In mental health care, there is no gold standard, no objective outcome measure, against which to assess the quality of diagnostic judgments. A classification of someone’s problems as, for example, a borderline personality disorder requires judging whether the diagnostic criteria of this disorder are present or not. These are not facts but presentations that require interpretation. Moreover, the disorder does not refer to an external reality that can be read out from these interpretations; that is, a correspondence criterion (Hammond, 1996) cannot be applied. To cope with this limitation, expert judgment is often considered to be the gold standard. Thus, an expert determines whether a candidate for the title is also an expert. This situation is unsatisfyingly circular. The lack of a gold standard means that it is unclear who really is an expert, because the expertise of the certifier has itself not been established by comparing performance to objective outcome measures. As a proxy, and for want of a more solid criterion, diagnostic expertise in clinical settings is often operationalized by years of clinical experience, peer nomination, or a combination of both (Goodyear, 1997), measures that are frequently contaminated by biases such as popularity (Shanteau, Weiss, Thomas, & Pounds, 2002) and familiarity.

Although it is often assumed that knowledge about and experience with a disorder are the main components in coming to an accurate classification (Custers, Regehr, & Norman, 1996; Hillerbrand & Claiborn, 1990; Shanteau, 1992), research has shown that in the clinical domain, expertise as indexed by years of experience is not always significantly associated with superior performance (Ágisdóttir et al., 2006; Garb, 2005; Spengler et al., 2009; Strasser & Gruber, 2004). Indeed, it was previously found that counselors with hardly any experience (0 to 2 years) and those with many (more than 10) years of experience classified equally well, and worse than counselors with an intermediate level of experience (2 to 10 years; Witteman & Van den Bercken, 2007).

On the basis of a suggestion by Cochran (1943), Weiss and Shanteau (2003) argued that expert judgment requires two key abilities: discrimination and consistency. Discrimination means that a counselor should be able to discriminate between a major depressive disorder and an anxiety disorder. Consistency means that the expert must make consistent decisions when repeatedly faced with the same or similar symptom patterns. Weiss and Shanteau quantified these notions in the form of the Cochran–Weiss–Shanteau (CWS) index, which defines expert judgment as the ratio of discrimination to (in) consistency. The higher the observed value of the CWS index, the better the performance. With this index, expertise can be evaluated without making use of an external gold standard; a coherence criterion (Hammond, 1996) is used instead.

Several studies have used the CWS index of expertise, including examinations of general practitioners (Skånér, Strender, & Bring, 1998), occupational therapists (Rassaiani, Ziviani, Rodger, & Dalgleish, 2009), ergonomists (Williams, Haslam, & Weiss, 2008), and auditors, agricultural judges, and personnel selectors (Shanteau et al., 2002). The index was externally validated using two tasks (mental calculation and golf putting) with known optimal responses (Weiss, Brennan, Thomas, Kirlik, & Miller, 2009). In our study, we used the CWS index to find differences in expertise in diagnostic classification, that is, in deciding whether a client suffers or does not suffer from major depression.

Major depression is a disorder marked by frequent comorbidity with other disorders and produces high variation...
Assessing Diagnostic Expertise With the CWS Index

in symptoms as well as interaction with many inter- and intrapersonal factors (Hasin, Goodwin, Stinson, & Grant, 2005). The syndrome is well known to most counselors and students in (clinical) psychology. A Diagnostic and Statistical Manual of Mental Disorders (4th ed., text rev.; DSM-IV-TR, American Psychiatric Association, 2000) classification is made by assessing and counting symptoms and comparing them to a predefined threshold. Counting is easy, but assessing symptoms is not so straightforward and requires clinical judgment. For the counselor, this is a clinically important judgment, because treatment decisions are directly linked to the client’s classification. We tested whether the CWS index is able to distinguish counselors with different levels of experience in classifying major depression.

Method

Participants

To have participants with different levels of experience, we recruited from three different groups: 1st-year students in clinical (child) psychology, clinical psychology master’s-level students, and practicing clinical psychologists. Master’s-level students had completed courses in psychopathology, psychodiagnosis, and intervention and were currently completing a clinical internship. Clinical psychologists had finished their master’s courses and were practicing counselors. Participants were recruited via e-mail (practitioners) and via the Behavioural Science Lab participation system at Radboud University Nijmegen, the Netherlands (students). Students who took part in the study could receive partial course credit or 5€ (approximately $7); practitioners were not offered compensation. There were 54 participants (mean age = 28.52 years, SD = 13.31, ranging from 18 to 70 years), consisting of 21 first-year students (17 female), 19 master’s-level students (17 female), and 14 practicing counselors (8 female).

Procedure

The study was conducted using the web version of Inquisit. Participants performed the study either in the Behavioural Science Lab at Radboud University Nijmegen or at their own computer at home. In either case, participants began the experiment by clicking on a link to a website. The program automatically downloaded and ran on the local computer. First, participants received instruction about the study. They were asked to complete the study in an environment where they would not be interrupted and to avoid any distractions such as mobile phones. They then performed the clinical judgment task, after which they were asked to provide additional demographic data. They were thanked for their participation and were offered the possibility for debriefing via e-mail. The total duration of the study was 10 to 15 minutes.

Classification Task

To identify diagnostic decision-making expertise, we used a classification task to assess each participant’s CWS index score, that is, participants’ ability to discriminate between cases of major depression and other cases, and the consistency of their performance. Short vignettes (see example below) were shown on the computer screen, and participants were asked to indicate their judgment of the probability that the patient described in each vignette suffers from major depressive disorder. In a slight departure from the usual elicitation, we did not ask our participants for a precise probability score but for a probability interval, because we believe this to be more ecologically valid than asking for a precise number (Renooij & Witteman, 1999; Witteman, Renooij, & Koele, 2007). It also reflects the inherent subjectivity of the judgments and the fact that there are no correct answers. The center of the probability interval was entered as the judgment. There were 24 vignettes, six of which were presented twice to assess consistency, resulting in 30 judgments from which the CWS index score was calculated (see Analysis section).

Bachmann et al. (2008) suggested that the number of attributes in a vignette should be limited to increase response reliability. Therefore, the vignettes included only two items about demographic information (all women, age between 30 and 65 years), five symptoms, and three yes/no items about context information (precipitating events, distress or motivation, and earlier treatment). The five symptoms were randomly taken from lists of the DSM-IV-TR criteria of depression (e.g., “Often feels guilty”), anxiety (e.g., “Repeatedly checks if all doors are closed”), or unspecific symptoms of mental health problems that may occur both with depression and with anxiety (e.g., “Sometimes seems absent”). There were equal numbers of vignettes (eight) that contained only symptoms of major depressive disorder, only symptoms of anxiety disorder, or only ambivalent symptoms. The vignettes thus varied considerably in whether they suggested a patient with major depressive disorder or not; such distinctiveness among the stimuli would help to distinguish experts from nonexperts (Dawson, Zeitz, & Wright, 1989; Garb, 1989). The 24 vignettes were pilot-tested for plausibility with 10 doctoral students in clinical psychology and slightly adapted until accepted as veridical. Six vignettes were repeated, two of each type (depression, anxiety, ambivalent).

The following is an example of a vignette suggesting depression rather than anxiety:

A woman, 62 years old, presents with the following symptoms: Does not enjoy life, often feels guilty without reason, often wakes up very early, has little interest in activities, often feels sad or empty. Further information: There have been no special events lately, she shows little motivation to change, she has not been treated before.

Analysis

The CWS index can be compared with an $F$ ratio and is computed in a similar way (Weiss & Shanteau, 2003). The numerator reflects the ability to differentiate between different stimuli and is the variance among the judgments of the 30 vi-
The first-year students and the master's-level students differed significantly (at the .05 level), with the master's-level students performing better. The counselors were not significantly different (at the .05 level), with the master's-level students performing better. The counselors were not significantly different from the 1st-year students and were worse than the master's-level students.

Results

The main question of this study was whether the CWS index differs with different levels of experience. Figure 1 shows that, indeed, it does. The mean CWS for the first-year students \( (n = 21) \) was 8.30 [95% CI 6.59–10.20]; for the master's-level students \( (n = 19) \), it was 31.53 [16.35–51.65]; and for the counselors \( (n = 14) \), it was 12.81 [8.68–17.73].

The first-year students and the master's-level students differed significantly (at the .05 level), with the master's-level students performing better. The counselors were not significantly different from the 1st-year students and were worse than the master's-level students.

Discussion and Conclusion

In this study, we applied the CWS index (Weiss & Shanteau, 2003) in the field of clinical judgment. We found clear differences of CWS expertise among our three groups of participants.

We believe that the task we used in this study is relevant to the assessment of clinical judgment (see also Hauser, Spada, Rummel, & Meier, 2006). Counselors are required to classify their clients' complaints in a DSM-IV-TR category, not so much because clients ask them to do so but because they are told to do so by their institutions and certainly by insurance companies, whose refunds are based on such classifications (Hohenshil, 1996).

It is important to note that both abilities must be simultaneously present. It is trivial for a counselor to appear skilled in one ability at the expense of the other. For example, one can exhibit high consistency by regarding all clients as similar. In the medical analogue of this task, a physician might engage
in “defensive medicine” by recommending that all patients be subjected to the same procedure without regard to their individual symptoms. That strategy would exemplify high consistency but not judgmental expertise. Similarly, one can exhibit high discrimination by regarding all clients as unique, thereby making it impossible to use either direct or vicarious experience as a guide to effective therapy. Expert judgment requires simultaneous exercise of consistency and discrimination. We do not recommend examining either ability in isolation; use of the CWS index forces the evaluator to acknowledge the trade-off.

From our perspective, the CWS index is a new and valuable tool to study expertise in clinical judgment. Using it, we were able to distinguish different levels of expertise in the clinical judgment task of assessing the probability that a client may be classified as suffering from a major depressive disorder, which is quite an important judgment because treatment decisions depend on it. Of course, in real clinical situations, expertise in classifying other mental disorders is required as well.

References


Wood, S., & J. Shanteau (Eds.), *Emerging perspectives on judgment and social policy: Interdisciplinary approaches.* January 2012


Inquisit (Web version) [Computer software]. Seattle, WA: Millisecond Software.


