DOI: 10.1002/mpr

Table 2. (Continued)

	New Soldier Study (Q2–4 2011)[2]				All Army Study (Q2–4 2011)[2]				Baseline Pre-Post Deployment Study			
	Population		Sample		Population		Sample		Population		Sample	
	Percent	(SE)	Percent	(SE)	Percent	(SE)	Percent	(SE)	Percent	(SE)	Percent	(SE)
No religion	29.2*	(0.1)	26.3	(0.6)	18.8*	(0.0)	22.3	(0.6)	21.3	(0.3)	21.8	(0.5)
Unknown	3.4	(0.0)	3.5	(0.3)	8.0*	(0.0)	3.9	(0.3)	8.6	(0.2)	8.6	(0.4)
χ^2_4	53.4*				154.7*				1.0			
Rank												
E1–4					43.8*	(0.0)	53.4	(0.7)	58.9	(0.4)	58.8	(0.6)
E5–9					38.0*	(0.0)	34.0	(0.6)	32.2	(0.3)	32.8	(0.6)
WO or CO					18.2*	(0.0)	12.6	(0.4)	8.9	(0.2)	8.3	(0.3)
χ^2_2					231.5*				2.1			
Time in Army												
0–24 months					16.9*	(0.0)	19.5	(0.5)	31.9	(0.3)	31.5	(0.5)
25–48 months					19.3*	(0.0)	22.6	(0.6)	23.4	(0.3)	23.8	(0.5)
49+ months					63.8*	(0.0)	57.8	(0.7)	44.8	(0.4)	44.7	(0.6)
χ^2_2					85.7*				0.6			
Deployed												
Never					31.7*	(0.0)	29.3	(0.6)	45.0	(0.4)	43.4	(0.6)
Previously 1					31.6*	(0.0)	36.0	(0.6)	31.0	(0.3)	31.1	(0.6)
Previously 2+					36.7*	(0.0)	34.6	(0.6)	24.0	(0.3)	25.5	(0.5)
χ^2_2					44.0*				7.0*			
(*n*)[4]	(212,797)		(11,802)		(3,528,477)		(5528)		(19,182)		(7425)	

*Significant difference between weighted (WT1) SAQ respondents who consented to record linkage and the population.

[1]The samples were weighted (WT1) to adjust for differences on SAQ variables between SAQ respondents who consented and provided linking information versus those that did not, making the weighted sample representative of all SAQ respondents on the SAQ variables. The population data are taken from contemporary administrative data for the populations from which the samples were selected: all soldiers in BCT for the NSS, all non-deployed Regular Army soldiers not in BCT for the AAS, and all soldiers in the three Brigade Combat Teams included in the survey for the PPDS.

[2]The NSS and AAS studies were piloted in Q1 2011 absent the questions about suicidality and the safety plan associated with those questions (which did not receive IRB approval until after the Q1 replicates were fielded). Full implementation started in Q2 2011, which is why this was the first replicate included in the weighting. Data for 2012 are not reported here because weighting of the 2012 NSS and AAS data are being carried out separately using an updated population post-stratification dataset for that year.

[3]Includes alternative education certificate, Army National Guard (ARNG) and General Educational Development (GED).

[4]The population for the NSS is defined as the pooled monthly snapshot of all soldiers in BCT in the time interval April–November 2011. The population for the AAS is defined as the pooled monthly snapshot of all Regular Army soldiers not in BCT and not deployed over the time interval May–December 2011. The population for the baseline PPDS is defined as the pooled monthly snapshot of all soldiers in the three Brigade Combat Teams in the sample in the months before and during baseline data collection.

Table 3. Design effects on selected 30-day outcome variable prevalence estimates due to survey weighting and clustering in the three main Army STARRS survey samples[1]

	New Soldier Study (Q2–4 2011)[2]	All Army Study (Q2–4 2011)[2]	Baseline Pre-Post Deployment Study
Generalized anxiety disorder	1.5	1.0	1.0
Intermittent explosive disorder	1.2	1.6	1.1
Major depressive episode	1.1	1.8	1.0
Panic disorder	1.3	1.3	1.2
Post-traumatic stress disorder	1.2	1.7	1.0
Suicide ideation	1.2	1.5	1.1
Any of the above	1.1	1.9	1.1
(*n*)	(11,802)	(5428)	(7425)

[1]The samples were doubly weighted to adjust for differences on SAQ variables between SAQ respondents who consented and provided linking information for administrative data versus those that did not (WT1) and between the weighted (WT1) sample of SAQ respondents with linked ADS data and the population (WT2).
[2]The NSS and AAS studies were piloted in Q1 2011 absent the questions about suicidality and the safety plan associated with those questions (which did not receive IRB approval until after the Q1 replicates were fielded). Full implementation started in Q2 2011, which is why this was the first replicate included in the weighting. Data for 2012 are not reported here because weighting of the 2012 NSS and AAS data are being carried out separately using an updated population post-stratification dataset for that year.

imprecision introduced by clustering and weighting into account. As a result, special design-based methods of estimating standard errors and significance tests are used in Army STARRS analyses to adjust for the effects of weighting and clustering. The TSL method is the main approach used here (Wolter, 1985), although we also use the more computationally intensive method of JRR (Kish and Frankel, 1974) for applications where a convenient software application using the TSL method is not readily available or for highly non-linear estimation problems in which the linearization of the TSL method might be problematic.

Although the effects of weighting and clustering can be described in a number of ways, a particularly convenient way is to calculate a statistic known as the design effect (DE; Kish, 1965) for a number of variables of interest. The DE is the square of the ratio of the design-based standard error (SE) of a descriptive statistic divided by the simple random sample SE. The DE can be interpreted as the approximate proportional increase in the sample size that would be required to increase the precision of the design-based estimate to the precision of an estimate based on a simple random sample of the same size. DEs due to clustering are usually a good deal larger in estimating means and other first-order statistics than more complex statistics, as the number of respondents having the same characteristics in the same SECU of a single stratum becomes smaller and smaller as the statistics become more complex. This leads to a reduction in the effects of clustering in the estimation of DE. DEs due to weighting are also usually somewhat smaller for multivariate than bivariate descriptive statistics because DEs are due not only to the variance in the weights but also to the strength of the association between the weights and the substantive variables under consideration. Because means typically have higher DEs than other statistics, evaluations of DEs typically focus on the estimation of means. We do the same here.

Seven dichotomous measures of 30-day prevalence of critical outcome variables were included in the evaluation of DEs: suicide ideation and DSM-IV disorder estimates for major depressive episode, generalized anxiety disorder, PTSD, panic disorder, intermittent explosive disorder, and any of the above six outcomes. DEs for these estimates are in the range 1.1–1.5 for the NSS, 1.0–1.9 for the AAS, and 1.0–1.2 for the PPDS. (Table 3) The fact that a number of DEs are 1.0 (i.e.,equal in efficiency to a simple random sample) or only slightly higher than 1.0 can be explained by the same general pattern of the samples with linked administrative data over-representing soldiers with the disorders that are the focus of interest in Army STARRS.

Trimming weights to reduce design effects

As DEs can be sensitive to extreme weights, weight trimming of various sorts is often used to reduce this

sensitivity. We investigated the implications of trimming the final consolidated weight (WT1 × WT2) in each survey. In doing this we took into consideration the fact that even though weight trimming usually reduces the variance of weights, and in this way improves the precision of estimates and the statistical power of tests, it can also lead to bias in estimates if the reduction in variance created due to added efficiency is less than the increase in variance due to bias. It is possible to study this trade-off between bias and efficiency empirically in order to evaluate alternative weight trimming schemes by making use of the equation

$$\mathrm{MSE}_{Yp} = B_{Yp}{}^2 + \mathrm{Var}\left(Y_p\right), \tag{1a}$$

$$= E\left[\left(\hat{B}_{Yp}\right)^2 - \mathrm{Var}\left(\hat{B}_{Yp}\right) + \mathrm{Var}\left(Y_p\right)\right], \tag{1b}$$

where MSE_{Yp} is the mean squared error of the prevalence of outcome variable Y at trimming point p, B_{Yp} is the bias of that prevalence estimate and $\hat{B}_{Yp}$, an unbiased estimate of that bias, $\widehat{\mathrm{Var}}\left(\hat{B}\right)_{Yp}$), is the estimated variance of $\hat{B}_{Yp}$, $\mathrm{Var}(\hat{Y}_p)$ is the estimated variance of estimate $\hat{Y}_p$, and $E[\ ]$ in Equation 1b indicates that the quantity in square brackets is an unbiased estimator of MSE.

Each of the three terms in Equation 1b can be estimated empirically for any value of p, making it possible to calculate MSE across a range of trimming points and select the trimming point that minimizes MSE. The first term, $(\hat{B}_{Yp})^2$, can be estimated directly as $(Y_p - Y_0)^2$, where Y_0 represents the weighted prevalence estimate of Y based on the untrimmed weight. The other two terms in Equation 1b can be estimated using a pseudo-replicate method in which separate estimates for each stratum-SECU are generated for Y_p at each value of p (Zaslavsky *et al.*, 2001). The separate estimates were obtained by sequentially modifying the sample and then generating an estimate based on that modified sample. The modification consisted of removing all cases from one SECU and then weighting the cases in the remaining SECU in the same stratum to have a sum of weights equal to the original sum of weights in that stratum. If we define Y_p as the weighted estimate of Y at trimming point p in the total sample and we define $Y_{p(sn)}$ as the weighted estimate at the same trimming point in the sample that deletes SECU n ($n = 1, 2$) of stratum s ($s = 1$–42), then $\mathrm{Var}(Y_p)$ can be estimated as

$$\widehat{\mathrm{Var}}\left(\hat{Y}_p\right) = \sum_s \left[\left(\hat{Y}_{p(s1)} - Y_p\right)^2 + \left(\hat{Y}_{p(s2)} - \hat{Y}_p\right)^2\right]/2. \tag{2}$$

$\mathrm{Var}(\hat{B}_{Yp})$ was estimated in the same fashion by replacing $\hat{Y}_{p(sn)}$ in Equation 2 with $\hat{B}_{Yp(sn)} = \hat{Y}_{p(sn)} - \hat{Y}_{0(sn)}$ and replacing $\hat{Y}_p$ with $\hat{B}_{Yp(sn)} = \hat{Y}_p - \hat{Y}_0$.

The analysis compared the design-based MSE of 30-day prevalence estimates for the same outcomes as considered in the last sub-section using the consolidated WT1 × WT2 weight and 10 successively more severely trimmed versions of these weights in which between 1% and 10% of cases were trimmed at each tail of the distribution. Trimming consisted of distributing the weights at each of these tails equally across all cases in that tail. MSE_{Y0} was arbitrarily set at 100.0 and all other values were defined in relation to that mean for ease of interpretation. Summary results for illustrative trimming points are presented in Table 4. In the cases of NSS and AAS, while weight trimming reduced MSE for some outcomes (most notably, generalized anxiety disorder in the NSS and major depressive episode in the AAS), it increased MSE for other outcomes, leading us to decide not to trim the consolidated weight for either survey. In the case of PPDS, while the effects of weight trimming were generally positive, they were so modest that we decided not to trim the consolidated weight. As with the weights themselves, it is possible that results regarding the value of weight trimming will differ in the remaining sample replicates. As a result, all weight trimming calculations will be repeated in future Army STARRS study replicates once data collection is completed. Consolidated weight trimming rules will then be created that allow for changes in optimal trimming procedures over the course of the study.

Discussion

As noted in the Introduction, our reading of previous methodological literature led us to expect that Army STARRS survey respondents who agreed to administrative record linkage would have lower rates of self-reported mental disorder than survey respondents who provided identifying information both because those with mental disorders would be less likely to consent to record linkage and because those who did consent would under-report emotional problems. Yet the opposite pattern was found in the data when we examined the predictors of WT1: SAQ respondents who consented to administrative record linkage had significantly higher, not lower, self-reported rates of mental illness than SAQ respondents who did not consent to record linkage.

Why this pattern occurred is unclear. One possibility is that it reflects a positive effect of the message used in respondent recruitment: that Army STARRS is an *independent* research project carried out by academic researchers *outside of the Army* that represents a *unique opportunity* for soldiers to let Army leadership know about issues they are experiencing in the realms of work-related stress and

Int. J. Methods Psychiatr. Res. 22(4): 288–302 (2013). DOI: 10.1002/mpr

Table 4. Effects of weight trimming on the bias-efficiency trade-off for selected outcome variable prevalence estimates in the three main Army STARRS survey samples[1]

	New Soldier Study Q2–4 2011[2]			All Army Study Q2–4 2011[2]			Baseline Pre-Post Deployment Study		
	Trimming point[3]			Trimming point[3]			Trimming point[3]		
	2	5	10	2	5	10	2	5	10
Generalized anxiety disorder	95.9	85.3	72.7	98.7	121.7	192.5	100.2	100.6	98.5
Intermittent explosive disorder	99.8	102.6	82.0	109.1	99.4	101.8	100.1	100.7	100.2
Major depressive episode	105.1	90.9	87.8	116.9	65.5	76.6	98.3	97.5	95.4
Panic disorder	101.4	93.4	94.4	102.9	93.5	101.9	99.2	99.7	97.0
Post-traumatic stress disorder	100.6	111.1	120.2	99.7	138.2	172.4	97.2	95.9	94.9
Suicide ideation	101.2	100.6	148.6	96.1	100.2	148.2	95.6	95.3	92.3
Any 30-day disorder	99.8	89.1	94.2	102.9	102.3	96.0	99.7	99.6	97.4
(*n*)		(11,802)			(5428)			(7425)	

[1]The samples were doubly weighted to adjust for differences on SAQ variables between SAQ respondents who consented and provided linking information for ADS data versus those that did not (WT1) and between the weighted (WT1) sample of SAQ respondents with linked ADS data and the population (WT2).
[2]The NSS and AAS studies were piloted in Q1 2011 absent the questions about suicidality and the safety plan associated with those questions (which did not receive IRB approval until after the Q1 replicates were fielded). Full implementation started in Q2 2011, which is why this was the first replicate included in the weighting. Data for 2012 are not reported here because weighting of the 2012 NSS and AAS data are being carried out separately using an updated population post-stratification dataset for that year.
[3]The trimming point is the proportion of respondents trimmed at each tail of the distribution. Four per cent of respondents (2% at the upper end of the distribution and 2% at the lower end of the distribution) were trimmed in the solution with a trimming point of 2, 10% at a trimming point of 5, and 20% at a trimming point of 10. See the text for a more detailed description of the procedures and rationale for trimming. Results for trimming points at each whole number value in the range 1–10 are available on request.

Int. J. Methods Psychiatr. Res. 22(4): 288–302 (2013). DOI: 10.1002/mpr

emotional problems. This recruitment message went on to say that only a small proportion of all soldiers were invited to participate in the survey, that each respondent's voice consequently speaks for many, and that it is important for those few soldiers who are invited to take advantage of this opportunity to have their voices heard by Army leadership in a fashion that protects confidentiality. This message was presented to all potential Army STARRS survey respondents both in a Study Information Sheet distributed prior to the informed consent session and in the informed consent session. The Army STARRS data collection team worked very closely with local Army Points of Contact to mount a campaign for survey participation while distributing Study Fact Brochures. They also emphasized the high-profile nature of Army STARRS and made it clear that survey results would be used at the highest levels of Army leadership. This recruitment message and the aggressive campaign mounted to disseminate this message might have encouraged both a high response rate and also encouraged soldiers with mental disorders to admit having these disorders, leading to the high reported rates of emotional problems among soldiers who agreed to administrative record linkage.

It is important to put the Army STARRS response rates in perspective by noting that these response rates are a good deal higher than those in a number of other major military surveys, including in surveys that offered complete anonymity to survey respondents (Bray *et al.*, 2006; Ryan *et al.*, 2007). As noted by Heeringa and colleagues in a companion paper in this issue (Heeringa *et al.*, 2013), these high Army STARRS response rates are due to higher proportions of pre-designated respondents in Army STARRS than previous surveys attending the consent sessions coupled with equally or higher proportions of those attending these sessions in Army STARRS than previous surveys agreeing to participate.

The high overall response rates in the Army STARRS surveys had an important implication for WT2, where we compared Army/DoD administrative record data in the population of all soldiers with those in the weighted (WT1) subset of soldiers who both completed the Army STARRS SAQ and provided administrative record linkage. This analysis failed to find evidence of significant differences between the weighted (WT1) sample and the population on a variety of administrative record variables. As a result, while it was important to weight the SAQ data for soldiers who consented to administrative data linkage, this was because failure to do so would have led to *over*-estimation rather than *under*-estimation of mental disorder prevalence in the de-identified survey data.

As we saw in the analysis of DEs, this over-representation of soldiers with mental disorders improved efficiency in estimating prevalence and correlates of these outcomes. Another important finding in this part of the analysis was that the distributions of the consolidated weights were fairly symmetrical and had a relatively narrow range. Taken together, these weight characteristics led to the finding, reported in Table 3, that DEs for the self-reported outcomes of central interest to the initiative are all quite modest.

One important limitation of the earlier analysis is that the weighting adjustments are based on the assumption that self-reports of mental disorders are as valid in the sample of respondents who provided de-identified SAQs as in the sample whose SAQ reports are completely anonymous. This need not be the case. The definitive evaluation of this issue would have required us to carry out an experiment in which a probability sub-sample of soldiers selected to participate in an Army STARRS survey were asked to provide completely anonymous survey data without the option to provide identifying information for administrative record linkage. We did not carry out that experiment. This means that even though prevalence estimates of the disorders assessed in the Army STARRS surveys are higher in the de-identified than anonymous SAQ sub-samples, it might still be the case that prevalence estimates would have been higher yet among respondents whose SAQs are not completely anonymous if they had never been asked to provide identifying information. There is no way to assess this possibility with the data available to us here, but it is a possibility that needs to be kept in mind when interpreting substantive results. To the extent that this bias exists, prevalence estimates of these disorders in the weighted Army STARRS survey data should be considered conservative.

Acknowledgments

On behalf of the Army STARRS Collaborators

Funding/Support

Army STARRS was sponsored by the Department of the Army and funded under cooperative agreement number U01MH087981 with the US Department of Health and Human Services, National Institutes of Health, National Institute of Mental Health (NIH/NIMH). The contents are solely the responsibility of the authors and do not necessarily represent the views of the Department of Health and Human Services, NIMH, the Department of the Army, or the Department of Defense.

Role of the Sponsors

As a cooperative agreement, scientists employed by NIMH (Colpe and Schoenbaum) and Army liaisons/consultants (COL Steven Cersovsky, MD, MPH USAPHC and Kenneth Cox, MD, MPH USAPHC) collaborated to develop the study protocol and data collection instruments, supervise data collection, plan and supervise data analyses, interpret results, and prepare reports. Although a draft of this manuscript was submitted to the Army and NIMH for review and comment prior to submission, this was with the understanding that comments would be no more than advisory.

Additional Contributions

The Army STARRS Team consists of Co-Principal Investigators: Robert J. Ursano, MD (Uniformed Services University of the Health Sciences) and Murray B. Stein, MD, MPH (University of California San Diego and VA San Diego Healthcare System); Site Principal Investigators: Steven Heeringa, PhD (University of Michigan) and Ronald C. Kessler, PhD (Harvard Medical School); NIMH collaborating scientists: Lisa J. Colpe, PhD, MPH and Michael Schoenbaum, PhD; Army liaisons/consultants: COL Steven Cersovsky, MD, MPH (USAPHC) and Kenneth Cox, MD, MPH (USAPHC). Other team members: Pablo A. Aliaga, MA (Uniformed Services University of the Health Sciences); COL David M. Benedek, MD (Uniformed Services University of the Health Sciences); Susan Borja, PhD (National Institute of Mental Health); Gregory G. Brown, PhD (University of California San Diego); Laura Campbell-Sills, PhD (University of California San Diego); Catherine Dempsey, PhD, MPH (Uniformed Services University of the Health Sciences); Richard Frank, PhD (Harvard Medical School); Carol S. Fullerton, PhD (Uniformed Services University of the Health Sciences); Nancy Gebler, MA (University of Michigan); Joel Gelernter, MD (Yale University); Robert K. Gifford, PhD (Uniformed Services University of the Health Sciences); Stephen E. Gilman, ScD (Harvard School of Public Health); Marjan G. Holloway, PhD (Uniformed Services University of the Health Sciences); Paul E. Hurwitz, MPH (Uniformed Services University of the Health Sciences); Sonia Jain, PhD (University of California San Diego); Tzu-Cheg Kao, PhD (Uniformed Services University of the Health Sciences); Karestan C. Koenen, PhD (Columbia University); Lisa Lewandowski-Romps, PhD (University of Michigan); Holly Herberman Mash, PhD (Uniformed Services University of the Health Sciences); James E. McCarroll, PhD, MPH (Uniformed Services University of the Health Sciences); Katie A. McLaughlin, PhD (Harvard Medical School); James A. Naifeh, PhD (Uniformed Services University of the Health Sciences); Matthew K. Nock, PhD (Harvard University); Rema Raman, PhD (University of California San Diego); Nancy A. Sampson, BA (Harvard Medical School); LCDR Patcho Santiago, MD, MPH (Uniformed Services University of the Health Sciences); Michaelle Scanlon, MBA (National Institute of Mental Health); Jordan Smoller, MD, ScD (Harvard Medical School); Nadia Solovieff, PhD (Harvard Medical School); Michael L. Thomas, PhD (University of California San Diego); Christina Wassel, PhD (University of Pittsburgh); and Alan M. Zaslavsky, PhD (Harvard Medical School). The authors would also like to thank John Mann, Maria Oquendo, Barbara Stanley, Kelly Posner, and John Keilp for their contributions to the early stages of Army STARRS development.

Additional Information

A complete list of Army STARRS publications can be found at http://www.ARMYSTARRS.org.

Declaration of interest statement

In the past five years Kessler has been a consultant for Eli Lilly & Company, Glaxo, Inc., Integrated Benefits Institute, Ortho-McNeil Janssen Scientific Affairs, Pfizer Inc., Sanofi-Aventis Groupe, Shire US Inc., and Transcept Pharmaceuticals Inc. and has served on advisory boards for Johnson & Johnson. Kessler has had research support for his epidemiological studies over this time period from Eli Lilly & Company, EPI-Q, GlaxoSmithKline, Ortho-McNeil Janssen Scientific Affairs, Sanofi-Aventis Groupe, Shire US, Inc., and Walgreens Co. Kessler owns a 25% share in DataStat, Inc. Stein has in the last three years been a consultant for Healthcare Management Technologies and had research support for pharmacological imaging studies from Janssen. The remaining authors report no competing interests.

References

Begin G., Boivin M., Bellerose J. (1979) Sensitive data collection through the random response technique: some improvements. *Journal of Psychology*, **101**(1), 53–65.

Bliese P.D., Thomas J.L., McGurk D., McBride S., Castro C.A. (2011) Mental health advisory teams: a proactive examination of mental health during combat deployments. *International Review of Psychiatry*, **23**(2), 127–134, DOI: 10.3109/09540261.2011.558834.

Bray R.M., Hourani L.L., Olmsted K.L.R., Witt M., Brown J.M., Pemberton M.R., Marsden M.E.,

Marriott B., Scheffler S., Vandermaas-Peeler R., Weimer B., Calvin S., Bradshaw M., Close K., Hayden D. (2006) 2005 Department of Defense Survey of Health Related Behaviors Among Active Duty Military Personnel: A Component of the Defense Lifestyle Assessment Program (DLAP), Research Triangle Park, NC, Research Triangle Institute.

Brink T.L. (1995) Sexual behavior and telling the truth on questionnaires. *Psychological Reports*, **76**(1), 218.

Campbell M.J., Waters W.E. (1990) Does anonymity increase response rate in postal questionnaire surveys about sensitive subjects? A randomised trial. *Journal of Epidemiology and Community Health*, **44**(1), 75–76.

Couper M.P., Singer E., Conrad F.G., Groves R.M. (2008) Risk of disclosure, perceptions of risk, and concerns about privacy and confidentiality as factors in survey participation. *Journal of Official Statistics*, **24**(2), 255–275.

Couper M.P., Singer E., Conrad F.G., Groves R.M. (2010) Experimental studies of disclosure risk, disclosure harm, topic sensitivity, and survey participation. *Journal of Official Statistics*, **26**(2), 287–300.

Edwards P.J., Roberts I., Clarke M.J., Diguiseppi C., Wentz R., Kwan I., Cooper R., Felix L. M., Pratap S. (2009) Methods to increase response to postal and electronic questionnaires. *Cochrane Database of Systematic Reviews*, **8**(3), MR000008, DOI: 10.1002/14651858.MR000008.pub4.

Fidler D.S., Kleinknecht R.E. (1977) Random responding versus direct questioning: two data-collection methods for sensitive information. *Psychological Bulletin*, **84**(5), 1045–1049.

Gadermann A.M., Engel C.C., Naifeh J.A., Nock M.K., Petukhova M., Santiago P.N., Wu B., Zaslavsky A.M., Kessler R.C. (2012) Prevalence of DSM-IV major depression among U.S. military personnel: meta-analysis and simulation. *Military Medicine*, **177**(8 Suppl), 47–59.

Heeringa S.G., Colpe L.J., Fullerton C.S., Gebler N., Naifeh J.A., Nock M.K., Sampson N.A., Schoenbaum M., Zaslavsky A.M., Stein M.B., Ursano R.J., Kessler R.C. (2013) Field procedures in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *International Journal of Methods and Psychiatric Research*, **22**(4), 276–287.

Heeringa S.G., West B.T., Berglund P.A. (2010) Applied Survey Data Analysis. Boca Raton, FL: Taylor and Francis.

Kessler R.C., Colpe L.J., Fullerton C.S., Gebler N., Naifeh J.A., Nock M.K., Sampson N.A., Schoenbaum M., Zaslavsky A.M., Stein M.B., Ursano R.J., Heeringa S.G. (2013) Design of the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *International Journal of Methods in Psychiatric Research*, **22**(4), 267–275.

Kish L. (1965) Survey Sampling. New York: John Wiley and Sons.

Kish L. (1976) Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society, Series A*, **139**(1), 80–95.

Kish L., Frankel M.R. (1974) Inference from complex samples. *Journal of the Royal Statistical Society, Series A*, **36**(1), 1–37.

Little R.J.A., Vartivarian S. (2005) Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, **31**(2), 161–168.

Ong A.D., Weiss D.J. (2000) The impact of anonymity on responses to "sensitive" questions. *Journal of Applied Social Psychology*, **30**(8), 1691–1708, DOI: 10.1111/j.1559-1816.2000.tb02462.x.

Rogers S.M., Miller H.G., Turner C.F. (1998) Effects of interview mode on bias in survey measurements of drug use: do respondent characteristics make a difference? *Substance Use and Misuse*, **33**(10), 2179–2200, DOI: 10.3109/10826089809069820.

Ryan M.A., Smith T.C., Smith B., Amoroso P., Boyko E.J., Gray G.C., Gackstetter G.D., Riddle J.R., Wells T.S., Gumbs G., Corbeil T.E., Hooper T.I. (2007) Millennium Cohort: enrollment begins a 21-year contribution to understanding the impact of military service. *Journal of Clinical Epidemiology*, **60**(2), 181–191, 10.1016/j.jclinepi.2006.05.009.

Schafer J.L. (1999) Multiple imputation: a primer. *Statistical Methods in Medical Research*, **8**(1), 3–15, DOI: 10.1177/096228029900800102.

Turner C.F., Ku L., Rogers S.M., Lindberg L.D., Pleck J.H., Sonenstein F.L. (1998) Adolescent sexual behavior, drug use, and violence: increased reporting with computer survey technology. *Science*, **280**(5365), 867–873, DOI:10.1126/science.280.5365.867.

Ursano R.J., Heeringa S., Stein M.B., Kessler R.C. (submitted for publication) The Army Study to Assess Risk and Resilience in Servicemembers (STARRS).

Warner C.H., Appenzeller G.N., Grieger T., Belenkiy S., Breitbach J., Parker J., Warner C. M., Hoge C. (2011) Importance of anonymity to encourage honest reporting in mental health screening after combat deployment. *Archives of General Psychiatry*, **68**(10), 1065–1071, DOI: 10.1001/archgenpsychiatry.2011.112.

Warner C.H., Appenzeller G.N., Mullen K., Warner C.M., Grieger T. (2008) Soldier attitudes toward mental health screening and seeking care upon return from combat. *Military Medicine*, **173**(6), 563–569.

Warner C.H., Breitbach J.E., Appenzeller G.N., Yates V., Grieger T., Webster W.G. (2007) Division mental health in the new brigade combat team structure: part II. Redeployment and postdeployment. *Military Medicine*, **172**(9), 912–917.

Werch C.E. (1990) Two procedures to reduce response bias in reports of alcohol consumption. *Journal of Studies on Alcohol*, **51**(4), 327–330.

Wolter K.M. (1985) Introduction to Variance Estimation. New York: Springer-Verlag.

Zaslavsky A.M., Schenker N., Belin T.R. (2001) Downweighting influential clusters in surveys: application to the 1990 Post Enumeration Survey. *Journal of the American Statistical Association*, **96**(455), 858–869, DOI: 10.1198/016214501753208889.