In this issue:

RATIONALE AND DEVELOPMENT OF THE CLINICIAN-ADMINISTERED PTSD SCALES

Dudley David Blake, PhD[1]
National Center for PTSD and
Palo Alto VAMC/Menlo Park Division

Several structured interviews are now available for assessing and diagnosing PTSD (for a review see Blake et al., in press). These interviews provide a critical element of the multi-component assessment of PTSD (Lyons et al., 1989; Wolfe et al., 1987). Among the available interviews are the Clinician Administered PTSD Scales (CAPS-1 and CAPS-2; Blake et al., 1990; Nagy et al., 1991), which were developed at the National Center for PTSD (NC-PTSD) to accommodate both research and clinical needs. This report describes the rationale behind and the development of the CAPS interviews.

The idea for a structured PTSD interview was introduced at the first meeting of the National Center in Baltimore, Maryland, in August, 1989. Dr. Dennis Charney, Director of the Clinical Neurosciences Division of the NC-PTSD, proposed that staff of the National Center develop a PTSD interview akin to the Hamilton Depression Scale (Hamilton, 1967). Recommendations by Dr. Terence Keane, Director of the Behavioral Science Division of the NC-PTSD, significantly shaped the eventual form of the interview. The first recommendation was to include a *range* of rating options per symptom, rather than simply determining symptom presence or absence. This convention effectively allowed the interview to be used as a continuous as well as a dichotomous measure of PTSD. The other available interviews typically provide limited, dichotomous information about PTSD symptom severity; a range of rating options allows for greater differentiation at both the symptom and disorder levels. A second recommendation involved the delineation of the frequency and intensity of each symptom, rather than making a gross determination of severity. Separating out frequency and intensity adds a level of measurement detail not ordinarily considered in PTSD assessment.

With these recommendations in mind, the first working drafts of the CAPS-1 and CAPS-2 were printed in February, 1990. Over the next 9 months, the interviews were field-tested and revised three times. The current versions of the CAPS and an accompanying Instruction Manual were released in October, 1990[2]. They were developed for clinicians and researchers who have a working knowledge of PTSD. Questions assess the 17 DSM-III-R symptoms of PTSD, as well as 8 PTSD-associated symptoms (derived from the DSM-III-R and the clinical research literature on PTSD). Items are also provided for rating: the impact of symptoms on social and occupational functioning; the status of PTSD symptoms relative to an earlier assessment or 6 months prior to the current assessment; estimated validity of the overall CAPS assessment; and overall PTSD severity. The CAPS interviews are compatible with DSM-IV.

The CAPS-1 was designed as a general diagnostic instrument by which determinations of current and lifetime PTSD status could be made. Preliminary data on the CAPS-1 suggest that it is a reliable and valid measure of PTSD symptomatology (see following article). The CAPS-2 was designed to assess PTSD symptoms during the previous week (instead of the previous month) and is primarily for use in repeated assessments over relatively brief assessment intervals. Accordingly, this interview has been employed successfully in evaluating treatment outcome (e.g., Boudewyns et al., 1993; Nagy et al., 1993).

Several features of the CAPS interviews set them apart from the other PTSD interviews. First, the *frequency* and *intensity* of each symptom are rated. The division of symptom severity into these two dimensions allows for a more fine-grained analysis of PTSD phenomenology. For example, one patient may experience frequent but moderately intense PTSD symptoms, whereas another may experience relatively infrequent but extremely intense symptoms. The frequency-intensity feature may prove to be especially important in tracking the course of PTSD symptoms across time, for example in outcome and follow-up studies in which the frequency or the intensity of symptoms can change differentially over time. The *presence* of each symptom (i.e., the dichotomous presence/absence judgment) is determined by weighing jointly the frequency and intensity ratings. One rationally derived convention for establishing endorsement is to count a symptom only when frequency is rated as a "1" (occurred at least once during the designated time frame) or greater and intensity is rated as a "2" (at least moderately intense or distressing) or greater. A more conservative strategy is to consider endorsement only if the frequency and intensity ratings sum to 4 or more. Recently, Weathers (1993) demonstrated that both of these rationally derived decision rules may overestimate PTSD status (see following article by Weathers & Litz).

A second CAPS feature is that all ratings are made on separate 5-point Likert scales. This convention results in finer gradations in the frequency and intensity of PTSD symptoms, as well as in the disorder as a whole. Not only can the CAPS be used for determining symptom or disorder presence or absence (see above); a range of rating options allows for finer discriminations to be made about PTSD severity. The total score for the CAPS PTSD ratings (frequency + intensity) can range from 0 to 136. Behaviorally anchored rating options for both frequency and intensity are provided to aid in making these symptom ratings.

Third, standard prompt questions are provided and follow-up questions are suggested for each item. The CAPS-1 provides standard prompt questions regarding both the frequency and the intensity of individual symptoms. This feature provides an explicit guide for the interviewer, who first asks these standard questions, and issues follow-up questions as needed. Guidelines for follow-up questioning are outlined in an accompanying CAPS Instruction Manual. Due in large part to the standard questions and behaviorally anchored rating options, the CAPS can be administered reliably and with less concern about error variance due to rater subjectivity. Nonetheless, all CAPS interviewers should undergo formal training and should: (a) have had previous experience with diagnostic interviews; (b) have a working familiarity with PTSD and associated symptoms; (c) observe actual CAPS interviews by experienced clinicians; and (d) practice using the interview in a role-play or mock-interview situation.

Since their introduction to the field, the CAPS interviews have gained wide acceptance in PTSD assessment. Work employing the CAPS in other countries (e.g., Lillywhite & Neal, 1993; Hovens et al., 1991) bears witness to its use internationally. It is hoped that the CAPS will set a high standard for PTSD assessment and provide an important means for advancing our understanding of PTSD.

REFERENCES

BLAKE, D.D., WEATHERS, F.W., NAGY, L.M., KALOUPEK, D.G., KLAUMINZER, G., CHARNEY, D.S. & KEANE, T.M. (1990). **A clinician rating scale for assessing current and lifetime PTSD: The CAPS-1.** *Behavior Therapist, 13,* 187-188.

BLAKE, D.D., WEATHERS, F.W., NAGY, L.M., KALOUPEK, D.G., GUSMAN, F., CHARNEY, D.S. & KEANE, T.M. (in press). **The development of a clinician-administered PTSD scale.** *Journal of Traumatic Stress.*

BOUDEWYNS, P.A., STWERTKA, S.A., HYER, L.A., ALBRECHT, J.W. & SPERR, E.V. (1993). **Eye movement desensitization for PTSD of combat: A treatment outcome pilot study.** *Behavior Therapist, 16,* 29-33.

HAMILTON, M. (1967). **Development of a rating scale for primary depressive illness.** *British Journal of Social and Clinical Psychology, 6,* 278-296.

HOVENS, J.E., SCHREUDER, B.J.N., RIVERO, V.V., VAN DER PLOEG, H.M., BRAMSEN, I. & KLAARENBEEK, M.T.A. (1991). **Assessment of PTSD with the Dutch CAPS and self-rating scales.** Paper presented at the 7th annual conference of the International Society for Traumatic Stress Studies, Washington, D.C.

LILLYWHITE, A.R. & NEAL, L.A. (1993). **The importance of severity as a factor in post-traumatic stress disorder.** *British Journal of Psychiatry, 163,* 837.

LYONS, J.A., GERARDI, R.J., WOLFE, J. & KEANE, T.M. (1988). **Multidimensional assessment of combat-related PTSD: Phenomenological, psychometric, and psychophysiological considerations.** *Journal of Traumatic Stress, 1,* 373-394.

NAGY, L.M., BLAKE, D.D., DAN, E., RINEY, S., MANGINE, W., SOUTHWICK, S.M., GUSMAN, F. & CHARNEY, D.S. (October, 1991). **Clinician Administered PTSD Scale - Weekly Version (CAPS-2): Reliability, validity, and sensitivity to change.** In D.D. Blake (Chair), *An update on the Clinician Administered PTSD Scales (CAPS-1 and CAPS-2).* Symposium paper presented at the 7th annual conference of the International Society for Traumatic Stress Studies, Washington, D.C.

NAGY, L.M., MORGAN, C.A., SOUTHWICK, S.M. & CHARNEY, D.S. (1993). **Open prospective trial of fluoxetine for posttraumatic stress disorder.** *Journal of Clinical Psychopharmacology, 13,* 107-113.

WEATHERS, F.W. (1993). **Rational and empirical scoring rules for the Clinician-Administered PTSD Scale.** Unpublished manuscript.

WOLFE, J., KEANE, T.M., LYONS, J.A. & GERARDI, R.J. (1987). **Current trends and issues in the assessment of combat-related post traumatic stress disorder.** *Behavior Therapist, 10,* 27-32.

PSYCHOMETRIC PROPERTIES OF THE CLINICIAN-ADMINISTERED PTSD SCALE, CAPS-1

Frank W. Weathers, PhD[1] and Brett T. Litz, PhD[1]
National Center for PTSD and Boston VAMC
Tufts University School of Medicine

The lifetime and current diagnostic form of the Clinician-Administered PTSD Scale (CAPS-1) is a structured interview for PTSD recently developed at the National Center for PTSD (see Blake, this issue). From the outset the primary goal was to create a comprehensive, psychometrically sound PTSD interview that could serve as a "gold standard" for the field of traumatic stress. We recognized that such an interview would have to combine rigor with clinical sensitivity in order to be useful in a wide variety of clinical and research contexts. The CAPS-1 has a number of features designed to address some of the limitations of other PTSD interviews. First, the CAPS-1 assesses the 17 DSM-III-R (and now DSM-IV) symptoms of PTSD, as well as 8 associated symptoms. It also includes five items assessing response validity, overall severity of symptoms, overall improvement since a prior assessment, and the impact of symptoms on social and occupational functioning. Second, the frequency and intensity of each symptom on the CAPS-1 are rated on separate 5-point scales, yielding both dichotomous (present or absent) and continuous scores for each symptom and for the disorder as a whole. Third, the CAPS-1 contains explicit, behaviorally anchored prompt questions and rating anchors which were intended to enhance reliability by increasing rating precision. Fourth, in

[1]Address for Drs. Weathers and Litz: National Center for PTSD (116B), VA Medical Center, 150 South Huntington Ave., Boston, MA 02130. Internet: WEATHERS.FRANK@BOSTON.VA.GOV, LITZ.BRETT.@BOSTON.VA.GOV

addition to the global rating of response validity, the CAPS-1 assesses the validity of responses to individual symptoms by means of a "Questionable Validity" (QV) rating for each item. Finally, the CAPS-1 provides guidelines and specific questions for assessing lifetime PTSD.

Soon after the CAPS-1 reached its final form we began investigating its reliability and validity. Regarding reliability, the primary concern for a diagnostic interview is the level of agreement between two interviewers conducting independent assessments. Since the CAPS-1 yields a PTSD diagnosis and a continuous measure of symptom severity, we examined the reliability of both types of scores. In studies of interrater reliability the interval between interviews must be chosen carefully. If the interval is too long, agreement between raters may be reduced by true changes in a patient's clinical status. Conversely, if the interval is too brief, agreement between raters may be inflated by patients' memory of their previous responses. We decided that a test-retest interval of 2-3 days would provide an optimal test of interrater reliability of the CAPS.

Although reliability is important, a reliable scale is not necessarily valid. In developing the CAPS-1 we were concerned with three types of validity: content validity, criterion-related validity, and construct validity. Content validity is the extent to which scale items correspond to essential aspects of a construct. With the CAPS-1 we addressed this by including items assessing core and associated features of the disorder and by relying on expert judgment in writing, piloting, and modifying probe questions and rating anchors. Criterion-related validity is the ability of the scale to predict something of interest. The type of criterion-related validity most relevant for diagnostic interviews such as the CAPS-1 is diagnostic utility, or the ability of a scale to identify individuals with and without the disorder. We opted to use the PTSD module of the Structured Clinical Interview for DSM-III-R (SCID) as the criterion because at the time it was the most widely used PTSD interview, and the only one shown to be reliable and valid (e.g., Kulka et al., 1988). Construct validity can be established by demonstrating that a scale correlates strongly with other measures of the same construct but not with measures of other constructs. We examined the correlation of the CAPS-1 with other measures of PTSD as well as with measures of depression, generalized anxiety disorder (GAD), and antisocial personality disorder (ASPD). We hypothesized that the CAPS-1 would show strong correlations with other measures of PTSD, moderate correlations with measures of depression and anxiety, since they are related constructs, and weak correlations with measures of ASPD.

In a preliminary investigation we administered the CAPS-1 to 25 combat veterans referred to the National Center (Blake et al., 1990). Subjects then completed the Combat Exposure Scale (CES; Keane et al., 1989), the Mississippi Scale for Combat-Related PTSD (Keane et al., 1988), and the PTSD scale of the MMPI (PK; Keane et al., 1984). For seven subjects a second clinician made independent ratings during the CAPS-1 interview. Interrater reliability

was excellent. Reliability coefficients for the frequency and intensity scores for each of the three PTSD symptom clusters (reexperiencing, numbing and avoidance, and hyperarousal) ranged from .92 to .99. Also, the two raters agreed on the diagnosis of all seven subjects, five of whom had PTSD. In addition, the three symptom clusters showed a high degree of internal consistency, with Cronbach's α coefficients ranging from .73 to .85. In terms of validity, the CAPS-1 showed strong correlations with the Mississippi Scale (.70) and the PK scale (.84), and a moderate correlation with the CES (.42).

Encouraged by these results we conducted a more stringent examination of the psychometric properties of the CAPS-1 in a larger sample of combat veterans (Weathers et al., 1992, 1994). We incorporated the methodological refinements noted earlier. First, subjects were administered the CAPS-1 twice by two clinicians working independently. This is a demanding test of reliability in that it involves multiple sources of potential error, including clinician variables such as how questions are asked or pursued, and subject variables such as response inconsistencies and genuine changes in clinical status. In contrast, the interrater reliability in the pilot study may have been inflated because ratings were made during the same interview and thus were based on identical information. Second, we evaluated the diagnostic utility of the CAPS-1 against a PTSD diagnosis derived from the SCID PTSD module. Finally, we included multiple measures of PTSD, as well as anxiety, depression, and ASPD.

Subjects were 123 service-seeking Vietnam theater veterans seen at the National Center for PTSD at the Boston DVA Medical Center. They were primarily white (73%), separated/divorced or never married (68%), unemployed or retired/disabled (61%), and had at least a high school education (87%). Most were veterans of either the Army (59%) or the Marines (29%). Mean age of the sample was 43.7. Measures included the SCID and the CAPS-1, as well as the MMPI-2 (Butcher et al., 1989) the Mississippi Scale, the Beck Depression Inventory (BDI; Beck et al., 1961), and the Beck Anxiety Inventory (BAI; Beck et al., 1988).

The first 60 subjects were administered the SCID by one clinician; followed 2-3 days later by the CAPS-1, administered by a second clinician; followed 2-3 days later by a second CAPS-1, administered by a third clinician. The other 63 subjects were administered the SCID, followed 2-3 days later by a single CAPS. Three clinicians served as CAPS-1 interviewers for the 60 subjects who were administered the CAPS-1 twice. There were three unique rater pairs (1, 2; 1, 3; 2, 3) and each pair interviewed 20 subjects. All clinicians worked independently and were blind to all other information about a subject.

The results confirmed that the CAPS-1 has excellent reliability. Test-retest reliability coefficients were calculated separately for each of the three rater pairs, based on CAPS-1 severity scores (Frequency + Intensity). Over the three symptom clusters these coefficients ranged from .77 to .96. For the CAPS-1 total severity score (summed over all 17 DSM symptoms), test-retest reliability ranged from .90

to .98. Also, the internal consistency of the CAPS-1 was strong, with alpha coefficients ranging from .85 to .87 for the three symptom clusters and an alpha coefficient of .94 for all 17 symptoms. This indicates a high degree of homogeneity among CAPS-1 items, suggesting that at least in this population of Vietnam combat veterans the CAPS-1 is measuring a unitary construct.

One disappointing finding was that the QV ratings had very low levels of reliability. The highest phi coefficient for an individual CAPS-1 item was .48, with others ranging from -.09 to .32. One reason for this is that interviewers did not use the "Yes" rating option very often. In fact, reliability could only be calculated for 14 of the 25 CAPS-1 symptoms because for the other 11 symptoms at least one of the two interviewers did not use the "Yes" option. Low reliability also may be due to clarification of questions occurring on the first interview. Interviewers are instructed to use the QV rating to indicate the presence of any factor that may have affected the validity of a subject's answer, including deliberate attempts at misrepresentation, but also including any confusion or misinterpretation of a question. If any confusion were cleared up on the first interview then the second interviewer would be less likely to make a QV rating. One solution for improving the reliability of the QV ratings is to supply more explicit guidelines for their use. However, it may be that ultimately such ratings will not prove to be useful.

In terms of diagnostic utility we found that the CAPS-1 was highly predictive of a PTSD diagnosis based on the SCID PTSD module. Using the CAPS-1 as a continuous measure we employed signal detection methods (Kraemer, 1992) to identify the optimally efficient cutoff score. A CAPS-1 total severity score of 65 had 84% sensitivity, 95% specificity, 89% efficiency, and a kappa of .78 against the SCID PTSD diagnosis. Using the CAPS-1 as a diagnostic measure a kappa of .77 was found between the CAPS-1 diagnosis and the SCID diagnosis. A significant advantage of the CAPS-1 is that cutoff scores can be adjusted for optimal prediction in different populations or against more or less stringent criteria.

Regarding construct validity the CAPS-1 was strongly correlated with other measures of PTSD including the Mississippi Scale (.91), the PK scale (.77), and the number of PTSD symptoms on the SCID (.89). It was almost as strongly correlated with measures of depression including the BDI (.74), the DEP content scale of the MMPI-2 (.69), and the number of depression symptoms on the SCID. Similarly, it was also strongly correlated with measures of anxiety including the BAI (.76), the ANX content scale of the MMPI-2 (.65), and the number of GAD symptoms on the SCID (.66). However, the CAPS-1 was only weakly correlated with measures of antisocial personality including a self-report checklist of antisocial behaviors (.14), the ASP content scale of the MMPI-2 (.34), and the number of ASPD symptoms on the SCID (.16). We then corrected these correlations for possible response bias by partialling out scores on the F scale of the MMPI-2 (a scale associated in part with overendorsement of psychopathology). The

CAPS-1 remained strongly correlated with the Mississippi Scale (.83) and the number of SCID PTSD symptoms (.82), but the correlations with the anxiety and depression measures dropped substantially, ranging from .36 to .55. The corrected correlations with the ASPD measures essentially dropped to zero, ranging from -.04 to .03.

Taken together, these results suggest that the CAPS-1 has very desirable psychometric properties. To date we have administered the CAPS-1 to hundreds of veterans at the National Center and we have found it very useful for a variety of clinical and research applications. If further research corroborates the findings to date, it appears that CAPS-1 can serve well as a standard assessment instrument in the field of traumatic stress. Undoubtedly, changes in the CAPS-1 will occur over time, propelled by new research, by feedback from clinicians and researchers who use the scale, and by the evolution of the construct of PTSD. In the remainder of this article we outline several issues and future directions in the development of the CAPS.

Scoring rules. Two of the features of the CAPS-1, the use of continuous rating scales and the separate assessment of frequency and intensity, although advantageous in many respects, present a challenge when the goal is to rate PTSD symptoms or diagnosis as either present or absent. When we first developed the CAPS-1 we proposed a rationally derived rule that a frequency of "1" or greater and an intensity of "2" or greater for a given CAPS-1 item would constitute the presence of a symptom. A diagnosis of PTSD then would be ascertained by following the DSM-III-R requirement of one reexperiencing symptom, three avoidance and numbing symptoms, and two hyperarousal symptoms. Based on our own clinical experience, as well as on feedback from colleagues and researchers in other settings, it appeared that this "1-2" rule was too lenient, leading to overdiagnosis of PTSD.

We since have explored several alternative scoring rules. Two of these rules also were derived rationally. According to the first such rule, a symptom is considered present if the severity of a CAPS-1 item (Frequency + Intensity) is 4 or greater. The second rule was based on the ratings of 25 doctoral-level clinicians at the National Center. We asked clinicians to consider every Frequency-Intensity combination for all CAPS-1 items, rating each combination as indicating that a symptom was either absent, subthreshold, or present. This rule was intended to be applicable to any traumatized population, not just to combat veterans, so clinicians were instructed to keep the DSM-III-R criteria in mind as they made their ratings. In the final scoring rule, a combination indicates the presence of a symptom if it was rated present by at least 60% of clinicians. A combination indicates the absence of a symptom if it was never rated as present, and was rated as subthreshold by no more than 25% of clinicians. All other combinations are considered subthreshold. The results of these ratings are available from the first author.

The last two scoring rules we have examined were derived empirically, again employing Kraemer's (1992) signal detection methods to identify optimal cutoff scores.

Table 1
Test-Retest Reliability of PTSD Diagnosis Based on Different CAPS Scoring Rules

Scoring rule	kappa
F1-I2 rule	.78
Clinician ratings	.80
Severity > or = 4	.78
Diagnosis calibration rule	.68
Symptom calibration rule	.89

Using the severity scores for each CAPS-1 item (which range from 0-8), we searched for the optimally efficient cutoff score for predicting either the corresponding SCID PTSD symptom (symptom calibration rule) or the SCID PTSD diagnosis (diagnosis calibration rule). Thus, each CAPS-1 item had its own cutoff score. Severity scores at or above the cutoff for a CAPS-1 item indicated the presence of a symptom. We then followed the DSM-III-R decision rule for diagnosing PTSD. (The cutoff scores for individual CAPS-1 items used in these two scoring rules are available from the authors.) Tables 1 and 2 present the reliability and diagnostic utility of these various scoring rules. The "best" scoring rule in this sample appears to be the symptom calibration rule. It has the highest test-retest reliability, as well as the highest kappa coefficient for predicting the SCID PTSD diagnosis. Interestingly, the clinician rating rule is the most stringent rule, yielding the lowest PTSD prevalence of any of the scoring rules and demonstrating a high level of specificity. This rule, which ensures few false positive diagnoses, would be useful, for example, in creating a homogeneous group of PTSD subjects for a research protocol.

New populations. The research to date has been conducted primarily on combat veterans. However, the CAPS-1 was designed to assess PTSD resulting from any type of psychological trauma. The CAPS-1 currently is being used by investigators studying other types of traumatized populations such as rape victims, and more research is needed to determine how the CAPS-1 performs in these populations. We have been using the CAPS-1 with substance abusers to assess PTSD resulting from non-military trauma, including childhood physical and sexual abuse, and a range of civilian traumas. Although this work is in its early stages we feel that the CAPS-1 is performing well, especially for adult-onset civilian traumas such as rape or other violent assault and severe car accidents.

For some populations, difficulties in assessment seem to lie not with the CAPS-1 but with the way PTSD is currently conceptualized. For example, if a strict interpretation of the DSM criteria is followed, then the avoidance and numbing and the hyperarousal symptom clusters require a change from a previous level of functioning. This may not be a reasonable requirement for PTSD resulting from early trauma, since the onset of the trauma may have occurred so early that there really was no "previous level of functioning." Also, items such as loss of interest or pleasure may not be applicable because of the obvious changes in age-appropriate activities. Our approach in this work has been to relax the requirement of a change from a previous level of functioning and simply to assess the symptom as is. For example, if a patient reports feeling little interest in anything or describes a clearly restricted range of affect, we code those symptoms as present even if there is no documented change in functioning.

Assessing other hypothesized features of PTSD. The structure of the CAPS-1, with its separate assessment of frequency and intensity of symptoms and its behaviorally anchored questions and rating scales, can serve as a useful framework for assessing unexplored features of PTSD. A recent example from our research investigating the parameters of emotional numbing (EN) in PTSD illustrates this point. We are interested in disaggregating the construct of EN and providing an empirical explication of this important phenomenon. Toward this end we have begun to ask trauma survivors not only if they experience disinterest, detachment, and restricted range of affect, but also if they have specific deficits in the disclosing, sharing, and outward expression of emotional reactions they are experiencing internally (Litz, 1992). We have labeled this behavior as "strategic withholding of emotional reactions," and have included it as item 10A in our CAPS-1 interviews. Following the format of other CAPS-1 items, the frequency question is worded: "Are there times when you deliberately choose to not show your feelings or let others know you are reacting?" Similarly, the intensity question is worded: "At times when you felt emotional, how much did you hold back or not show your emotional reactions? What emotions did you not express? Did you find yourself holding back more positive emotions, like happiness, negative

Table 2
Diagnostic Utility of CAPS Scoring Rules With SCID-Based PTSD Diagnosis as Criterion

Scoring rule	Sensitivity	Specificity	Efficiency	kappa
F1-I2 rule	.88	.73	.81	.62
Clinician ratings	.68	.98	.81	.64
Severity > or = 4	.87	.74	.81	.62
Diagnosis calibration rule	.90	.82	.86	.72
Symptom calibration rule	.91	.86	.89	.77

emotions, like sadness, or did you hold back all emotions equally?" Our preliminary findings suggest that asking these questions offers a more dimensional way of understanding the emotional behavior of trauma survivors. We have found that many individuals with PTSD report that they still have a broad range of subjective emotional experiences, but that they often deliberately suppress overt expression of their internal experience.

DSM-IV. Changes in how PTSD is conceptualized are inevitable. The CAPS-1 can be modified easily to incorporate these changes. For example, in the DSM-IV there are several modifications of the PTSD criteria. One is the change in Criterion A, the definition of a traumatic stressor. This change does not affect the CAPS-1, which is a symptom measure only, and does not measure the presence or absence of traumatic life events. Two other changes will impact on the CAPS-1, but will require very minor alterations of the scale. First, the symptom of physiologic reactivity has been moved from the hyperarousal cluster to the reexperiencing cluster. For the CAPS-1, this will simply involve moving items around. The other significant change is the addition of Criterion F, which is the requirement that the person experience either significant distress or that the PTSD symptoms cause marked impairment in an important area of functioning. The CAPS-1 already contains items tapping impairment in social and occupational functioning. These are two of the global ratings, CAPS-1 items 18 and 19. Adopting a cutoff of 2 or greater for either item, indicating at least moderate functional impairment in at least one of these two areas, should be sufficient for satisfying this requirement. In terms of distress, this can be inferred from patients' reports regarding their symptoms. It seems straightforward that if there is marked distress during reexperiencing, for example, then this requirement would be satisfied. It might be useful to develop a CAPS-1 global rating which directly assesses the subjective distress at having PTSD symptoms, similar to the questions in the SCID regarding phobias.

In summary, our research has shown that the CAPS-1 is a reliable and valid structured interview for PTSD, which has proven useful for a variety of clinical and research tasks. In general, advances in the field of traumatic stress depend on the widespread adoption of standardized assessment procedures using psychometrically sound instruments. This will greatly increase comparability of results across populations and clinical research laboratories. The adoption of the CAPS-1 can facilitate such standardization.

REFERENCES

BECK, A.T., EPSTEIN, N., BROWN, G. & STEER, R.A. (1988). **An inventory for measuring clinical anxiety: Psychometric properties.** *Journal of Consulting and Clinical Psychology, 56,* 893-897.

BECK, A.T., WARD, C.H., MENDELSON, M., MOCK, J.E. & ERBAUGH, J.K. (1961). **An inventory for measuring depression.** *Archives of General Psychiatry, 4,* 561-571.

BLAKE, D.D., WEATHERS, F.W., NAGY, L.N., KALOUPEK, D.G., KLAUMINZER, G., CHARNEY, D.S. & KEANE, T.M. (1990). **A clinician rating scale for assessing current and lifetime PTSD: The CAPS-1.** *Behavior Therapist, 13,* 187-188.

BUTCHER, J.N., DAHLSTROM, W.G., GRAHAM, J.R., TELLEGEN, A. & KAEMMER, B. (1989). **Minnesota Multiphasic Personality Inventory-2 (MMPI-2): Manual for administration and scoring.** Minneapolis: University of Minnesota Press.

KEANE, T.M., CADDELL, J.M. & TAYLOR, K.L. (1988). **Mississippi Scale for Combat-Related Posttraumatic Stress Disorder: Three studies in reliability and validity.** *Journal of Consulting and Clinical Psychology, 56,* 85-90.

KEANE, T.M., FAIRBANK, J.A., CADDELL, J.M., ZIMERING, R.T., TAYLOR, K.L. & MORA, C.A. (1989). **Clinical evaluation of a measure to assess combat exposure.** *Psychological Assessment, 1,* 53-55.

KEANE, T.M., MALLOY, P.F. & FAIRBANK, J.A. (1984). **Empirical development of an MMPI subscale for the assessment of combat-related posttraumatic stress disorder.** *Journal of Consulting and Clinical Psychology, 52,* 888-891.

KRAEMER, H.C. (1992). **Evaluating medical tests: Objective and quantitative guidelines.** Newbury Park, CA: Sage Publications.

KULKA, R.A., SCHLENGER, W.E., FAIRBANK, J.A., HOUGH, R.L., JORDAN, B.K., MARMAR, C.R. & WEISS, D.S. (1988). **National Vietnam veterans readjustment study (NVVRS): Description, current status and initial PTSD prevalence estimates.** Research Triangle Park, NC: Research Triangle Institute.

LITZ, B.T. (1992). **Emotional numbing in combat-related posttraumatic stress disorder: A critical review and reformulation.** *Clinical Psychology Review, 12,* 417-432.

SPITZER, R.L., WILLIAMS, J.B., GIBBONS, M. & FIRST, M.B. (1990). **Structural Clinical Interview for DSM-III-R - Patient Edition (SCID-P, Version 1.0).** Washington, DC: American Psychiatric Press.

WEATHERS, F.W., BLAKE, D.D., KRINSLEY, K.E., HADDAD, W.H., HUSKA, J.A. & KEANE, T.M. (1992, November). **The Clinician-Administered PTSD Scale: Reliability and construct validity.** Paper presented at the annual meeting of the Association for Advancement of Behavior Therapy, Boston, MA.

WEATHERS, F.W., BLAKE, D.D., KRINSLEY, K.E., HADDAD, W.H., HUSKA, J.A. & KEANE, T.M. (1994). **The reliability and validity of the Clinician-Administered PTSD Scale.** Manuscript submitted for publication.

ITEM RESPONSE THEORY AND PTSD ASSESSMENT

Lynda A. King, PhD[1], and Daniel W. King, PhD[1]
Central Michigan University

The approach to test development and evaluation known as item response theory (IRT) is rapidly gaining prominence in applied psychometrics. IRT is based on a series of mathematical functions or item-characteristic curves that describe the relationship between an individual's standing

[1]Address for Drs. King and King: Central Michigan University, Mount Pleasant, MI 48859-0001.
Internet: 3AJCQ3Q@CMUVM.CSV.CMICH.EDU

on a construct or attribute of interest (the X axis) and the probability that the individual will answer a test item or interview question in a certain way (the Y axis). When IRT was first proposed, the item-characteristic curve was the normal ogive, or an S-shaped cumulative normal density function that represents the likelihood of responding in the keyed direction on a dichotomously scored ("correct"/ "incorrect") test item. Later, a logistic function was introduced that made the IRT computational procedures more manageable. This function, the basis for most IRT applications today, is determined by up to three parameters or item characteristics: a *difficulty* or *attribute threshold level* at which persons responding to a test item will tend to choose one option over another, a *discrimination index* that tells how well the item differentiates among respondents at the threshold, and a *guessing* or *faking index* that reflects the likelihood an individual having no amount of the attribute will nonetheless endorse the item in the keyed direction. The mathematics underlying the estimation of IRT-based item characteristics are complex, and only since the availability of high-speed computers have the techniques become accessible to the general research community.

The first task in using item response theory is to determine which of many IRT models seems most suited to one's data. The decision depends on several factors, including assumptions about the dimensionality of the construct being measured, the type of item, and which combination of the three item characteristics best reflects the nature of the item. At present, the most commonly used models require unidimensionality, although there is growing interest in multidimensional IRT. Different IRT models have been developed to accommodate various item types (e.g., dichotomous, multiple-category nominal, ordinal Likert-type). Some estimate only a difficulty parameter and assume equivalent discrimination across items and no guessing; others estimate both difficulty and discrimination indices, and yet others estimate all three.

Once a model is chosen, data are submitted to maximum likelihood estimation procedures. These statistical methods derive parameter estimates for each item that have the greatest probability of having produced the observed data, given the selected IRT model. The parameter estimates for each item may then be used to graph the item-characteristic curves and go further to compute what is known as item and test information functions. The item information function can be expressed as a graphical representation of the precision of the item (Y axis) across a broad range of the measured attribute (X axis). At any given attribute score, the square root of the inverse of the information value is the standard error of measurement. Thus, the higher the information, the lower the standard error of measurement, and the more reliable is the item for that point on the attribute dimension. The sum of the information functions for items comprising a test will yield the test information function, a map of the precision of the full test (Y axis) across the attribute dimension (X axis). By examining the information functions for an item or a test, one can determine at which point on the attribute continuum the item or test is most

reliable and therefore most useful in discriminating among persons. The test information function is more revealing than its counterpart classical test theory-based reliability coefficient, which documents precision only at the mean of the attribute distribution.

In addition to supplying parameter estimates, item-characteristic curves, and item and test information functions, IRT also provides a method for scoring individuals. Again, the statistical basis is maximum likelihood estimation; an individual receives the score with the highest probability of having produced his or her pattern of responses to the items. The scores assigned to persons, therefore, are not the simple sums or averages of item scores that are commonly computed using a classical test theory psychometric approach. It can be demonstrated mathematically that IRT-based attribute estimates are efficient, or closer to an individual's true score, than those derived from any other scoring procedure.

Two advantages are most often cited for using IRT. First, item characteristics are sample-invariant or sample-independent. This means that the estimate of a parameter does not depend upon whether the individuals who provide the data possess generally high or generally low amounts of the attribute or are spread across the whole range of possible attribute scores. Second, the estimate of an individual's standing on a particular attribute does not depend on the specific set of items to which the individual responds. This means that persons can be scored on the same underlying attribute scale when they are administered different sets of items. Neither of these features can be claimed by the more familiar classical test theory, for which item characteristics are tied to the particular sample upon which they are computed and an individual's test score is a consequence of the particular items that are administered. All in all, IRT appears to provide more information about items, better understanding of how items represent the construct they are intended to measure, and greater flexibility in the use of items for various purposes.

We believe that IRT can contribute to PTSD assessment and research. One application is the detection of item bias or differential item functioning, defined as a situation in which an item's parameter estimates and characteristic curve are not the same for different populations. For example, one could conduct a differential item functioning study to determine if a collection of test items, such as a PTSD symptom checklist, manifests equivalent characteristics across different trauma populations. If differential item functioning were detected, the interpretation would be that the item carries alternative meanings for different groups and perhaps that the PTSD construct is not the same across populations. Similarly, differential item functioning methods could assist in the translation of a PTSD measure to another language. Because, as noted previously, item characteristics are independent of the sample upon which they are computed, item sets with equivalent characteristics are judged to have the same meaning across groups and IRT-based attribute estimates or scores are on

the same scale. Thus, one can assure equivalence of PTSD measurement across translations and, in turn, the validity of cross-cultural trauma inquiry would be enhanced.

An exciting potential future application of IRT to PTSD assessment is in the realm of interactive computerized adaptive testing. This innovation in test administration, scoring, and interpretation uses IRT-based item characteristics to provide an optimally precise subset of items tailored to the person being assessed. The essence of computerized adaptive testing is that IRT-based attribute estimates for individuals are not dependent on the items that are administered. The individual's responses to initial items yield a first approximation of location on the attribute dimension, with subsequent items computer-selected to be optimally precise around that location, and with each iteration of this process using items that triangulate on the most likely attribute score. Thus, each person is administered only those items necessary to reach some predetermined and acceptable level of score accuracy. This results in far fewer items being needed and much more efficient assessment: PTSD computerized "short forms." Recent research with affective measures has shown that item pools containing 30 to 35 items may be sufficient for developing a computerized adaptive testing strategy, and that the successful implementation of such a system can reduce the number of items required by about 50% with no loss in decision accuracy. Furthermore, new techniques are being developed to make certain that content breadth is represented in any computer-selected item set. This feature would be important for PTSD assessment, which requires evidence from three symptom categories.

As a part of our NIMH-funded (Violence and Traumatic Stress Program) project, we have undertaken several IRT studies of Mississippi Scale data from the NVVRS. An initial study examined the items in the original military version of the scale, a follow-up study examined the items in the civilian version, and we are currently conducting a study of what appears to be differential item functioning across the two forms. For our analyses, we have relied on David Thissen's MULTILOG program (distributed by Scientific Software, Chicago, 800-247-6113; available in mainframe and PC versions), which is suitable for the Mississippi Scale's Likert response format as well as for dichotomous and multiple-category nominal item formats. Other software available for IRT analyses includes: BILOG (also available from Scientific Software, mainframe and PC, exclusively for dichotomous items); and RASCAL, ASCAL, and XCALIBRE (Assessment Systems Corporation, 612-647-9220; PC only, dichotomous items only).

Recommended sources of further information:

HAMBLETON, R. K., SWAMINATHAN, H., & ROGERS, H. J. (1991). **Fundamentals of item response theory.** Newbury Park, CA: Sage.

HOLLAND, P. W., & WAINER, H. (Eds.). (1993). **Differential item functioning.** Hillsdale, NJ: Erlbaum.

KING, D. W., KING, L. A., FAIRBANK, J. A., SCHLENGER, W. E., AND SURFACE, C. R. (1993). **Enhancing the precision of the Mississippi Scale for Combat-Related Posttraumatic Stress Disorder: An application of item response theory.** *Psychological Assessment, 5,* 457-471.

LORD, F. M. (1980). **Applications of item response theory to practical testing problems.** Hillsdale, NJ: Erlbaum.

WAINER, H. (Ed.). (1990). **Computerized adaptive testing: A primer.** Hillsdale, NJ: Erlbaum.

The Pacific Islands Division of the National Center for PTSD has developed a self-administered, computerized version of the CAPS that runs under the Windows 3.1 operating environment. Clients are presented questions on the screen and use the mouse for input. If you are interested in obtaining a copy of the program and/or being involved in testing the Computerized CAPS (C-CAPS), please contact the program developers at the following address: David C. Richard, Pacific Center for PTSD, 1132 Bishop Street #307, Honolulu, HI 96813. Phone (808) 566-1649. Fax: (808) 566-1885.

National Center for PTSD (116D)
VA Medical and Regional Office Center
White River Junction, Vermont 05009