Received 09/27/10
Revised 01/05/11
Accepted 01/24/11

Research

Assessing Diagnostic Expertise of Counselors Using the Cochran–Weiss–Shanteau (CWS) Index

Cilia L. M. Witteman, David J. Weiss, and Martin Metzmacher

Counseling studies have shown that increasing experience is not always associated with better judgments. However, in such studies performance is assessed against external criteria, which may lack validity. The authors applied the Cochran–Weiss–Shanteau (CWS) index, which assesses the ability to consistently discriminate. Results showed that novice counselors performed almost on the same level as very experienced counselors. The authors thus replicated earlier findings with a novel approach: applying an internal coherence criterion.

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 (Ægisdóttir 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 (Rassafiani, 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

Cilia L. M. Witteman and **Martin Metzmacher,** Behavioural Science Institute, Radboud University Nijmegen, The Netherlands; **David J. Weiss,** Department of Psychology, California State University, Los Angeles. Correspondence concerning this article should be addressed to Cilia L. M. Witteman, Behavioural Science Institute, Radboud University Nijmegen, PO Box 9104, 6500 HE Nijmegen, The Netherlands (e-mail: C.Witteman@socsci.ru.nl).

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-

gnettes. Six vignettes were presented twice. The denominator captures the ability to make consistent judgments for the six vignettes that had been presented twice and is the between-trials variance over the six repeated vignettes. To compare the three groups of participants, we tested for overlapping 84.3% confidence intervals (CIs) of the mean CWS of each group. According to Payton, Greenstone, and Schenker (2003), those intervals allow for pairwise comparisons at the .05 level of significance.

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 [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).

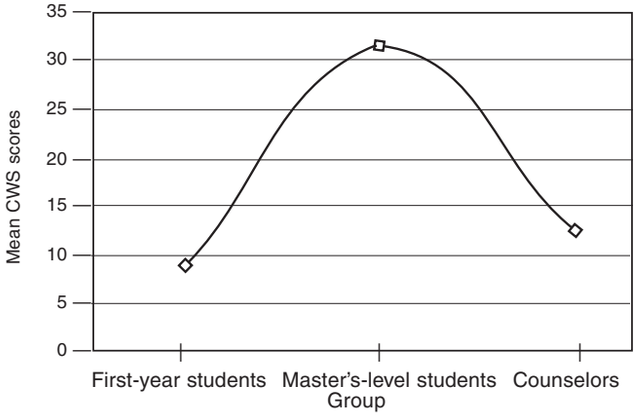

FIGURE 1

Mean Cochran–Weiss–Shanteau (CWS) Index Scores for the Three Experience Groups

It is interesting to note that the professional counselors in this study hardly performed better than the novices. This result has also been obtained with other methods, such as asking counselors to predict adjustment or prognosis or to judge the severity of a client's pathology (e.g., Ægisdóttir et al., 2006; Spengler et al., 2009). Comparing this study's results with those of a previous study (Witteman & Van den Bercken, 2007), we again saw an intermediate effect, whereby performance of a group with levels of experience between the novice and the more experienced (10+ years in the profession), the master's-level students, deviated significantly from the other groups' performance. This, and the large range of master's-level students outcomes, may have resulted from large variability in performance of professionals with in-between levels of experience: They shifted back and forth between rule-based (novices) and memory-based (experienced) judgments (Dougherty, Gronlund, & Gettys, 2003).

The sample used in this research was quite diverse, from first-year students to licensed counselors. An interesting next step would be to compare results of the CWS method with peer nomination. Another point for consideration is the question of how the CWS index increases with training. Shanteau, Friel, Thomas, Raacke, and Weiss (2010) showed that air traffic controllers increased their CWS score within a few sessions while in training. The method used in the current study could be suited to evaluate the effect of different training programs. In counseling, as in other domains in which judgment is paramount, being consistent and able to discriminate stimuli are necessary, if not sufficient, components of expertise. Training could focus on upgrading these crucial abilities.

In the current study, we had presumed that there is such a thing as expertise in the domain of clinical judgment. However, this is a point that is still up for debate. Accurately diagnosing mental disorders involves dynamic stimuli, which lack predictability (Shanteau, 1992). The development of expertise in such tasks depends on the existence of positive (social) feedback processes (Gaines, 1988). It is precisely the lack of feedback that is one of the main problems in the counseling field. Feedback is often so ambiguous that the development of expert skills is extremely difficult or outright impossible (Dougherty et al., 2003). This is an additional reason to focus on the ability to judge consistently and to discriminate between different disorders, for example, in models of supervision for counselors-in-training (Lambie & Sias, 2009).

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

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., & Nichols, C. S. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist, 34,* 341–382. doi:10.1177/0011000005285875

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

Bachmann, L. M., Mühleisen, A., Bock, A., Ter Riet, G., Held, U., & Kessels, A. G. H. (2008). Vignette studies of medical choice and judgment to study caregivers' medical decision behaviour: Systematic review. *BMC Medical Research Methodology, 8,* 50–58. doi:10.1186/1471-2288-8-50

Cochran, W. G. (1943). The comparison of differential scales of measurement for experimental results. *Annals of Mathematical Statistics, 14,* 205–216.

Custers, E. J., Regehr, G., & Norman, G. R. (1996). Mental representations of medical diagnostic knowledge: A review. *Academic Medicine, 71,* S55–S61.

Dawson, V. L., Zeitz, C. M., & Wright, J. C. (1989). Expert–novice differences in person perception: Evidence of experts' sensitivities to the organization of behavior. *Social Cognition, 7,* 1–30.

Dougherty, M. R. P., Gronlund, S. D., & Gettys, C. F. (2003). Memory as a fundamental heuristic in decision making. In S. L. Schneider & J. Shanteau (Eds.), *Emerging perspectives on judgment and decision research* (pp. 125–164). Cambridge, England: Cambridge University Press.

Gaines, B. R. (1988). Positive feedback processes underlying the formation of expertise. *IEEE Transactions on Systems, Man & Cybernetics, 18,* 1016–1020. doi:10.1109/21.23101

Garb, H. N. (1989). Clinical judgment, clinical training, and professional experience. *Psychological Bulletin, 105,* 387–396.

Garb, H. N. (2005). Clinical judgment and decision making. *Annual Review of Clinical Psychology, 1,* 67–89. doi:10.1146/annurev.clinpsy.1.102803.143810

Goodyear, R. K. (1997). Psychological expertise and the role of individual differences: An exploration of issues. *Educational Psychology Review, 9,* 251–265. doi:10.1023/A:1024787208551

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice.* New York, NY: Oxford University Press.

Hasin, D. S., Goodwin, R. D., Stinson, F. S., & Grant, B. F. (2005). Epidemiology of major depressive disorder. *Archives of General Psychiatry, 62,* 1097–1106.

Hauser, S., Spada, H., Rummel, N., & Meier, A. (2006). Expertise development in clinical psychology. In R. Sun, N. Miyake, & C. D. Schunn (Eds.), *Proceedings of the 28th annual conference of the Cognitive Science Society* (pp. 1458–1463). Mahwah, NJ: Erlbaum.

Hillerbrand, E., & Claiborn, C. D. (1990). Examining reasoning skill differences between expert and novice counselors. *Journal of Counseling & Development, 68,* 684–691.

Hohenshil, T. H. (1996). Role of assessment and diagnosis in counseling. *Journal of Counseling & Development, 75,* 64–67.

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

Lambie, G. W., & Sias, S. M. (2009). An integrative psychological developmental model of supervision for professional school counselors-in-training. *Journal of Counseling & Development, 87,* 349–356.

Payton, M. E., Greenstone, M. H., & Schenker, N. (2003). Overlapping confidence intervals or standard error intervals: What do they mean in terms of statistical significance? *Journal of Insect Science, 3,* 34–40.

Rassafiani, M., Ziviani, J., Rodger, S., & Dalgleish, L. (2009). Identification of occupational therapy clinical expertise: Decision-making characteristics. *Australian Occupational Therapy Journal, 56,* 156–166. doi:10.1111/j.1440-1630.2007.00718.x

Renooij, S., & Witteman, C. L. M. (1999). Talking probabilities: Communicating probabilistic information with words and numbers. *International Journal of Approximate Reasoning, 22,* 169–194. doi:10.1016/S0888-613X(99)00027-4

Shanteau, J. (1992). The psychology of experts: An alternative view. In G. Wright & F. Bolger (Eds.), *Expertise and decision support* (pp. 11–23). New York, NY: Plenum.

Shanteau, J., Friel, B., Thomas, R. P., Raacke, J., & Weiss, D. J. (2010). Assessing expertise when performance exceeds perfection. In E. S. Patterson & J. Miller (Eds.), *Macrocognition metrics and scenarios: Design and evaluation for real-world teams* (pp. 85–93). Burlington, VT: Ashgate.

Shanteau, J., Weiss, D. J., Thomas, R., & Pounds, J. (2002). Performance-based assessment of expertise: How can you tell if someone is an expert? *European Journal of Operational Research, 136,* 253–263. doi:10.1016/S0377-2217(01)00113-8

Skånér, Y., Strender, L., & Bring, J. (1998). How do GPs use clinical information in the judgment of heart failure? *Scandinavian Journal of Primary Health Care, 16,* 95–100. doi:10.1080/028134398750003241

Spengler, P. M., White, M. J., Ægisdóttir, S., Maugherman, A. S., Anderson, L. A., Cook, R. S., . . . Rush, J. D. (2009). The meta-analysis of clinical judgment project: Effects of experience on judgment accuracy. *The Counseling Psychologist, 37,* 350–399. doi:10.1177/0011000006295149

Strasser, J., & Gruber, H. (2004). The role of experience in professional training and development of psychological counsellors. In H. P. A. Boshuizen, R. Bromme, & H. Gruber (Eds.), *Professional learning: Gaps and transitions on the way from novice to expert* (pp. 11–28). Dordrecht, The Netherlands: Kluwer Academic.

Weiss, D. J., Brennan, K., Thomas, R., Kirlik, A., & Miller, S. M. (2009). Criteria for performance evaluation. *Judgment and Decision Making, 4,* 164–174.

Weiss, D. J., & Shanteau, J. (2003). Empirical assessment of expertise. *Human Factors, 45,* 104–116. doi:10.1518/hfes.45.1.104.27233

Williams, C. A., Haslam, R. A., & Weiss, D. J. (2008). The Cochran–Weiss–Shanteau performance index as an indicator of upper limb risk assessment expertise. *Ergonomics, 51,* 1219–1237. doi:10.1080/00140130802087094

Witteman, C. L. M., Renooij, S., & Koele, P. (2007). Medicine in words and numbers: A cross-sectional survey comparing probability assessment scales. *BMC Medical Informatics and Decision Making, 7,* 1–8. doi:10.1186/1472-6947-7-13

Witteman, C. L. M., & Van den Bercken, J. H. L. (2007). Intermediate effects in psychodiagnostic classification. *European Journal of Psychological Assessment, 23,* 56–61. doi:10.1027/1015-5759.23.1.56