

People Who Judge People

DAVID J. WEISS^{1*}, JAMES SHANTEAU² and PRISCILLA HARRIES³

¹*California State University, Los Angeles, California, USA*

²*Kansas State University, Kansas, USA*

³*Brunel University, West London, UK*

ABSTRACT

Experts who judge people usually provide opinions. It can be challenging to evaluate the professional performance of those experts, because for many domains there is no applicable external standard against which to verify the opinions. We review traditional methods for assessment and propose the purely empirical CWS approach as an alternative. Expert judgment entails discriminating among the various stimuli within the domain as well as being consistent when judging similar stimuli. We combine observed measures of these two components to form a ratio that we call the CWS index of expertise. We demonstrate the value of the index in an analysis of prioritization judgments made by occupational therapy students before and after they received specific training. The students' CWS scores improved considerably after training. The promise of the index as a selection tool is supported by the positive correlation of pre-training scores with both post-training scores and with course grades. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS expertise; prioritization; professional performance assessment; clinical judgment

INTRODUCTION

In the courtroom and in the clinic, in the office and in social settings, people make judgments about other people. The judgments may be assessments of current states, such as health status, ability, attractiveness, or veracity, or they may encompass predictions of future behaviors. People assess achievements by other people, for example by assigning grades or by voting for Academy Awards. These judgments can have dramatic consequences for the target person, so whether they are done well is an important issue. It is worrisome, therefore, that judgments about people are usually opinions. Opinions expressed within professional settings

* Correspondence to: David J. Weiss, Department of Psychology, California State University, Los Angeles, CA, USA.
E-mail: dweiss@calstatela.edu

Contract/grant sponsor: U. S. Air Force Office for Scientific Research; contract/grant number: #FA9550-04-1-0230.

are referred to as clinical judgments, but that impressive-sounding label cannot mask the fact that they are opinions. Are there people whose opinions are especially worthy of respect?

One of the enduring themes in Paul Meehl's writings was the superiority of actuarial methods over clinical judgment (Meehl, 1954; Dawes, Faust, & Meehl, 1989). Judges tire, they behave idiosyncratically, and in general they fail to follow the rules of the field consistently. A formula, on the other hand, remains true to its stated principles. Meehl's tactic to demonstrate the superiority of actuarial methods was to show that the formula predicted observed outcomes better than did predictions made by even the "best" professionals.

This analytic approach has great appeal for assessing expertise. If we knew the right answers, it would be easy to evaluate the judgmental performance of someone who claimed expertise in a domain. The best judge would be the one whose opinions were closest to the right answers. Ideally, one could verify expressed clinical judgments, and if someone made consistently accurate evaluations, then we would be comfortable in attributing expertise to that person. For example, assessment of violence risk among incarcerated offenders can be compared to post-release violent behavior (Swets, Dawes, & Monahan, 2000).

Unfortunately, for many professions of interest, no objective criterion that would permit verification is available. How do we know whether assigning custody to a particular parent was the right decision? Can we tell whether a rejected applicant would have been successful? Was the patient truly depressed? In the absence of known "correct answers," alternative means of assessing the expertise of the judge are needed.

In this paper, we first review non-empirical methods for identifying experts. Adherents of the traditional perspective have little need for verification; a judgment made by an expert is presumed to have been made expertly. We then discuss empirical methods that have appeared in the literature over the past few decades. Proponents of these approaches have recognized the problem of verification, but in our view their solutions have major limitations.

Our own preferred approach is a purely empirical one that looks only at expressed judgments. We call this methodology CWS¹ (Weiss & Shanteau, 2003), and have proposed the CWS index as a general solution to the problem of identifying expertise in the absence of external standards. That methodology is recounted here and illustrated with a study that explored expertise among occupational therapists. In shifting the focus from identifying experts to identifying expertise, we emphasize performance rather than capability. When we acknowledge someone as an expert, we are employing a shorthand phrase that strictly speaking is applicable only to the class of behaviors that have been assessed.

NON-EMPIRICAL APPROACHES

In everyday life, finding an expert is usually seen as a simple problem. The "Yellow Pages" in the telephone book provide lists of people who establish their expertise by listing their names under professional categories. The putative expert is automatically presumed to judge proficiently, since that is her job. When a layperson seeks supporting evidence before hiring a candidate, the evidence is likely to consist of three kinds of information: experience, titles, and personal recommendations.

Experience

A judge's number of years of job-relevant experience is often used as a surrogate measure of expertise. That is, judges with many years of experience are classified as "experts," while others with fewer (or no) years of

¹W and S are initials of the first and second authors of this paper, the creators of the index, while C commemorates the statistician William Cochran, who many years earlier had proposed a similar formulation to evaluate a response instrument (Cochran, 1943).

experience are labeled “novices.” Experience is an objective index that is easy to measure. Unfortunately, while the argument can be made that experts usually have considerable experience, the converse does not necessarily follow. There are many examples of professionals with considerable experience who are less than expert (Oskamp, 1962). Such individuals may even work in the same setting as top experts, but they never rise to the performance levels that mark the true expertise.

For example, Goldberg (1968) asked clinical psychologists with varying degrees of experience to diagnose psychiatric patients. He found no relation between years of experience and accuracy of the diagnoses; however, the confidence of clinicians in their diagnoses did increase with experience.

Although there are undoubtedly instances where a positive relationship exists between experience and performance, there is little reason to expect this relationship to apply universally. At best, experience is an uncertain predictor of degree of expertise. At worst, experience may reflect seniority and nothing else.

Titles

In many professions, the practitioner acquires a title that connotes skill. We are more likely to have confidence in someone who is a certified professional. Titles often accrue as a result of education. A doctor with an MD plus a PhD may be an “advanced specialist,” or a therapist with a master’s degree may be a “marriage and family counselor.” Generally, it seems safe to say that a titled individual is more likely to be an expert than someone who does not bear the title. But the relationship is not universal. Titular hierarchies are usually unidirectional; people move up the ladder, but rarely down. Once certified, the recipient is accredited for life. Even if the skill level of the individual suffers a serious decline, the title remains.

Acclamation

A very plausible approach has been to rely on identification of “experts” by subject matter experts (SMEs). That is, professionals in a domain are asked whether someone else is, or is not, an expert. When there is agreement among SMEs, that person is then labeled an expert “by acclamation.” For example, in an analysis of livestock judges, Phelps (1977) asked professionals in the field to nominate those they thought were the best. From the responses, Phelps selected four top livestock judges who served as “experts” in the investigation.

The use of acclamation is an attractive strategy. It is unlikely that everyone working in a field would identify the same unqualified candidate as an expert. If they agree, it seems safe to assume that the agreed-upon person is an expert. One problem with this approach is a “popularity effect”—someone better known by peers is more likely to be identified as an expert. Meanwhile, someone outside the peer group is unlikely to be viewed as an expert—although that person may be at the cutting edge of new knowledge. Indeed, those who make new discoveries in a field are frequently challenged by peers at the time of the breakthrough (Kahneman, 2003; Pledge, 1959).

EMPIRICAL APPROACHES TO EVALUATING EXPERTISE

Clinical judgment calls for the judge to integrate the information inherent in the target person, along with relevant background information. The expert selects the dimensions that merit attention, and then evaluates and combines the target person’s scores on those dimensions. The response is a projection that places the target person either on a continuum or in a defined category, as determined by the task. Although process analyses are sometimes feasible, empirical assessment generally implies that the only information available for performance evaluation is contained within overt responses.

Quantitative assessment

Einhorn (1972, 1974) proposed identifying experts by virtue of two forms of reliability measures. This was a pioneering approach that rested entirely upon quantitative analysis of behavioral data. A candidate could be evaluated using two necessary conditions for expertise. The first is intra-individual reliability. That is, an expert's judgments should be *consistent* over time. Conversely, inconsistency would be *prima facie* evidence that the person is not an expert. Similarly, Bolger and Wright (1992) proposed that assessing consistency when no gold standard of objective validation is available.

One limitation of this approach is that high consistency can be obtained by someone following a simple, but incorrect, rule. As long as the rule is followed precisely, the person's behavior will exhibit high consistency. For example, a physician treating patients complaining of chest pain might recommend surgery for all patients over 60 and drugs for all patients under 60. Although the doctor's judgments would be perfectly consistent, it is unlikely that such a policy would be optimal. Thus, internal consistency is a necessary condition—an expert could hardly behave randomly—but it is not sufficient for defining expertise.

Einhorn's (1972, 1974) other necessary condition is *consensus* between experts. That is, the experts in a given field should agree with each other (Ashton, 1985). If they do not, this suggests that at least some of the would be experts are not really what they claim to be.

On the surface, consensus appears to be a compelling property for experts. After all, patients feel comfortable when two or more experts (such as medical doctors) agree about which procedure to follow. When the experts disagree, on the other hand, patients feel uncomfortable committing to a course of action. Goldberg and Werts (1966) employed this logic to disparage the profession of clinical psychology, reporting a correlation less than 0.40 for the agreement among a set of experienced practitioners evaluating MMPI profiles.

While Einhorn (1972, 1974) proposed that responses from different experts should be highly correlated, Uebersax (1993) and Uebersax and Grove (1990, 1993) sought agreement in the latent structure underlying judgments. The idea is that experts should be assessing the same thing, although they may have unique perspectives that lead them to evaluate differently.

Our view is that consensus in any guise is an inappropriate criterion for expertise (Shanteau, 2001; Weiss & Shanteau, 2004). One problem with consensus is that the agreement can result from overly hasty closure, as in *groupthink* (Janis, 1972) where pressure to conform inhibits airing of dissident views.

There is a deeper concern as well. Bertrand Russell phrased it succinctly: "Even when the experts all agree, they may well be mistaken" (Russell, 1993). Such historical examples as phrenology in medicine and the Rorschach test highlight the danger of reliance upon consensus as a basis for determining truth (Gardner, 1957). Generally, the problem is that workers in the domain may reach agreement without understanding the underlying mechanism.

The inference of expertise from consensus is a logical error. To be sure, when people have the correct solution to a problem, their answers must agree; but the converse does not logically follow. Agreement does not imply correctness.² Russell's concern is that of a logician as well as a social critic.

Discrimination

Hammond (1996) has argued that experts should be able to make fine distinctions. That is, they should be attentive to subtle differences that are relevant in particular contexts. This ability was described as "selective

²A popular book (Surowiecki, 2004) advances the provocative thesis that, under specified conditions, the mean judgment of a group of people will often be more accurate than that of the most knowledgeable individual. Our view is that this version of the "law of large numbers" may be applicable when little true differential expertise exists, so that the so-called experts really do not know much more than other folks. We are not surprised that the crowd is wiser than the professional when it comes to predicting future economic or political events.

attention” by Abdolmohammadi and Shanteau (1992), that is, making fine discriminations between similar stimulus cases.

Discrimination cannot be blind. It must be driven by the specific decision situation or task. People often have difficulty separating attributes, leading to halo effects. Weiss and Shanteau (2000) examined the ratings for Technical Merit and Presentation given to contestants at the 1999 U. S. figure skating championships. For the Ladies’ Singles Event, the average correlation between ratings for the two attributes was 0.95, with the individual judges exhibiting correlations between 0.91 and 0.98. However, there was no way for us to distinguish true association in the two aspects of skating performance from inability of the judges to maintain distinct criteria.

It is possible to be discriminating—but on the wrong attributes. For instance, looking at the age of potential job candidates allows for fine distinctions to be made between applicants. However, age is seldom a predictor of on-the-job performance. Moreover, it is a legally inappropriate criterion for hiring, promotion, etc. Still, that does not prevent some personnel managers from using age as a factor in evaluating candidates (Nagy, 1981).

Confounded assessment can be difficult to detect. Imagine a culture that needs to estimate weights of people, perhaps in order to establish occupancy limits for elevators, but has no tool for this measurement. However, suppose the hypothetical culture does have a good technology for measuring heights of people, and uses those measures instead. Most of their elevators would not fall from misestimating the load, because height and weight are positively correlated. They might never realize that a more effective measure could be achieved.

CWS approach

Weiss and Shanteau (2003) argued that expert judgment requires different responses to different stimuli (*discrimination*) while responding similarly to similar stimuli (*consistency*). These two properties each are necessary, but not sufficient, for expertise. The additional sufficient property is *validity*, which cannot be evaluated without “ground truth.” Only validation can guarantee that the judge is performing the assigned task correctly. However, assessing validity requires a criterion measure, and our concern is with situations in which no such measure is available.

We combine empirical measures of discrimination and consistency to construct a ratio, yielding the index we call CWS. While the ratio format and the consequent tradeoff between discrimination and consistency are new ideas, the two components of the index are familiar aspects of expertise. Using empirical criteria avoids the circularity inherent in approaches that rely on expert knowledge to identify expertise.

The intuition underlying the index is that a good measuring instrument necessarily has a high CWS ratio, that is, exhibits high discrimination and high consistency. Like a good measuring instrument, an expert judge must be both discriminating and consistent. It is fairly easy to display either quality using a simple response strategy, one that requires little knowledge of the task. One can show discrimination simply by generating a wide variety of responses; one can exhibit consistency by responding similarly to all cases. But adopting either of these strategies alone guarantees that the other quality will be lost. To be able to incorporate both qualities simultaneously, on the other hand, requires accurate and consistent assessment of stimuli, the essence of expert judgment.

The equation that defines the index incorporates both key properties in a ratio format, providing that assessed expertise will be high when the judge discriminates effectively, and will be reduced if the judge is inconsistent. The trade-off inherent in a ratio ensures that one cannot appear to behave expertly merely by following either of the simplistic strategies that emphasizes one property at the expense of the other.

To implement the measure, we ask candidates to evaluate a common set of stimuli. The estimates of discrimination and inconsistency are dependent on the particular stimuli that were judged, so comparisons must be based only on responses to stimuli judged by all candidates. To afford the estimate of inconsistency,

some, if not all, of the stimuli must be evaluated more than once. Each judge's data are analyzed on a single-subject basis. We analyze the responses in two ways: the first to estimate discrimination and the second to estimate inconsistency. The CWS value for an individual is computed from estimates of the two key quantities.

$$\text{CWS} = \frac{\text{Discrimination}}{\text{Inconsistency}} \quad (1)$$

Discrimination means that as the stimulus changes, the evaluation changes accordingly. Inconsistency means that repeated evaluations of the same stimulus differ considerably. Both of these constructs may be operationalized as statistical dispersion, since it is the extent of differences that is crucial. Any summary measure of dispersion—variance, standard deviation, or mean absolute deviation—might yield plausible CWS ratios. We have relied upon variance measures (literally, mean squares) in our work so far, and consider them the default option. We use MS_{Stimuli} as the estimate of discrimination and $MS_{\text{Replications}}$ as the measure of inconsistency. Variances, with their heavy weighting of large discrepancies, have traditionally been employed by statisticians to capture precision of measurement (Grubbs, 1973).

We previously illustrated (Weiss & Shanteau, 2003) the merits of the approach by reanalyzing a study of physicians estimating the probability that patients had chronic heart failure (Skånér, Strender, & Bring, 1998). We also reanalyzed data from students and personnel analysts rating job applications characterized by relevant or irrelevant information (Shanteau & Nagy, 1984). The professionals were able to ignore age and appearance, which are legally irrelevant, but the students could not. A guide for the use of the CWS index may be found at <http://www.k-state.edu/psych/cws/pdf/using-cws.pdf>.

APPLICATION TO OCCUPATIONAL THERAPY

We now show how CWS can capture acquisition of expertise among occupational therapists whose task is to prioritize patients for community mental health services (Harries, 2004). While this study was not designed with CWS in mind, we offer it as a new illustration because we performed post hoc analyses that brought out previously unappreciated features of the data.

The domain of concern for the occupational therapist pertains to the patient's capacity to engage in self-care, work, or leisure occupations. Occupational therapists view effective occupational engagement as integral to health. In Britain, referrals received by community health teams have to be prioritized, as the demand for services far exceeds the available provision. Referrals are prioritized in order to ensure that services are provided to the most needy, with the proviso that the patient will be able to benefit from the treatment. There is no gold standard for prioritization. The judgments have to integrate both medical and circumstantial factors with the patient's occupational dysfunction.

METHOD

Participants

Forty occupational therapists from England, Scotland, and Wales constituted the experienced group. Five were male, 35 were female, 39 were Caucasians and 1 was African Caribbean. Thirty-five had worked for more than 3 years in their community. The majority of their work was carried out in deprived neighborhoods.

Thirty-seven students in the final year of an urban British university's curriculum in occupational therapy constituted the student group. All were white females and their mean age was 23.

Community Occupational Therapy Mental Health Referral Form (Adult Mental Health Services)

Client's name	Mr xx	Address	Within catchment area.
Age	48	D.o.b.	xx/xx/xx
Date of referral	xx/xx/xx (Recent)	Telephone	(xxxx) xxx xxxx
Name of referrer	GP	GP	Dr xx
Consultant	Dr xx		
Diagnosis	Anxiety		Five year history.
Current living situation	Home with family		
Reason for referral	Psychological and physical disabilities. Functional assessment needed to identify level of support required.		
Other services involved	Counsellor		
Any known history of violence?	Physically abusive		
Is the client aware of the referral?	Yes.		

Low priority High priority

Figure 1. Sample profile, with visual analog response scale.

Design

In the first phase of the study, the experienced participants were asked to prioritize individually a set of 120 referrals presented in the form of profiles on single pages, similar to actual referral forms used in England (Figure 1). There were 90 different profiles plus 30 repetitions (to check for consistency). Each profile conveyed nine pieces of information (cues) about the hypothetical client. A large set of potential cues was previously ranked by an independent panel of eight experienced therapists, and the nine cues they ranked highest were used in the profiles. The cues are listed here in the order of their contributions to the judgments of the experienced therapists as assessed by beta weights. The cues were (1) reason for referral, (2) history of violence, (3) diagnosis, (4) living situation, (5) support available, (6) referrer, (7) gender, (8) age, (9) length of history. The number of levels of the cues varied from two (for gender) to eight (for reason for referral). To keep the design to manageable size, profiles were constructed by random selection among combinations, with the constraint that the cues remained uncorrelated across the set.

Responses were made by putting a mark on a 112 mm line at the foot of each referral. One end of the line was labeled low priority and the other end was labeled high priority (a visual analog scale). The response was the number of centimeters, recorded to two decimal places, between the left end of the line and the mark.

Prior to training, the students judged 72 profiles using the same response mode. Eighteen of the profiles they judged twice were also in the set judged by the experienced therapists. Responses to those common profiles were analyzed for the CWS comparison of experienced therapists versus students.

The training was a 1-hour session during which the students received written and graphical instruction regarding a prioritization strategy derived from a lens model analysis of the experienced judges' policies. The gold standard policy was derived using a Ward's cluster analysis (Cooksey, 1996) of the experienced therapists' judgments. A subgroup ($n=9$) whose policies followed the occupational therapy profession's recommended methods was identified. The mean of their weightings was defined as the gold standard policy. The policy assigned high importance to the reason for referral, moderate importance to diagnosis, and lesser importance to history of violence. Other cues were not given appreciable weight.

After training, the students judged 48 profiles. Ten of the profiles from the initial set were included in this second test. Responses to those common profiles, each judged twice, were the basis of the CWS analysis of student performance before and after training.

RESULTS

Although most stimulus sets used for testing before and after training differed, we confine the discussion of training effects to the 20 profiles that were common to both sets. Two interesting results emerged from the comparison. First, the mean CWS rose from 4.84³ (84.3% CI⁴ = 3.84–5.96) to 20.34 (84.3% CI = 11.92–30.40) after training. The increase meant that on average, students were able to discriminate more consistently following training; on an individual basis, the CWS ratio rose for 29 of the 36⁵ students. What the trainees apparently learned was to use the prioritization strategy taught in the writeup. Prior to instruction, the correlation between the judgments of the students and the lens model strategy based on group representation of the experienced therapists was 0.23 (95% CI = –0.10 to 0.52). After training, the correlation of students' judgments with those of the experienced therapists rose to 0.70 (95% CI = 0.49–0.83).

Another interesting result highlights the promise of CWS as a selection tool. The students' CWS scores prior to training were positively correlated with their post-training CWS scores ($r=0.65$, 95% CI = 0.41–0.80) and with course grades ($r=0.52$, 95% CI = 0.24–0.72). So even though few of the students followed the recommended strategy prior to training, those who initially were better able to discriminate and were consistent were likely to be the students who were more successful. They were also favorably positioned to transfer that ability when using the information provided during training.

Graphical analysis

By examining the plots of individual judges in Figure 2, we can gain insight into what the CWS ratio captures even though we do not know the “right” answers. There were 18 profiles that were judged twice by 38⁶ experienced occupational therapists and by 37 students prior to training. For each judge, the plot shows the two responses to each of those profiles. The highest scoring experienced therapist, #40, whose CWS

³Statistical processing of CWS ratios was carried out using square roots, as recommended by Weiss and Edwards (2005). The reported mean CWS is the square of the sum of the individual square roots. In comparing individual CWS ratios, square roots of the values reported convey their relative magnitudes appropriately.

⁴We follow Payton, Greenstone, and Schenker (2003) in employing 84.3% confidence intervals rather than the usual 95% confidence intervals when comparisons are intended. The 84.3% confidence level allows comparisons with a Type I error rate of 0.05. Here the lack of overlap implies that, at the 0.05 level, there is a significant difference between the mean CWSs.

⁵One student did not provide two replicates during the post-training evaluation. The CWS ratio could not be computed, so the student was omitted from the post-training analyses.

⁶Two of the experienced therapists did not evaluate all of the stimuli and accordingly were omitted from the CWS analyses.

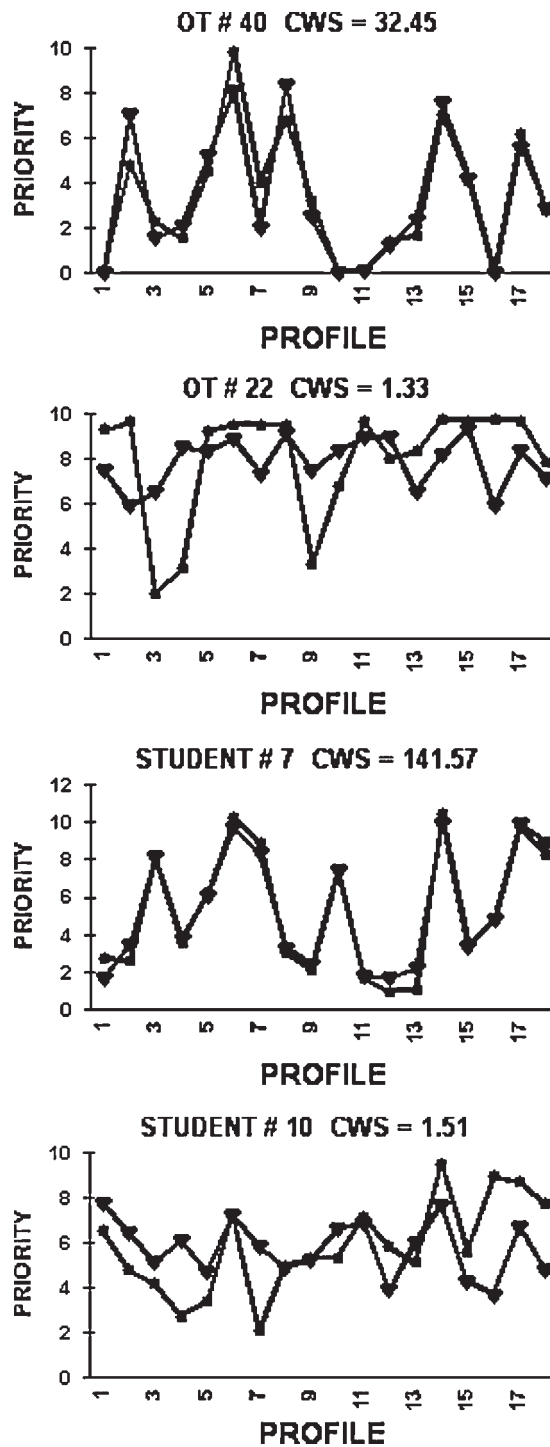


Figure 2. Prioritizing recommendations for two replications of 18 profiles, for each of four judges (two experienced occupational therapists and two students).

is 32.45, is quite discriminating and consistent. On the other hand, the lowest scoring experienced therapist, #22, has a much lower CWS of 1.33; #22 showed little discrimination on the first set of judgments and was thereby inconsistent as well. This is not a pattern that an expert should exhibit. Student #7 is an outstanding judge according to the CWS criterion ($CWS = 141.57$), showing both discrimination and consistency. This was the highest scoring student, both before and after training. However, the student's pattern of responses does not match that of experienced therapist #40, and the CWS criterion cannot tell us who is more accurate. The lowest scoring student, #10, is neither discriminating nor consistent.

The results from the CWS analysis suggest that experience need not guarantee expert performance. The mean CWS obtained for the experienced group was 6.35 (84.3% CI = 5.14–7.68), and surprisingly, the mean CWS for the students was 7.10 (84.3% CI = 4.63–10.04). While the students did not weight the information in the profiles in the same way that the experienced therapists did, they were on average no less capable of consistently discriminating among cases.

DISCUSSION

Although the Harries (2004) study was designed prior to the development of CWS methodology, nevertheless it proved feasible to apply the analytic procedure to subsets of the data and thereby illuminate the original findings. Perhaps the most interesting result was the promise of CWS as a screening tool. Students who were initially better at consistently discriminating among profiles also proved to be better students in terms of class grades and in how they learned to emulate the strategy derived from that employed by the experienced therapists. CWS also demonstrated that most students did improve after training. Moreover, that demonstration did not require assumptions about correct judgments. The CWS assessment ignores the question of whether the gold standard policy taught to the students was truly optimal.

It was surprising that the experienced therapists as a group were not better judges, at least in CWS terms, than the students prior to training. Some therapists, such as OT #22 (see Figure 2), are simply not very good at this prioritization task. This observation is consistent with similar results summarized by Shanteau (1999), who remarked that professionals in domains such as clinical psychology and medicine are often neither internally consistent nor in agreement with their colleagues, especially in comparison to professionals in auditing, weather forecasting, and livestock judging. Clinical assessments, wherein people judge people, are inherently challenging.

CONCLUSION

There has been intense debate over what expert clinicians do, and over the proper methodology for analyzing their thought processes (Anderson, 1972). If we knew what ought to be done, an alternative to assessing expertise in situations where we cannot ask about correct outcomes would be to see whether a candidate's thought processes were appropriate. Unfortunately, employing a methodology of this type requires the strong assumption that the evaluator knows what ought to be done.

In contrast, the CWS approach does not require a cognitive processing analysis of the judgmental process in order to assess expertise. The CWS analysis does not incorporate any information regarding the stimuli; because correct answers are not known, stimuli are treated as purely nominal. An operational advantage of this neutrality is that the evaluation can be carried out for clinicians who are judging real people. In contrast, methods that delve into thought processes generally rely upon analyzing judgments of paper people, who have characteristics that are systematically varied so that analysis is feasible. With the CWS approach, all that is required is a set of people to be judged. These can be real people about whom no knowledge need be presumed.

An important element of a judge's expertise is extracting accurate information from the stimulus environment (Dawes, 1979), but those who are evaluating the judges cannot be expected to share this specialized expertise. In general, the analyst does not know how to deconstruct the judgmental task to set up codable inputs (Dawes & Corrigan, 1974). Therefore, the CWS methodology, which presumes only that the stimuli can be identified, avoids a strong assumption needed for cognitive process analysis.

Thinking about CWS from what we take to be Meehl's perspective suggests that the new index will be most useful when the behavior of the target people is difficult to decompose, when there is not a clear understanding of the relevant cues. If we wish to evaluate the work of experts where there is no objective correlate of good performance, professionals such as figure skating judges, movie critics' or journal reviewers, then CWS may offer the most powerful tool available. It is not so much these professionals evaluate holistically—we do not know how to tell whether they do—, it is that we have only vague comprehension of the judgmental tasks. Looking for the necessary features, discrimination and consistency, in their judgments, may be as much as we can do.

Judgments about people are inherently difficult. As stimuli to be judged, real people have the undesirable property that they care about the expressed judgment. They may try to look good and in doing so may provide false reports about their attitudes, history, and capabilities. In medical settings, on the other hand, some people may be motivated to look bad, exaggerating their symptoms for social or financial reasons. Even when the target person tries to cooperate by disclosing honestly, there may be limits on self-knowledge.

Meehl's (1954) challenge to clinical prediction did not meet with universal approval from clinicians (Holt, 1961). Similarly, we have encountered experts who have expressed opposition to the application of objective methods in a subjective domain. This resistance is understandable, as one whose expertise has already been established has little incentive to cooperate in the exploration of a new tool that has the potential to question the existing hierarchy.

How would Meehl's approach do from a CWS perspective? For any task in which people are judged, including diagnosis of current status or prediction of future action, evaluations made according to a formula, for example, multiple regression or expected value, would yield high performance according to CWS. The use of any deterministic formula would guarantee that every judgment of a given person would be the same, so there would be no inconsistency at all. Whether the formula is optimal, or even sensible, would not affect the CWS ratio.

This line of thinking exposes an inherent limitation of our methodology, one that we have previously labeled "the catch" (Weiss & Shanteau, 2003); a person may score well while doing the wrong thing. Therefore, high CWS cannot guarantee expert performance. In practice, the catch has not caused problems in the previous studies in which we have empirically distinguished experts from non-experts.

CWS is capable of capturing distinctions between relevant and irrelevant stimulus information. In a previous reanalysis (Weiss & Shanteau, 2003) of Nagy's (1981) data, we demonstrated that experienced personnel selectors exhibited large CWS ratios for experience and education of job applicants, but had values clustering around 1.0 for age, attractiveness, and gender. This is an appropriate pattern, as the latter attributes are legally irrelevant and should not be considered. In contrast, student participants showed large CWS values for all five cues.

Because in this case the researchers had external knowledge about the stimuli, we were able to evaluate the inappropriately high CWS values on irrelevant attributes as evidence of lack of expertise on the part of the students. Similarly, the low CWS values shown by the experienced analysts on those irrelevant attributes were considered supportive of their expertise. Of course, ground truth allows strong inferences. If we had no background knowledge, we might have incorrectly inferred that the students were behaving more expertly than the professionals because they were consistently extracting more information from the cues.

In many studies of people doing their jobs, detailed knowledge will not be available. At a minimum, we recommend that the researcher ascertain that the stimuli used in the study are relevant and span the range encountered in everyday work. This advice is not specific to CWS, but applies to any empirical assessment.

Generally, consultation with domain experts is the most practical route to effective stimulus selection. In our view, this is where subject matter experts are most valuable in the assessment process.

A pragmatic limitation of the CWS methodology is that in order to measure consistency in judging people, at least some of the target people have to be assessed by the expert more than once. In real life, people change over time, so that judging them repeatedly may in a strict sense be impossible. For research purposes, people as stimuli can be made repeatable by using videotape or other recorded information. Even then, people may be so memorable that special care is needed to keep the judgments independent.

CWS ratios cannot be compared across testing situations; the ratio depends upon the particular stimuli that were involved. All empirical assessment of expertise is subject to this limitation, though the conditionality is not always acknowledged explicitly. For example, a baseball player's batting average depends upon the quality of the pitching and defense with which he contended. Comparing performance statistics across different levels, say between major and minor leagues, is dubious.

As illustrated here, CWS was useful in evaluating expertise in health-care diagnosis. We have also applied CWS to reanalyzes of studies of cardiology, nursing judgment, and pathology. Such diagnoses are often guided by medical decision support systems. While support systems can certainly inform the diagnostician, they depend vitally on the skills of individual health-care specialists. Rarely there are clearly defined best ways to utilize such systems, especially at the level of the individual patient.

CWS offers an approach that can be used to enhance performance, with or without the assistance of decision support systems. People can be taught to increase their CWS scores, which means they will discriminate more precisely and be more consistent. The issues addressed by CWS are of importance not only to researchers, but also to health-care workers and to a public that relies heavily upon the judgments of experts.

ACKNOWLEDGEMENT

Preparation of this manuscript was partially supported by grant #FA9550-04-1-0230 from the U. S. Air Force Office of Scientific Research.

REFERENCES

- Abdolmohammadi, M. J., & Shanteau, J. (1992). Personal characteristics of expert auditors. *Organizational Behavior and Human Decision Processes*, *58*, 158–172.
- Anderson, N. H. (1972). Looking for configularity in clinical judgment. *Psychological Bulletin*, *78*, 93–102.
- Ashton, A. H. (1985). Does consensus imply accuracy in accounting studies of decision making? *Accounting Review*, *60*, 173–185.
- Bolger, F., & Wright, G. (1992). Reliability and validity in expert judgment. In G. Wright, & F. Bolger (Eds.), *Expertise and decision support* (pp. 47–76). New York: Plenum Press.
- Cochran, W. G. (1943). The comparison of different scales of measurement for experimental results. *Annals of Mathematical Statistics*, *14*, 205–216.
- Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and application*. New York: Academic Press.
- Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*, 571–582.
- Dawes, R. M., & Corrigan, B. M. (1974). Linear models in decision making. *Psychological Bulletin*, *81*, 95–106.
- Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*, 1668–1674.
- Einhorn, H. J. (1972). Expert measurement and mechanical combination. *Organizational Behavior and Human Performance*, *7*, 86–106.
- Einhorn, H. J. (1974). Expert judgment: some necessary conditions and an example. *Journal of Applied Psychology*, *59*, 562–571.

- Gardner, M. (1957). *Fads and fallacies in the name of science*. New York: Dover.
- Goldberg, L. R. (1968). Simple models or simple processes? Some research on clinical judgments. *American Psychologist*, 23, 482–496.
- Goldberg, L. R., & Werts, C. E. (1966). The reliability of clinicians' judgments: A multitrait-multimethod approach. *Journal of Clinical Psychology*, 30, 199–206.
- Grubbs, F. E. (1973). Errors of measurement, precision, accuracy and the statistical comparison of measuring instruments. *Technometrics*, 15, 53–66.
- Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. New York: Oxford University Press.
- Harries, P. A. (2004). *Occupational therapists' judgement of referral priorities: Expertise and training*. Unpublished doctoral dissertation, Brunel University, West London, England.
- Holt, R. R. (1961). Clinical judgment as a disciplined inquiry. *The Journal of Nervous and Mental Disease*, 133, 389–382.
- Janis, I. L. (1972). *Victims of groupthink*. Boston: Houghton-Mifflin.
- Kahneman, D. (2003). Experiences of collaborative research. *American Psychologist*, 58, 723–730.
- Meehl, P. E. (1954). *Clinical versus statistical prediction*. Minneapolis: University of Minnesota Press.
- Nagy, G. F. (1981). *How are personnel selection decisions made? An analysis of decision strategies in a simulated personnel selection task*. Unpublished doctoral dissertation, Kansas State University, Manhattan, Kansas.
- Oskamp, S. (1962). The relationship of clinical experience and training methods to several criteria of clinical prediction. *Psychological Monograph*, 76 (28, Whole No. 547), 1962, p. 27.
- 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 6p. Available online at insectscience.org/3:34.
- Phelps, R. H. (1977). *Expert livestock judgment: A descriptive analysis of the development of expertise*. Unpublished doctoral dissertation, Kansas State University, Manhattan, Kansas.
- Pledge, H. T. (1959). *Science since 1500*. New York: Harper Torchbooks.
- Russell, B. (1993). *Sceptical essays*. London: Routledge.
- Shanteau, J. (1999). Decision making by experts: the GNAHM effect. In J. Shanteau, B. Mellers, & D. Schum (Eds.), *Decision research from Bayes to normative systems: Reflections on the contributions of Ward Edwards*. Norwell, MA: Kluwer Academic Publishers.
- Shanteau, J. (2001). What does it mean when experts disagree? In E. Salas, & G. Klein (Eds.), *Linking expertise and naturalistic decision making* (pp. 229–244). Mahwah, NJ: Erlbaum.
- Shanteau, J., & Nagy, G. F. (1984). Information integration in person perception: Theory and application. In M. Cook (Ed.), *Issues in person perception* (pp. 48–86). London: Methuen.
- Skånér, Y., Strender, L., & Bring, J. (1998). How do GPs use clinical information in the judgements of heart failure? *Scandinavian Journal of Primary Health Care*, 16, 95–100.
- Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*. New York: Doubleday & Company.
- Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, 1, 1–26.
- Uebersax, J. S. (1993). Statistical modeling of expert ratings on medical treatment appropriateness. *Journal of the American Statistical Association*, 88, 421–427.
- Uebersax, J. S., & Grove, W. M. (1990). Latent class analysis of diagnostic agreement. *Statistics in Medicine*, 9, 559–572.
- Uebersax, J. S., & Grove, W. M. (1993). A latent trait finite mixture model for the analysis of rating agreement. *Biometrics*, 49, 823–835.
- Weiss, D. J., & Edwards, W. (2005). A mean for all seasons. *Behavior Research, Methods, Instruments, and Computers*, 37, 677–683.
- Weiss, D. J., & Shanteau, J. (2000). *Consensus: The hobgoblin of little minds*. Paper presented at the OK Judgment and Decision Making Meeting, Oklahoma City, OK.
- Weiss, D. J., & Shanteau, J. (2003). Empirical assessment of expertise. *Human Factors*, 45, 104–116.
- Weiss, D. J., & Shanteau, J. (2004). The vice of consensus and the virtue of consistency. In C. Smith, J. Shanteau, & P. Johnson (Eds.), *Psychological investigations of competent decision making* (pp. 226–240). Cambridge, UK: Cambridge University Press.

Authors' biographies:

David J. Weiss received his PhD in experimental psychology in 1973 from the University of California, San Diego. He is a professor of psychology at California State University, Los Angeles.

James Shanteau received his PhD in experimental psychology in 1970 from the University of California, San Diego. He is a professor of psychology at Kansas State University.

Priscilla A. Harries received her PhD in human sciences in 2004 from Brunel University. She is the course leader for the MSc in occupational therapy in the School of Health Sciences and Social Care at Brunel University.

Authors' addresses:

David J. Weiss, Department of Psychology, California State University, Los Angeles, CA 90032, USA.

James Shanteau, Department of Psychology, Bluemont Hall 492, Kansas State University, Manhattan, KS 66506, USA.

Priscilla A. Harries, School of Health Sciences and Social Care, Brunel University, Uxbridge, Middlesex UB8 3PH, England.