Available online at www.sciencedirect.com
ScienceDirect
Psychiatry Research 166 (2009) 269–280
PSYCHIATRY RESEARCH
www.elsevier.com/locate/psychres

Evaluating the reliability of multiple assessments of PTSD symptomatology: Multiple examiners, one patient

Domenic Cicchetti [a,*], Alan Fontana [b,c], Donald Showalter [b]

[a] Yale University, Child Study Center and Departments of Psychiatry and Biometry, New Haven, CT 06512, United States
[b] North East Program Evaluation Center (NEPEC), West Haven, CT 06516, United States
[c] Yale University, Department of Psychiatry, New Haven, CT, United States

Received 14 February 2007; received in revised form 28 June 2007; accepted 24 January 2008

Abstract

The objective of this investigation was to assess the inter-examiner reliability of PTSD symptomatology by 12 clinical examiners who evaluated independently a single-case Vietnam-Era veteran, using videotaped clinician interviews with The Clinician Administered PTSD Scale-1 (CAPS-1). A second patient was utilized for cross-validation purposes. Data were analyzed using a specially designed Kappa statistic. In previous reliability assessments of the CAPS-1, a pair of examiners assessed multiple patients, and demonstrated evidence of high reliability and validity. As in previous reliability assessments, reliability was assessed both for *frequency* and *intensity* of PTSD symptomatology in both patients. Results indicated that the reliability levels of the CAPS-1 were consistently and almost exclusively in the excellent to perfect levels of inter-examiner agreement, as based upon both global assessments and on a symptom-by-symptom basis. The results of this investigation are interpreted in the broader framework of their applicability to assessing inter-examiner agreement in clinical trials or other large multi-site studies.

Keywords: PTSD symptomatology; Multiple examiners; Reliability; Single case

1. Introduction

The Clinician Administered PTSD Scale-1 (CAPS-1; Blake et al., 1990) has been shown to be a highly reliable clinical instrument for assessing both the frequency and intensity of PTSD, when administered independently, by two examiners, to large groups of combat veterans (Weathers and Litz, 1994). The CAPS-1 comprises 19 PTSD items that are first scored for frequency and then for the intensity of symptomatology.

Each item is scored on a five-category dichotomous-ordinal scale (Cicchetti, 1976; Cicchetti and Sparrow, 1981) in which 0 denotes that the symptom is "absent" and 4 denotes that it is maximally "present."

In the development of the CAPS-1, the authors (Blake et al., 1990), devised five scoring rules: (1) The first scoring rule (or "1-2" rule) was "rationally derived" and used as the criterion for "presence" of a PTSD symptom, namely, a combined *frequency* score of "1" or greater, with an *intensity* score of "2" or greater. (2) The second of these criteria considers a PTSD symptom present if the severity

* Corresponding author. Tel.: +1 203 488 6563; fax: +1 203 483 1123.
E-mail address: dom.cicchetti@yale.edu (D. Cicchetti).

0165-1781/$ - see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.psychres.2008.01.014

of a CAPS-1 item, based upon the sum of its *Frequency* and *Intensity* rating, is 4 or greater. (3) The third rule requires examiners to consider each item's *Frequency–Intensity* combination as indicating whether a given PTSD symptom is "absent," "subthreshold," or "present," based on DSM-III-R criteria. The fourth and fifth scoring rules (4-5) consist, respectively, of a symptom calibration rule as well as a diagnostic calibration rule. Here "optimally efficient cut-off scores" (Blake et al., 1990, p. 5), based on signal detection methods (Kraemer, 1992), are used to define cut-off scores for "presence" or "absence" of a given PTSD symptom.

The five scoring rules of the CAPS have been shown to have kappa (*K*) values ranging between 0.68 and 0.89. Values this high define the levels of reliability to range between GOOD (0.60–0.74) and EXCELLENT (0.75 and above), as given by the criteria of Cicchetti and Sparrow (1981) and Cicchetti (2001). Similarly, these same kappa values are considered Substantial and Almost Perfect, respectively, by the earlier criteria of Landis and Koch (1977).

The dual purposes of this report are: (1) to determine whether multiple independent clinical examiners can reliably evaluate the extent of PTSD symptomatology in a single veteran patient; and (2) to determine whether the results can be successfully replicated on a second veteran patient.

To our knowledge, this type of reliability assessment, though often necessary, is seldom undertaken. The value of the current multi-rater study is that it examines the level of inter-examiner reliability using a constant stimulus, here a videotape that is based on a live clinician and patient interview.

As was true of the Weathers and Litz, (1994) investigation, in the typical inter-examiner reliability study, a small number of examiners (i.e. 2) evaluate a relatively large sample (*n*) of subjects (say, 30 or more — in the study of Weathers et al., the *n* was actually 123). Depending upon the specific rater by subject configuration pattern, and the specific scale upon which the variable is measured, an appropriate model of kappa, weighted kappa or the mathematically similar intraclass *r* is applied (e.g., Fleiss, 1981; McGraw and Wong, 1996; Shrout and Fleiss, 1979). In fact, once the overall and specific kappa or weighted kappa coefficients are derived, it is even possible to derive a coefficient for each subject in the sample (e.g., Cicchetti et al., 1990). The problem is not as straightforward in the case of multiple examiners rating independently a single case.

In a larger theoretical research framework, and from a historical perspective, there are two basic types of research orientations, nomothetic and ideographic. The terms were borrowed from the Kantian Philosopher, Wilhelm Windelbrandt (1894) by the American psychologist Gordon Allport, in 1937 (for a recent scholarly treatise on the subject, see also Grice et al., 2006). The first and far more common is the nomothetic or group focused research in which large numbers of subjects are studied and summary statistics, such as means or averages, or omnibus correlational or agreement statistics are applied. In the ideographic approach to science, the individual subject is the unit of investigation. Nonetheless, the approach has resulted in some major scientific discoveries. In the field of biostatistics, the classic case of the woman tasting tea comes to mind. The famous Sir Ronald Fisher used the single-case method to assess the extent to which a lady was able to differentiate successfully cups of tea in which the milk had been poured before the tea was poured, from those in which the tea was poured before the milk. From this study, Fisher was able to derive the concepts of research design, in general, and randomization, sensitivity, and tests of statistical significance, in particular (Holschuh, 1980; Salsberg, 2001). In another famous single-case study, Broca evaluated his patient, Tan, and thereby discovered that specific area of the brain responsible for human speech, and known eponymously as Broca's area. As noted by Wilson (1987), the discovery of Broca's area marked the beginning of modern day neuropsychology.

Perhaps a major reason why multiple assessments of a single patient have not been routinely reported, either in the PTSD literature, in particular, or in the behavioral and biomedical literature, more generally, is that, heretofore, an appropriate statistical methodology has not been available to assess inter-examiner reliability. In the next section, we will describe such a methodology. We will then apply it to a single patient and provide and discuss the meaning and implications of the obtained results.

2. Methods

2.1. A brief history of multiple assessments of a single case

One of the earliest research articles that focused upon the issue of multiple assessments of a single case was published by Erika Chance in 1963. A group of 542 clinicians, consisting of psychoanalysts, psychiatrists, psychologists, and social workers each analyzed both a psychoanalytical and a psychotherapy hour of a single case. For each case, the effect of different disciplines upon both gender concepts and clinical interpretations,