Interviewer Effects on Epidemiologic Diagnoses of Posttraumatic Stress Disorder


In an epidemiologic study of 641 interviewed subjects in the Australian Vietnam Veterans Health Study, three diagnoses of Vietnam combat-related posttraumatic stress disorder (PTSD) were obtained: lifetime prevalence using a variant of the Diagnostic Interview Schedule and lifetime and current (1-month) PTSD prevalence using the Standardized Clinical Interview for the Diagnostic and Statistical Manual of Mental Disorders, Third Revision. Prevalence estimates using the Standardized Clinical Interview varied according to interviewer characteristics (female vs. male, clinician vs. nonclinician) but not for the Diagnostic Interview Schedule. The authors use a simple variant of logistic regression to distill estimates of two informative parameters characterizing interviewers' judgments: severity threshold (related to the individual interviewer's criterion of "case-ness") and reliability (related to degree of classification error of the individual interviewers). Examination of these estimates shows that female clinicians adopted lower severity thresholds for diagnosis of PTSD than other interviewers and hence had higher prevalence estimates while being relatively reliable in their judgments. Examination also shows that nonclinician interviewers can perform at least as reliably as clinicians. The Diagnostic Interview Schedule measure of PTSD was not moderated by these interviewer aspects. This use of threshold and reliability parameters is offered for routine use in epidemiologic field studies to examine potential interviewer effects.


One of the prime determinants of prevalence in epidemiologic field studies is the interviewer. Reduction of discrepancy among interviewers has been recognized as an important goal, with modern large scale studies (1–3) often incorporating interview reliability studies and regularly reporting measures of agreement among interviewers, particularly between clinicians and lay interviewers. Recent developments in the measurement of psychiatric status have ushered in what Bruce Dohrenwend (4) has termed the "third generation" of studies of the incidence and prevalence of psychiatric disorder. The development of standardized interviews, beginning with the Present State Examination (5) and progressing through the Diagnostic Interview Schedule (DIS) (6), the Structured Clinical Interview for the Diagnostic and Statistical Manual of Mental Disorders, Third Revision (SCID) (7), the Composite International Diagnostic Interview (8), and the Schedules for Clinical Assessment in Neuropsychiatry (9), has occurred as a response to the search for methods to reduce the biases that arise from the use of different settings (clinic or community) and different types of interviewers (psychiatrists, other clinicians, and nonclinician interviewers).

Mostly in such studies, individual diagnosis and ratings by a trained clinician remain the "gold standard" with which other interviewers are compared. The same series of subjects must be passed before both interviewer and gold standard clinician, the latter usually being an expensive additional process.

In any inquiry leading to diagnosis that relies on symptomatology, research subjects must be asked whether they have experienced the relevant symptoms. This can be accomplished in a self-administered for-
In this paper, we deal with interviewer effects on PTSD diagnoses made from the Diagnostic and Statistical Manual of Mental Disorders, Third Revision. Instead, each interviewer's caseness judgments are evaluated against some dimensional criterion measure related to the illness in question (risk factor, comorbidity, symptom severity), and the interviewer is characterized by his or her diagnosing behavior in relation to this "internal" dimensional criterion. This approach may have several advantages over others: It does not require data from an independent gold standard source (typically clinician judgment, itself subject to problems and uncertainty); it can be used with any correlated dimensional variable (frequently already available elsewhere in an epidemiologic dataset); and it provides some insight into the nature of interviewer differences.

MATERIALS AND METHODS

Materials

The data examined here are derived from the Australian Vietnam Veterans Health Study (11–13). A random sample of 1,000 Vietnamese War veterans was drawn from Australian Army lists of military postings to Vietnam during the involvement of Australia in that conflict (1962–1972). These veterans were located, sought out, and interviewed nationally during the period July 1990 through February 1993. Completed interviews were obtained from 641 of this sample, but Army data of the era were also taken from the Central Army Records Office on all 1,000 subjects. A personal interview was administered to each of the participating veterans. It is important to note the sequence in which interview data were collected. The interviews began with an initial, standardized (physical) health status interview followed by a self-completion battery of measures of psychological health. This battery included the 28-item General Health Questionnaire (GHQ) (14), a self-reported measure of psychiatric morbidity, and the Mississippi Scale for Combat-related Posttraumatic Stress Disor-
Interviewer Effects and Posttraumatic Stress Diagnosis

In figure 1, these concepts of diagnostic reliability and diagnostic severity threshold in relation to a dimensional variable correlated with the illness of inter-

Am J Epidemiol Vol. 144, No. 6, 1996
FIGURE 1. Diagnostic reliability and diagnostic severity in relation to some dimensional variable correlated with a particular illness. Prob, relation of the dichotomous criterion of caseness given continuous morbidity score. PDF, population distribution of a fictitious continuous measure of morbidity.

RESULTS
Establishing interviewer effects

The prevalence of lifetime combat-related PTSD according to the DIS was 11.7 percent. The prevalence est are represented. A more thorough discussion of dimensional and categorical alternative views of illness is presented elsewhere (17). See (18) for a more detailed description of these concepts in relation to a latent dimension of morbidity. In figure 1A, subjects with morbidity scores of 13 or less have zero probability of being judged as a case (Y axis), whereas those with scores of 19 or more are almost certain to be judged as cases. Subjects with morbidity scores of 16 have 50 percent chance of being judged as either a case or a noncase. (Conventionally, the point at which 50 percent of subjects are classified as "cases" and 50 percent as "noncases" is the threshold for judgment of caseness, and the score at this point represents the severity threshold.) Because only about 10 percent of subjects in the figure 1 example score 16 or more on morbidity, this diagnosis may be called "severe." It might also be referred to as a "reliable" diagnosis in the sense that subjects outside a tight region (which is centered on the severity level of 16) are almost certain to be appropriately diagnosed according to the following scores: less than 13, always noncases; more than 19, always cases. The diagnosis represented by the unbroken line in figure 1B takes the same severity threshold level as in figure 1A. However, it is an "unreliable" diagnosis in that subjects who differ somewhat significantly from this severity level of 16, either above or below, still have substantial chances of being wrongly classified. Note that the prevalence reported with the diagnosis in figure 1B would be higher than that of the diagnosis in figure 1A, despite the fact that both are operating at the same threshold severity level of morbidity. Finally, in figure 1C, a diagnosis that is unreliable yet operating at a mild level can be seen; severity levels of 13 (the population average in this example) have caseness probability of 0.5.

Using the approach described above, each interviewer's performance on each PTSD diagnosis is characterized below. These two concepts of diagnostic reliability and diagnostic severity are more informative aspects of epidemiologic diagnoses than the simpler (confounded) measure of prevalence. Similarly, these two parameters appear to be more informative than the usual single parameters that 1) index only discordance between raters, like Cohen's kappa (19); 2) provide little or no insight into which, if either, interviewer is better and why; and 3) require that at least some subjects undergo interviews twice.
of the same illness according to the AUSCID was 20.7 percent, whereas current PTSD by the AUSCID yielded 11.4 percent. In other words, the SCID interview gave rise to almost twice as many (comparable) lifetime cases as the DIS, which was administered soon afterward. Recent studies of US veterans show the same pattern of SCID and DIS PTSD prevalence rates (2, 3). To investigate whether this difference can be accounted for by different behavior among the interviewers (with some being more willing than others to accept symptom reports on the AUSCID as "real"), interviewers were classified by sex and by counseling experience. Of the 641 interviews, complete data were available on 638.

As can be seen in table 1, the highest AUSCID-lifetime rate (30.3 percent) is associated with female counsellors. This rate varies significantly (logistic regression $\chi^2$ with 3 df, $p = 0.002$) across the four interviewer groups. For the AUSCID-current diagnosis, counsellors (particularly female) seem to yield cases more readily than noncounsellors ($p = 0.002$), whereas the higher female rates for the DIS diagnoses did not achieve statistical significance. Prima facie, interviewer characteristics appear to influence AUSCID but not DIS prevalences.

It is possible that each interviewer group saw a different series of veterans, and thus the interviewer effect seen in table 1 could be attributable to confounding by subject characteristics. As can be seen from the bottom section of table 1, the averages of some related scores of the veterans did differ significantly across the four groups. The age at which the veteran first served in Vietnam significantly differs across the four groups ($p = 0.006$), with more at-risk (younger) veterans on average in the female counsellor group—the group with highest case rates. The traumatic response score also differed significantly, albeit the pattern of results did not relate to the case rates in a clear manner. The Mississippi and combat scores, although not differing significantly, did yield highest averages for female counsellors.

Logistic regressions for each diagnosis were fitted to investigate whether the interviewer effects remained when the case rates in each group were "adjusted" for such differences. The initial models had 8 degrees of freedom, 3 for the four interviewer groups and 5 for the covariates. After computing these models for each of the three diagnoses, no model yielded a significant partial coefficient for age in Vietnam or for the GHQ scale, so these were dropped and the models were recomputed with only three covariates. The results from these final models are presented in table 2.

As shown in table 2 for the AUSCID-lifetime diagnosis, all three covariates are highly significant with (partial) odds ratios in the predicted direction (greater than unity, indicating that an increase in the scale is associated with higher caseness). The three "dummy" variables coding the four interviewer groups were defined as follows: "noncounsellor" = 1 for noncounsellors, 0 for counsellors; "female" = 1 for females, 0 for males; and "fem-cns" = 1 for female counsellors, 0 otherwise. For the AUSCID-lifetime diagnosis, the only significant dummy variable was the one indexing female counsellors: an odds ratio of 3.367 is associ-

---

**TABLE 1. Interviewer effects on unadjusted case rates of posttraumatic stress disorder (PTSD) based on data from the Australian Vietnam Veterans Health Study**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Men</th>
<th>Women</th>
<th>Overall</th>
<th>$p$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Counselors (n = 133)</td>
<td>Noncounselors (n = 225)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Counselors (n = 178)</td>
<td>Noncounselors (n = 103)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUSCID*-lifetime</td>
<td>19.5</td>
<td>16.0</td>
<td>30.3</td>
<td>15.5</td>
</tr>
<tr>
<td>AUSCID-current*</td>
<td>13.5</td>
<td>6.7</td>
<td>18.0</td>
<td>7.8</td>
</tr>
<tr>
<td>DIS*-lifetime</td>
<td>10.6</td>
<td>8.1</td>
<td>14.9</td>
<td>14.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PTSD case rates (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age Vietnam*</td>
<td>24.5</td>
<td>23.4</td>
<td>23.2</td>
<td>25.1</td>
</tr>
<tr>
<td>GHQ-28*</td>
<td>4.3</td>
<td>4.1</td>
<td>4.0</td>
<td>3.7</td>
</tr>
<tr>
<td>Mississippi*</td>
<td>72.4</td>
<td>73.3</td>
<td>75.3</td>
<td>73.2</td>
</tr>
<tr>
<td>Trauma response*</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
<td>5.3</td>
</tr>
<tr>
<td>Combat*</td>
<td>4.9</td>
<td>5.1</td>
<td>5.6</td>
<td>5.1</td>
</tr>
<tr>
<td><strong>Covariate means</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* AUSCID, current authors' adaptation of the Standardized Clinical Interview for the Diagnostic and Statistical Manual of Mental Disorders, Third Revision, PTSD module; AUSCID-current, diagnosis within the past month; DIS, Diagnostic Interview Schedule; Age Vietnam, age of embarkation on first tour of Vietnam; GHQ-28, 28-item General Health Questionnaire; Mississippi, Mississippi Scale for Combat-related Posttraumatic Stress Disorder; trauma response, 16-item scale to measure the initial reaction to a worst traumatic event; combat, 21-item scale to measure combat experience in Vietnam.
TABLE 2. Interviewer effect* on adjusted case rates of posttraumatic stress disorder (PTSD) based on data from the Australian Vietnam Veterans Health Study

<table>
<thead>
<tr>
<th></th>
<th>AUSCID-*Lifetime</th>
<th>AUSCID-current*</th>
<th>DIS-*Lifetime</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OR*</td>
<td>p value</td>
<td>OR*</td>
</tr>
<tr>
<td>Mississippi*</td>
<td>1.067</td>
<td>0.000</td>
<td>1.090</td>
</tr>
<tr>
<td>Combat*</td>
<td>1.083</td>
<td>0.018</td>
<td>1.058</td>
</tr>
<tr>
<td>Trauma response*</td>
<td>1.235</td>
<td>0.000</td>
<td>1.151</td>
</tr>
<tr>
<td>Noncounsellor</td>
<td>0.601</td>
<td>0.166</td>
<td>0.185</td>
</tr>
<tr>
<td>Female</td>
<td>0.624</td>
<td>0.280</td>
<td>1.124</td>
</tr>
<tr>
<td>Fem-cns*</td>
<td>3.367</td>
<td>0.033</td>
<td>1.097</td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Counsellor</td>
<td>1.000</td>
<td></td>
<td>1.000</td>
</tr>
<tr>
<td>Noncounsellor</td>
<td>0.601</td>
<td></td>
<td>0.185</td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Counsellor</td>
<td>2.101</td>
<td></td>
<td>1.234</td>
</tr>
<tr>
<td>Noncounsellor</td>
<td>0.375</td>
<td></td>
<td>0.207</td>
</tr>
</tbody>
</table>

* AUSCID, current authors' adaptation of the Standardized Clinical Interview for the Diagnostic and Statistical Manual of Mental Disorders, Third Revision, PTSD module; AUSCID-current, diagnosis within the past month; DIS, Diagnostic Interview Schedule; OR, odds ratio; Mississippi, Mississippi Scale for Combat-related Posttraumatic Stress Disorder; combat, 21-item scale to measure combat experience in Vietnam; trauma response, 16-item scale to measure the initial reaction to a worst traumatic event; Fem-cns, female counsellor.

Ated with moving (only) from 0 (not both female and a counsellor) to 1 (female counsellor) on the fem-cns dummy variable. For AUSCID-current, only the variable noncounsellor was significant, whereas for the DIS-lifetime diagnosis, none of the interviewer-group dummy variables was significant.

Because each of the three interviewer group dummy variables is zero for the male counsellor, we can reconstruct odds ratios (relative to this group) for each of the other three gender-profession groups by multiplying the corresponding dummy variable odds ratio in the top panel (significant or not). For instance, male counsellors have scores 0, 0, and 0 on the dummy variables and female counsellors have scores 0, 1, and 1 on these dummy variables; so the product \((0 \times 0.601) \times (1 \times 0.624) \times (1 \times 3.367) = 2.101\) yields the odds ratio for SCID-lifetime caseness associated with being interviewed by a female rather than a male counsellor. In the second panel, these reconstructed model-predicted odds ratios for the four interviewer groups are shown. Note that for subjects with the same values on the covariates, the model predicts AUSCID-lifetime case rates in the same rank order as in table 1—highest for female counsellors, next highest for male counsellors, and male and female noncounsellors with the lowest case rate. The same statement can be made about the AUSCID-current diagnosis, but not about the DIS-lifetime diagnosis.

The significance of the interviewer group classification when the veterans are "matched" on the covariates can be tested by examining the 3-df deviance tests contrasting the full models (three covariates plus three interviewer group dummy variables) with reduced models omitting the interviewer group dummy variables. This 3-df deviance test for AUSCID-lifetime was highly significant \((\chi^2 = 23.55, p = 0.000)\), showing that the interviewer groups should not be ignored in modeling case rates on this diagnosis. The same can be said for AUSCID-current \((\chi^2 = 24.87, p = 0.000)\); but we increase deviance only within the bounds of chance \((\chi^2 = 5.15, p = 0.161)\) when we omit these effects with the DIS-lifetime diagnosis. The reduced models had correct classification rates of 87.1, 92.5, and 90.9 percent, indicating that the fitted covariates were “controlling” for veteran differences adequately.

In summary, even controlling for important covariates, the interviewer effects remain for both AUSCID diagnoses, with counsellors (particularly female counsellors) diagnosing more cases, even among "identical" (adjusted) series of veterans. No such effect emerges with the more structured DIS-lifetime diagnosis. At this time, it can be asserted more confidently that the interviewer effects on the AUSCID diagnoses of PTSD are genuine. But to what extent are they due to differences among interviewers in severity thresholds and to what extent, differences in reliability?

The data in table 3 characterize reliability and severity for each of the PTSD diagnoses for each of the four interviewer groups. To obtain these indices, a simple 1-df logistic regression model was fitted to the diagnostic data using data from all veterans:

\[
\log \left[ \frac{p}{1 - p} \right] = a + bX \quad \{\text{or odds} = \frac{p}{1 - p} = \exp(a)\exp(bX)\},
\]
where $X$ represents a single dimensional criterion morbidity measure that we would expect to be strongly related to the PTSD diagnoses—either the Mississippi, the GHQ, or the traumatic response scale. The same data were then fitted with a 7-df logistic regression model, corresponding to fitting different $a + bX$ models within each interviewer group (three group intercept parameters plus four morbidity ($X$) subvariables, one for each group, taking values of $X$ for veterans in that group and 0 for veterans not in that group). Comparison of these two models (deviance test with 6 df) tested whether separate regressions were required within each interviewer group: the $p$ value for the AUSCID-lifetime-Mississippi combination was 0.057; for the other five AUSCID-diagnosis $X$ morbidity combinations, these deviance test $p$ values were all less than 0.005, justifying within-group regressions. For the DIS-lifetime-$X$ combinations, no test achieved significance at 0.05, indicating that separate interviewer-group regressions were not required for this diagnosis.

If we restate the linear regressions from such a model as

$$[a] + [b]X = r(X - s) = [-sr] + [r]X,$$

whence

$$a = -sr \text{ and } b = r,$$

then $r$ and $s$ represent the reliability and severity, in precisely the sense discussed earlier, of the diagnosis being modeled. The parameter $s$ is the value of the morbidity variable $X$ at which $p = 0.5$ and so represents severity threshold for case judgment, and the steepness (maximized at the point where $X = s$) of the logistic cumulative function is equal to $r/4$ (obtained by differentiating the cumulative logistic function and maximizing this derivative). If the diagnosis is oriented or scored to ensure a positive relation with $X$, then $r$ will range between 0 and infinity; and the bigger the value of $r$, the greater the reliability. When $r$ is 0, the curves in figure 1 are flat and horizontal, indicating that the diagnosis is totally unrelated to the dimensional criterion $X$. When $r$ is infinite, the curves become step-functions indicating perfect accord—$X$ below $s$ corresponding exactly to noncases, and $X$ above $s$, to cases. (As logistic regression is maximum likelihood, this 1-1 reparameterization of the estimates of $a$ and $b$ into estimates of $r$ and $s$ preserves maximum likelihood estimation properties.)

In table 3, $s$ and $r$ parameter estimates within interviewer groups on each dimensional criterion can be seen. However, recall that for DIS-lifetime diagnoses, we do not have statistical justification to trust such interview group differences. Taking each interviewer group in turn and restricting attention to the AUSCID diagnoses, we can deduce from table 3 that male counsellors were the least reliable interviewers, having the minimum values for $r$ on four of the six diagnosis-$X$ morbidity combinations, and with almost the worst on the remaining two combinations. Their thresholds were intermediate. The male noncounsellors appeared to be very reliable, particularly when the Mississippi and traumatic response scales were con-
sidered as the dimensional criteria. Their threshold for caseness appeared to be relatively high, although they were not the most demanding interviewers in this sense. The female counsellors showed reasonable reliabilities, particularly for the Mississippi scale. Perhaps the most dramatic finding was their uniformly low severity (threshold) criteria: In all six combinations, they were diagnosing at the mildest levels. The female noncounsellors appeared to have low reliabilities with the Mississippi scale and otherwise, moderate. Their thresholds were the most extreme in five of the six combinations, especially the Mississippi. Averaging these indices across validators and AUSCID diagnoses produces data that are shown in table 4 and that are summarized in the following three major observations: 1) Female noncounsellors assigned caseness at lower levels of morbidity (while doing so with reasonable reliability); 2) male counsellors were the least reliable interviewers; and 3) noncounsellors performed as reliably as the counsellor clinicians.

Although the focus in the present paper is on interviewer differences, this approach applies to any diagnosis-dimensional-criterion combination, and we can also compare the DIS and SCID diagnoses using \( s \) and \( r \) estimates. Comparing the SCID-lifetime and DIS-lifetime data in table 3 reveals that the relatively high prevalences obtained with the SCID are predominantly due to the lower thresholds that it has (either because interviewers adopt lower thresholds with the SCID or because interviewees are more ready to admit symptomatology in the context of the SCID). It is important that the lower case rates obtained with the DIS do not appear to be associated with any marked increase in reliability. If we average these lifetime diagnoses over dimensional criteria and interviewer groups, these comments are summarized as follows: SCID-lifetime—\( r = 0.183, s = 41.41 \); DIS-lifetime—\( r = 0.186, s = 46.74 \).

**DISCUSSION**

Our results show that different diagnostic criteria appear to be used by different interviewers with the AUSCID instrument, despite clinical experience and despite attempts to standardize criteria by training. The possibility that the interviewers are behaving consistently and that the effects are due to the veterans presenting differently as a result of different levels of interviewer-veteran rapport is unambiguously ruled out by the following observations. First, the effects are not present on the DIS-lifetime diagnosis, which was based on data collected at the close of the interview after any such change in veteran presentation would be expected to have occurred. Second, the effects were present on the trauma response scale, which immediately followed the relaxation of more “rigid” interview structure (self-report or standardized interview) and immediately preceded the gathering of the AUSCID symptom data on which the affected diagnoses are algorithmically based. In other words, if the veterans had “changed” (and the interviewers were consistent), then we would expect female counsellors, in particular, to have had subjects with higher trauma response scores and to have had similar severity thresholds to the other interviewer groups, albeit on the artifactually elevated trauma response data of their veterans. It can be seen from tables 1 and 3 that neither of these propositions hold.

Epidemiologic diagnostic instruments like the DIS attenuate if not remove such effects. However, in instruments like the AUSCID (and its parent, the SCID), where the required amount of symptom-level clinical interpretation is larger, such effects more readily manifest. It must be emphasized that it cannot be said from such data that (female) counsellors were “overdiagnosing”; it may be that others “underdiagnosed.” Similarly, readers should be cautioned against generalizing about either interviewer gender or profession. Many of the interviews were completed by interviewers who accumulated 60 or more interviews in total. Such individual interviewers, while of necessity being male or female, counsellor or noncounsellor, are bound to have had large individual effects on the analyses reported herein.

The data reported here show that interviewer-derived prevalence estimates may be affected strongly by interviewer characteristics. This is not a new finding, yet such phenomena as discussed here are seldom reported. In cases in which interviewer error is suspected (either threshold or diagnostic reliability), we recommend including (if not already included) morbidity-related measures that are continuous. We also recommend examining the dynamics of interviewer differences in terms of diagnostic reliability and severity threshold, parameters directly providing insight (beyond that provided by the simple observation of discordance) that may be of use in addressing and removing such unwanted influences from future epidemiologic studies.
ACKNOWLEDGMENTS

This research was supported by the National Health & Medical Research Council of Australia (NH&MRC), the Public Health Research & Development Committee of the NH&MRC, the Australian Army, the Australian Vietnam War Veterans’ Trust Limited, the Australian Department of Veteran Affairs, the Vietnam Veterans’ Counselling Service, the Westmead Research Institute, the Australian War Memorial, and a wide variety of ex-service and veteran unit organizations.

REFERENCES