SPECIAL SECTION

Empirical Assessment of Expertise

David J. Weiss, California State University, Los Angeles, California, and **James Shanteau,** Kansas State University, Manhattan, Kansas

The assessment of expertise is vital both in practical situations that call for expert judgment and in theoretical research on the psychology of experts. It can be difficult, however, to determine whether a judge is performing expertly. Our goal was to develop an empirical measure of expert judgment. We argue that two necessary characteristics of expertise are discrimination of the various stimuli in the domain and consistent treatment of similar stimuli. We combine measures of these characteristics to form a ratio we call the Cochran-Weiss-Shanteau (CWS) index of expertise. The proposed index was demonstrated using two studies that distinguished experts from nonexperts based on their judgmental performance. The index provides new insights into expertise and offers a partial definition of expertise that may be useful in a variety of theoretical and applied settings. Potential applications of this research include selection, training, and evaluation of experts and of expert-machine systems.

INTRODUCTION

All people depend on experts to make life safe (e.g., by providing basic resources) and interesting (e.g., by entertaining with music and art). Most people would claim to be experts, at least at something. But are they really expert? How is the claim to be substantiated?

Experts have often been identified by self-proclamation or acclamation by other experts as well as by experience, titles, and degrees. However, these methods can be misleading when searching for an expert. We prefer instead to cast the problem in empirical terms: An *expert* is someone who carries out a specified set of tasks expertly. Because it emphasizes behavior, this apparent tautology is not devoid of content. We propose to compare the job performance of candidate experts. In this paper we offer a new methodology for evaluating, on a relative basis, the degree of expertise demonstrated on a particular task.

At first glance, one might hope to evaluate expertise by looking at outcomes. The ideal is to correlate action with a *gold standard,* an unequivocally valid, universally accepted outcome measure that directly reflects the behavior under scrutiny. The expert surgeon's patients are more likely to survive than are those of the poor surgeon; the expert air traffic controller's planes are more likely to arrive safely. Survival and safe arrival seem to be relevant gold standards.

Where gold standards exist, there are well-established procedures for assessing expertise. When a judge makes dichotomous decisions, the correctness of which can be determined objectively, d' provides a measure of accuracy (Swets, 1986). For numerical responses, the Brier score (Brier, 1950) penalizes errors in relation to the square of their magnitudes.

The expert performance approach (Ericsson & Lehmann, 1996) has been used with considerable success in finding behavioral assessments that generalize and thus suggest expertise. Someone who excels when tested in the laboratory is likely to excel in other settings as well. A fast sprinter outruns slower counterparts under most conditions. A chess master will select superior

Address correspondence to David J. Weiss, Department of Psychology, California State University, Los Angeles, 5151 State University Dr., Los Angeles, CA 90032; dweiss@calstatela.edu. **HUMAN FACTORS,** Vol. 45, No. 1, Spring 2003, pp. XXX–XXX.

moves in unfamiliar positions. Reproducible success in controlled settings predicts success in real-world applications.

When it is clear that an outcome measure captures expertise, it is appropriate to use it as a means to identify the expert. A potential problem is that a process may be more complex than use of the obvious outcome measure presupposes. Would it be surprising if the "best" surgeons generated poor survival rates? If patients and surgeons were randomly paired, medical outcome might be an effective assessment tool, but selection biases can render the correlation meaningless. A test that scales surgeons according to survival rates among their patients might be capturing the ability to attract easy cases rather than true surgical skill. The obvious gold standard may be tarnished.

One must be very careful to select tasks for which meaningful comparisons are feasible. In the laboratory the investigator can ask doctors or trainees to diagnose cases for which correct designations are known (Ericsson & Smith, 1991). In the field one might compare the success rates of emergency room physicians when patients are assigned to the first available doctor (Ericsson & Lehmann, 1996). In contrast to most medical settings, in this case the assignment of patient to practitioner can be regarded as essentially random.

For many tasks at which experts make a living, no measurable outcome exists. How is one to know if the wine taster has judged accurately or if the professor has graded the essays well? Adherents of the expert performance approach would question the merits of studying such domains. Although there is no hint of an objective external criterion, we believe that some people do these tasks better than others and that people improve their performance. We would like our assessment scheme to include such expertise. We propose a more general approach to assessing expertise, one that looks at the behavior itself rather than its association with an outcome. We do not reject the idea of gold standards; they provide an ultimate solution when available. Our view is that they are not easy to find, and this is not a coincidence; experts are needed in precisely those domains where no correct answers exist (Gigerenzer & Goldstein, 1996).

Our proposed methodology is built on two premises. The first is that evaluative skill, or judgment, is the heart of expertise. Whatever an expert is called on to do entails recognition and evaluation of the crucial stimuli in the situation. We present a classification of expert tasks – a partial theory of expertise – in which other skills are superimposed on judgment. This classification is not yet supported by much evidence but represents a research agenda. The cornerstone is the measurement of judgmental competence, and it is in the judgment arena that we expect the most effective assessments. As other skills occasioned by additional task requirements overlay judgment, the contribution of that core component may be obscured, and our appraisals may correspondingly be less satisfactory.

Our second premise is that an expert judge tries to function as a measuring instrument. A measuring instrument accurately evaluates stimuli; that is what it is built to do. The instruments used in everyday practice accomplish technologically something that a human expert may have done professionally in the past (Hoffrage & Gigerenzer, 1998). The thermometer supplants sensitive hands; sonar replaces dowsers. The instrument is usually better than the human, at least for its limited purpose, in that it does not exhibit inconsistency caused by fatigue or bias. Everyone's instrument can be built to the same specifications. The instrument may become a true gold standard, but once that acceptance has taken place, there is little need to assess human experts on that task because their skill has become obsolete (Shanteau, 1995).

CATEGORIES OF EXPERTISE

Tasks that call for expertise can be divided into four categories: Expert judges award medals, audit companies, assign grades, or make diagnoses. Experts in prediction are the best at forecasting the weather, hiring an employee, recommending medical treatment, or advising whether parole is a worthwhile risk. Expert instructors train novices, develop computationally aided "expert systems," set criteria for testing, or mentor aspiring experts. Performance experts do something better than most people can do it; their task may be playing an instrument, fixing a car, shooting a basketball, or painting a landscape.

We argue that evaluative skill is the basic cognitive ability that characterizes all these areas of expertise. Whatever the task, therefore, the expert must attend to relevant aspects of the situation and decide what needs to be done. It is this common element, evaluation, that our index is designed to capture. What distinguishes the categories is what the expert must do after the evaluation has been carried out:

- Evaluation + qualitative or quantitative expression = expert judgment.
- Evaluation + projection = expert prediction.
- Evaluation + communication = expert instruction.
- Evaluation + execution = expert performance.

An expert judge classifies stimulus cases into appropriate groups or categories. The first decision a physician faces is whether or not the patient's condition is serious (i.e., conducts triage). If the condition is deemed serious, then the doctor identifies the disease. Next, a course of treatment must be selected. Each of these decisions – triage, diagnosis, and treatment – involves an evaluative judgment. To do this well, the judge must be able to maintain appropriate criteria across a set of cases.

The expert predictor has the challenge of incorporating evaluations into a projected future scenario. Changes in conditions that will occur in the future must be anticipated. To be an expert predictor, one must not only evaluate but also be able to extrapolate evaluations into an unobserved future environment. The penal expert, for example, must evaluate the current status of potential parolees and decide how they will fare when faced with the temptations of the outside world (of course, the temptations that an individual will face cannot be specified precisely; Swets, Dawes, & Monahan, 2000).

The expert instructor needs to be able to communicate judgment strategies to novices. The required skills include breaking down the process into comprehensible subunits, explaining the requisite steps, illustrating the appropriate behavior, observing student performance and providing feedback, and motivating students. An expert instructor, then, must be able to both evaluate and communicate. An expert critic requires similar skills.

The performance expert must add execution to the requisite evaluation. In general, motor skills (e.g., strength, coordination, dexterity, and stamina) as well as evaluation are required to exhibit performance expertise.

Although those who are expert in one category may be asked to serve in another, expertise is generally highly specific. A great surgeon (performance category) may make poor recommendations (prediction category) that do not consider the patient's personal values. A skilled teacher (instruction category) may do poor laboratory work (performance category). A great coach (instruction category) may ask players to execute maneuvers that the coach can envision but not execute (performance category). Thus expertise may not transfer; each category calls for different specific talents beyond the evaluative skills required of all experts.

Expertise is also domain specific. Michael Jordan could not hit the curve ball when he tried to be a professional baseball player. An expert weather forecaster has no claim to predicting the stock market. The evaluative skills, as well as the additional requirements of each category, are the result of specific domain abilities, training, and experience.

BEHAVIORAL ASSESSMENT

Our predecessor, Einhorn (1972, 1974), proposed two empirical criteria he deemed necessary for expertise. The first is intraindividual reliability: An expert's judgments should be consistent over repeated trials. Reasoning similarly, Bolger and Wright (1992) proposed the method of assessing reliability when no gold standard of objective validation is available. R. H. Ashton (2000) observed that there is not much evidence bearing on the reliability of experts' judgments.

High consistency can be obtained by someone's following a simple, but incorrect, rule. As long as the rule is followed precisely, the person's behavior will be consistent. Consistency is a necessary condition – an expert could hardly behave randomly – but, as Einhorn (1972, 1974) acknowledged, 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 (A. H. Ashton, 1985). If they do not, then it 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 doctors agree on diagnoses and recommendations. When the physicians disagree, however, patients feel uncomfortable in committing to a course of action.

Although consensus is likely when various experts are judging in accord with a common latent structure (Uebersax, 1993), our view is that it is an inappropriate criterion for expertise (Shanteau, 2001; Weiss & Shanteau, in press). The confusion has arisen because consensus is the basis for terminology. Constructs, such as the defining characteristics of a disease, must be shared by the linguistic community that employs them. Doctors need to agree on what is meant by a term such as *myocardial infarction.* However, identifying and interpreting a particular patient's symptoms calls for perceptual and integrative skills. The judgment depends on more than merely knowing what the diagnostic category entails. Perhaps a crucial symptom is hard to detect, so that only someone with superior vision or sense of smell notices it. Whether the judgment is correct cannot be determined by agreement among judges. Bertrand Russell phrased it succinctly: "Even when the experts all agree, they may well be mistaken" (Russell, 1993).

To be sure, when people have the correct solution to a problem, their answers must agree. However, the converse does not follow; agreement does not imply correctness. Russell's concern was that of a logician as well as a social critic.

THE CORE OF EXPERTISE

We propose that expert judgment must satisfy two essential criteria. These constitute necessary, but not sufficient, conditions for expertise. The first is that expertise calls for discriminating among the stimuli within the domain. The ability to differentiate between similar but not identical stimuli is a hallmark of expertise (Hammond, 1996). Second, we follow Einhorn's (1974) suggestion that internal consistency is a requirement of expertise. (Creativity may be a valued characteristic for a performance expert, generating "inspired inconsistency." However, our focus here is on expert judgment; inconsistency in evaluation is capricious or random and produces chaos.) Furthermore, we propose that although the two criteria are assessed separately, they are linked psychologically in that they trade off. Consider a judge who is urged to emphasize consistency, as might be the case if an internist were asked to triage patients in an emergency room. We would expect to see less discrimination, as compared with that associated with the diagnoses made in the internist's normal practice. Conversely, a judge who is asked to be more discriminating, as might happen if a university called on faculty members to switch from letter grades to numerical grades, will show less consistency.

The two criteria are necessary to establish expertise. Both are empirical, so an index of expertise can be constructed purely from data. Using empirical criteria avoids the circularity inherent in approaches that rely on an expert's identification of the gold standards by which expertise is defined.

A study by Skånér, Strender, and Bring (1998) illustrates how expertise can be seen in a set of judgments. Twenty-seven Swedish general practitioners (GPs) judged the probability of heart failure for 45 cases based on real patients. Five of the cases were repeated, although the GPs were not informed of that. The case vignettes stated that each patient came to the clinic because of fatigue. Case-specific information was then provided for 10 cues: age, gender, history of myocardial infarction, dyspnea, edema, lung sounds, cardiac rhythm, heart rate at rest, heart X-ray, and lung X-ray.

For each vignette, the GPs were instructed to assess the probability that the patient suffered from any degree of heart failure (Skånér et al., 1998). The assessments were made on a graphic scale, with *totally unlikely* at one end and *certain* at the other; these were converted into 0- to 100-point values. The authors found wide, unexplained individual differences in the pattern of results. After inconclusive analyses of demographic variables, the authors concluded that the large variation among the GPs could not be readily explained.

Results for 4 of the GPs (identified by number) are shown in Figure 1. As can be seen, there was considerable variation between and within the 4 GPs. Still, each GP showed a distinctive pattern in terms of discrimination and reliability. Doctor #18 was highly discriminating (sizable differences among patients) and consistent (little difference between first and second presentations). Doctor #8 showed some discrimination but lacked consistency (especially for Patient B). Doctor #16 was consistent but treated all patients rather similarly – all were seen as having moderately high chances of heart failure. Doctor #23 showed no uniform pattern of discrimination or consistency.

Based on these data alone, one can gain considerable insight into the judgment strategies and abilities of the GPs. Doctors #18 and #16 were consistent, but one discriminated and the other did not. Doctors #8 and #23 were inconsistent and varied in their discriminations. We believe that most clients, without knowing anything further, would prefer someone like Doctor #18, who can make clear discriminations in a consistent way. In effect, our proposed measure (see next section) quantifies this intuition with a single index.

THE CWS INDEX

We propose the ratio of discrimination over inconsistency as an index of expertise (see Equation 1). *Discrimination* refers to the judge's differential evaluation of the various stimuli within a set. *Consistency* refers to the judge's evaluation of the same stimuli similarly over time; inconsistency is its complement. The ratio will be large when a judge discriminates effectively, and it will be reduced if the judge is inconsistent.

$$\text{CWS} = \frac{\text{Discrimination}}{\text{Inconsistency}}. \tag{1}$$

Our construction of the performance index parallels Cochran's (1943) suggestion that a ratio be used to assess the quality of a response instrument. Cochran argued that an effective dependent measure should allow the participant to express perceived differences among stimuli in a consistent way. We view expert judgment in the same way. We acknowledge our intellectual debt to Cochran by referring to our performance-based approach as the Cochran-Weiss-Shanteau (CWS) index.

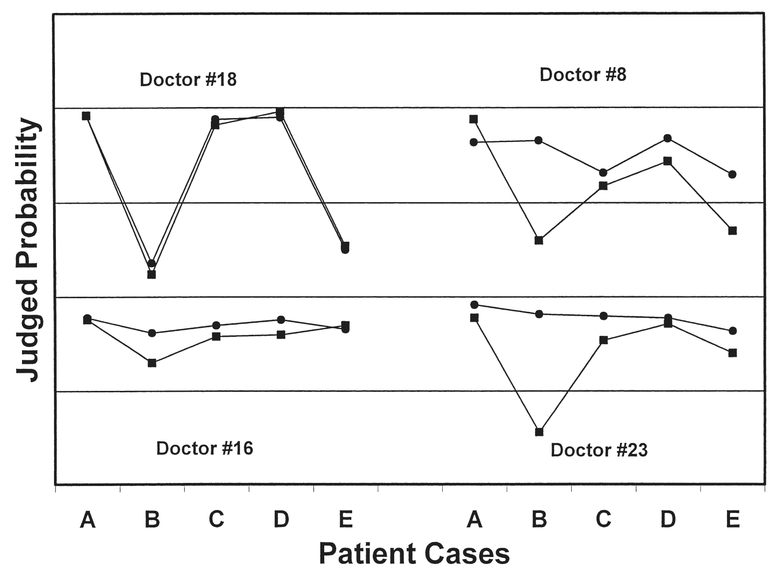

Figure 1. Judgments of the probability of heart failure for five patients made by four doctors, based on data from Skånér et al. (1998). The vertical axis for each pair of doctors (#18 and #8, #16 and #23) is the same. The five repeated cases are represented by the letters A-E on the horizontal axis. The circles are the judgments for the first presentation and the squares are the judgments for the second presentation. Thus the first judgment of Patient A by Doctor #18 is near 100; the second judgment is similar.

The intuition underlying the index is that a good measuring instrument necessarily has a high CWS ratio. A properly deployed instrument yields different measures for different objects and yields the same measure whenever it is applied to a given object. A ruler, for example, discriminates among objects of varying length and produces the same score for the same object. Accurate measurements necessarily yield high CWS.

Like a good measuring instrument, an expert judge must be both discriminating and consistent. It is easy to display either quality by using a simple response strategy, one that requires little knowledge of the stimulus objects. One can show discrimination simply by generating a wide variety of responses; one can exhibit consistency by making the same response for all cases. Adopting either of these strategies alone, however, guarantees that the other quality will be lost. The ability to incorporate both qualities simultaneously requires accurate and consistent assessment of stimuli, the essence of expert judgment. The CWS index tries to capture what physicists (Taylor, 1959) call the *resolving power* of the expert judge.

CALCULATING CWS

To implement the measure, we ask putative experts to evaluate a common set of stimuli. Some, if not all, of the stimuli must be evaluated more than once. We analyze each candidate's responses in two ways: The first is to estimate discrimination and the second is to estimate inconsistency. By forming the ratio of these estimates, we can determine whose judgmental performance is better for that set of stimuli.

We wish to impose three requirements on our index. First, it should have a "zero point." This fixed value represents the absence of expertise. Because we employ the ratio format, the starting point has the value of one. Although in principle values as low as zero may occur, a candidate who is completely insensitive to the stimuli (i.e., responds randomly) will have an expected CWS value = 1.

The second requirement is scale invariance across the constituent elements. If the responses are linearly transformed, the measures of discrimination and of inconsistency may change, but the CWS ratio should remain unchanged. We would like response instruments that might be expected to produce ratings approximately linearly related to one another, such as a category scale, a percentage scale, or a graphic rating, to yield comparable CWS scores.

For the sake of coherence, the third requirement is that both numerator and denominator be estimated using the same summary statistic. That is, the measure of discrimination should have the same basis as that of the measure of inconsistency.

Effective discrimination implies that as the stimulus changes, the evaluation changes accordingly. *High inconsistency* implies that repeated evaluations of the same stimulus differ considerably. Both of these constructs may be viewed as statistical dispersion, as it is the extent of differences that is crucial. Any summary measure of dispersion can yield CWS ratios that apply to all the candidate judges. Because dispersions are never negative, their ratio will always be nonnegative as well. Three dispersion measures – variance, standard deviation, and mean absolute deviation – might be used.

Because we see no clear theoretical advantage for any of the three dispersion measures, we do not wish to be dogmatic about the choice. (CWS ratios using any of the proposed dispersion measures satisfy our three requirements. It should be noted that scale invariance under linear transformation depends on the ratio formulation of CWS; it does not obtain for other ways of integrating discrimination and inconsistency, such as a difference measure. The dispersion measures are not linearly related to one another and do not, in general, produce comparable placements along a continuum of expertise. The two measures that square differences — variance and standard deviation — are monotonically related and so produce rankings that agree.) However, we have relied on variance measures (literally, mean squares) in our work so far and consider them the default option.

We have used the variance among mean responses to different stimuli (MS_{Stimuli}) as the estimate of discrimination and the variance among responses to the same stimulus ($MS_{\text{Replications}}$) as the measure of inconsistency. These quantities are easily obtained using standard statistics programs. Variances, with their heavy weighting of large discrepancies, have traditionally been used by statisticians to capture precision of measurement (Grubbs, 1973), with a ratio format the usual arrangement for comparison. Furthermore, variances afford a statistical advantage in that an estimate of their ratio is an asymptotically efficient estimator of the underlying ratio (I. R. Goodman, personal communication, February 1999). An additional consideration is that a procedure developed by Schumann and Bradley (1959), discussed later, can be used to determine whether or not two CWS ratios are significantly different.

We illustrate the computations for the 4 doctors selected from the Skånér et al. (1998) data in Table 1. As can be seen, Doctor #18, who shows high discrimination (3365.15) and low inconsistency (5.80), has a CWS value of 580.20. We cannot say whether this value, in isolation, is of a high or low magnitude. Therefore, the CWS values for the other doctors need to be considered. Doctor #8, with moderate discrimination and high inconsistency, has a CWS value of 1.21. Doctor #16, with low discrimination but low inconsistency, has a CWS value of 1.81. Doctor #23, with low discrimination and high inconsistency, has a CWS value of 0.76. Thus Doctor #18 stands apart from the other three doctors with a considerably higher CWS score.

CWS ratios computed using the alternative dispersion measures are also included in the table. As one would expect with the impact of squaring reduced, the range of obtained CWS ratios becomes smaller. Using mean absolute deviations, Doctor #16 moves slightly ahead of Doctor #8 in the rankings. No matter which measure is used, it is clear that Doctor #18 stands far apart from the others. However, the relative ranking of the other doctors depends to some extent on the dispersion measure selected.

SCHUMANN-BRADLEY PROCEDURE

When CWS estimates of discrimination and inconsistency are variances, a statistical comparison is available. Schumann and Bradley (1959) developed a procedure for determining whether one *F* ratio is significantly larger than another. The technical requirement is that the designs be identical in structure. This requirement will routinely be satisfied in a study comparing candidates judging the same stimuli and thereby legitimizes treating CWS ratios as if they were *F* ratios. (As noted, however, other measures of dispersion satisfy CWS but do not generate *F* ratios.)

The ratio of two *F* ratios constitutes a test statistic, w. This w is compared with w_0, a critical value found in the table presented by Schumann and Bradley (1959). The test can be employed either directionally or nondirectionally. The one-tailed test determines whether the candidate is significantly less capable than is a designated expert. The two-tailed test asks whether there is a significant difference between two judges. Each judge is considered as a separate "experiment." A computer program incorporating the Schumann and Bradley procedure and table of critical values has been published by Weiss (1985). Obtained ws allow comparison of the expertise exhibited by the various candidates as they judge a particular set of stimulus objects. Pairwise comparisons express how each candidate compares with the others. Alternatively, one may compare candidates with a reference expert.

TABLE 1: CWS for 4 Doctors

Dispersion Measure	CWS			
	Dr. #18	Dr. #8	Dr. # 16	Dr. #23
Variance	3365.15/5.80 = 580.20	490.75/404.60 = 1.21	65.40/36.10 = 1.81	330.40/434.00 = 0.76
SD	58.0/2.41 = 4.07	22.15/20.11 = 1.10	8.08/6.01 = 1.34	18.18/20.83 = 0.87
Mean absolute deviation	35.76/1.40 = 25.54	14.4/11.4 = 1.16	3.84/3.30 = 1.16	10.04/9.8 = 1.02

Two-tailed Schumann-Bradley significance tests were carried out on a pairwise basis using the variance ratios for the 4 doctors studied by Skånér et al. (1998), as presented in the first row of Table 1. The results, shown in Table 2, provide statistical confirmation of Doctor #18's expertise. As can be seen, Doctor #18 is significantly different from each of the other three doctors, with no differences among the remaining GPs.

When a judge's evaluations are expressed ordinally, as is typical in animal judging (Phelps & Shanteau, 1978), little efficacy is lost. As has been shown by Weiss (1986), sufficiently dense ordinal data may be subjected to analysis of variance with essentially no loss of power. Even yes/no responses cause no problem; Lunney (1970) demonstrated that carrying out analysis of variance on dichotomous responses yields results essentially equivalent to those obtained with continuous scales.

Policy recommendations are sometimes expressed qualitatively, with no hint of ordinal information. The air traffic controller selects one airplane over another and issues a control instruction. The bridge player chooses a bid. In each case the response is a label. For these nominal evaluations a CWS ratio can also be defined. The discrimination component is based on the proportion of nonmatching responses to different stimuli, whereas the inconsistency component is based on the proportion of nonmatching responses to the same stimulus. Our procedure for computing CWS for nominal data is presented in Appendix A.

STIMULUS OBJECTS

The basic task for our expert is to appraise each of a set of stimulus objects repeatedly. For ephemeral stimuli, such as an athletic performance, we could make a recording that preserves the information needed for expert judgment. This allows the same objects to be presented more than once to each individual. A factorial design is often convenient but is not required.

It is also necessary that stimuli vary in perceptible ways. This may not be trivial to achieve, given that determination of the extent of variation requires expertise and one does not wish a priori to presume it. These details of stimulus variation are crucial to one's ability to establish expertise. If the objects vary too little, then no one will be able to discriminate among them; if the objects vary too much, then all candidates will discriminate perfectly and no one will appear to be any better than anyone else. The variation issue is not a unique concern for CWS; the effect size a researcher obtains is always tied to the choice of stimuli (O'Grady, 1982). The best course of action may be to begin with a wide-ranging set of stimuli and plan to refine the selection subsequently.

An obtained CWS index depends on both the candidate's expertise and the particular set of stimuli presented. The more the stimuli differ from each other, the easier they are to discriminate. It is therefore not meaningful to compare CWS scores for candidates who have judged different stimulus sets, just as it is not meaningful to compare across different domains. An alternative perspective on this interaction is that when the same judge evaluates several stimulus sets, the index reflects task difficulty; higher CWS scores for a particular set imply that those stimuli were easier to distinguish (Thomas & Pounds, 2002).

THE VALIDITY CHALLENGE

We recognize that looking solely at internal properties of the data cannot yield ultimate satisfaction. When we conclude that Doctor #18

TABLE 2: Schumann-Bradley *w* Values for Each Pair of Doctors

	Dr. #18	Dr. #8	Dr. #16	Dr. #23
Dr. #18		479.50*	320.55*	763.42*
Dr. #8			1.50	1.59
Dr. #16				2.38

* = Significant at .05 level, two-tailed test.

is a more expert diagnostician, we are making an assumption that goes beyond the data – namely, that there is variation among the patients. The assumption of variation (in this case, of extent of illness) within the population is a customary one for social scientists. We acknowledge the logical possibility that Doctor #16 is in fact the most accurate, that all of the patients have equally severe conditions, and that Doctor #18 is seeing differences where none exists.

We would like to know what ultimately happened to the patients, but reality does not always provide such comfort. Even if follow-up were possible (in this real-life example, Dr. Skånér could track only some of the patients), it would take years before definitive results became available. Furthermore, the treatment the patients received depended on the diagnoses they received and thus would have differentially affected the outcomes.

A different approach to validity is to show that the index distinguishes between acknowledged experts and novices. We reanalyzed the data from Ettenson (1984; see also Ettenson, Shanteau, & Krogstad, 1987), who asked two groups of auditors to evaluate a set of financial cases. One group of 15 expert auditors was recruited from Big Six accounting firms. The group consisted of audit seniors and partners with 4 to 25 years of audit experience. For comparison, 15 novice accounting students were obtained from two large Midwestern universities.

Each financial case was described using 16 cues, and the value of each cue was selected to be high or low. For example, net income was set at either a high or low number. For every case, the participant judged the extent to which the firm was a going concern; this is a typical evaluation an auditor is asked to make.

Based on feedback from a senior faculty auditor, the cues were classified as either relevant (e.g., net income), partially relevant (e.g., aging of receivables), or irrelevant (e.g., prior audit results) for this task. Using a CWS analysis, we estimated discrimination variance from the mean square values for each cue. Inconsistency was estimated from the average of within-cell variances – high variance implies high inconsistency. The ratio of discrimination variance divided by inconsistency variance was computed to form separate CWS values for relevant, partially relevant, and irrelevant cues.

The results in Figure 2 show that CWS values for the expert group drop systematically as the relevance of the cues declines. For the novice group there is a similar but less pronounced decline. More important, there is a sizable difference between experts and novices for relevant cues. This difference is smaller for partially relevant cues and nonexistent for irrelevant cues.

We also reanalyzed data from Nagy (1981), who asked participants to evaluate summary descriptions of applicants for the position of computer programmer at a real (name deleted upon request) company in Seattle, Washington. The study employed both professional personnel selectors (experts) and management students (novices). Each applicant was described by legally relevant attributes (recommendations from prior employers and amount of job-relevant experience) and legally irrelevant attributes (age, physical attractiveness, and gender). Filler information from local phone books was used for background information, such as home address.

Four professionals in personnel selection and 20 business students made two evaluations of 32 applicants (generated from a 2 × 2 × 2 × 2 × 2 factorial design, in which there were two values for each of the five attributes listed previously). Before the evaluations participants

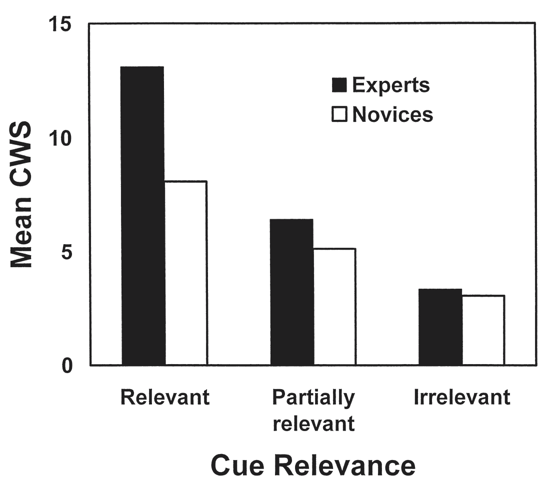

Figure 2. Mean CWS values for expert and novice auditors for three categories of cue relevance, based on data from Ettenson et al. (1987).

were reminded about the company's written policy for hiring (i.e., what information should and should not be used). Based on the 64 (32 × 2) evaluations, the importance of the five attributes was determined for each participant on a 0- to 100-point normalized scale; average CWS values were generated for each group.

For the two relevant attributes (recommendations and experience), the CWS values are nearly identical for the two groups (see Figure 3). This should not be surprising, given that participants were reminded immediately before the study about relevant hiring criteria. In contrast, CWS values for the three irrelevant attributes (age, attractiveness, and gender) reveal a very different pattern. For professionals, CWS values cluster around 1.0 (as they should, for it is inappropriate to discriminate on these aspects). In contrast, the values are considerably larger for students. Despite being told that age, attractiveness, and gender are not legally allowable, business students had sizable CWS values for these "irrelevant" attributes. Clearly, it is not easy to ignore something as obvious as age or gender, even if required by the guidelines. Professionals, however, followed strategies to do precisely that.

The results from these reanalyses show that CWS can distinguish between levels of expertise, especially when the focus is on the most relevant cues. We are aware of the circularity in this reasoning. How do we know the auditors were really "experts"? We have relied on the judgment of the researchers, trusting that at least roughly they were able to distinguish experts from nonexperts. We also used external information (the knowledge of cue relevancy) supplied by a subject matter "expert" to help validate the index.

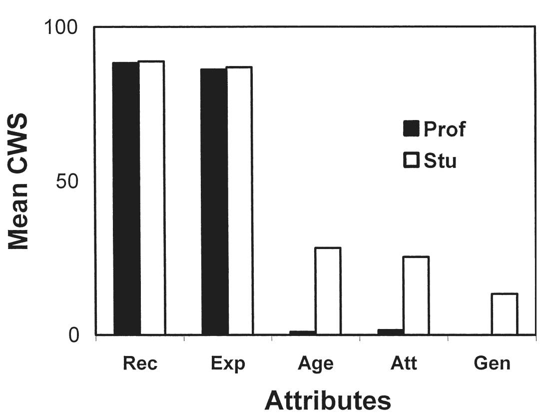

Figure 3. Mean CWS values for professional (PROF) and student (STU) personnel selectors, based on data from Shanteau and Nagy (1984). Two attributes – recommendations (REC) and experience (EXP) – were legally relevant, whereas the other three attributes – age, attractiveness (ATT), and gender (GEN) – were not.

LIMITATIONS

Because in general the index does not incorporate domain knowledge, the user can be misled if a candidate attends consistently to inappropriate stimulus features. Our criteria are necessary but not sufficient. That is, expert judgment yields high CWS, but high CWS does not guarantee expertise. A figure skating judge who evaluates the contenders primarily on the basis of, say, appearance (weighting costume and hair style heavily) would be deemed to show expertise according to our index if those attributes were used to discriminate consistently among the athletes. Clearly, this is not real expertise for the task of judging athletic performance. A general approach toward resolving the ambiguity is to ask for several kinds of judgments using the same stimuli, a strategy reminiscent of the classic multitrait-multimethod approach to construct validity (Campbell & Fiske, 1959).

A CWS index can be interpreted only relatively, not absolutely. That is, CWS is meaningful only in a comparative sense (i.e., it can be used to say which of two candidates is performing better with this particular set of stimuli). Any empirical index will have this character unless stimulus-specific performance norms are available. If true expertise is rare for the requested judgments, there may be no experts included in the study. Hence the identified "experts" may not really be very expert.

FUTURE DIRECTIONS

CWS may be valuable in longitudinal studies on the development of expertise. One can separate the components to see if the bulk of the improvement comes from improved consistency, as has been suggested by Ashby and Maddox

(1992). If it turns out that those with higher CWS at the beginning of training maintain their superiority, then a selection tool is available. We have already seen CWS scores improving with practice (Friel, Thomas, Shanteau, & Raacke, 2001).

The overarching challenge is to apply our index to the other categories of expertise. Does the categorization presented in the Categories of Expertise section have more than heuristic value? Although the utility of the CWS index does not depend on the proposed hierarchy, understanding of the nature of expertise would be enhanced by cross-task, within-subject comparisons. A practical implication of the structure is that training in judgment might contribute more to expertise in instruction than would a similar investment in training of specialized performance skills.

We have focused on tasks for which there is no clear-cut objective performance index. When a valid, unconfounded outcome measure is available, how will expertise as assessed by CWS compare with expertise measured by deviations from correct answers? If the assessments were in accord, perhaps skeptics would be willing to trust CWS. We recognize that lack of sufficiency inhibits acceptance.

CONCLUSION

CWS has been useful for evaluating judgmental expertise in medical diagnosis, auditing decisions, and personnel selection. We have also had some success in applying CWS to performance expertise in air traffic control (Thomas, Willems, Shanteau, Raacke, & Friel, 2001). The latter deserves some elaboration.

The air traffic system depends on the skills of controllers to manage the order and spacing of flights across the country and around the world. Their expertise is an important practical concern. The gold standard of safe arrival provides an insensitive criterion for discriminating among controllers. Although we are grateful that so few accidents or incursions occur, the paucity of data need not imply that all controllers are equally proficient.

CWS is an approach that can be used to select, train, evaluate, and enhance performance. When the controller's performance is dependent on equipment, CWS can be considered to be evaluating the combination of human and machine. Therefore, studying the expertise displayed by a given controller while using different supporting equipment can allow comparison of the relative contribution of the tools. The issue addressed by CWS is of importance not only to researchers but also to a public that relies heavily on skilled professionals working with complex apparatus.

As members of groups, we are all called on to make important decisions. We may or may not be qualified to make those decisions. Democratic tendencies prevail when differential expertise has not been established or is philosophically objectionable. The democratic approach considers the judgments of everyone, with equal merit given to all. For example, academic departments often employ democratic procedures, with equal votes for everyone, in personnel selection and retention decisions. Is it similarly appropriate to employ democracy to decide on the department's next computer system?

Reliance on a select group of knowledgeable individuals is the antithesis of democracy. When people agree to surrender their decision-making power to the elite, they are entitled to assurance that those designated are truly capable. We believe that empirical assessment of expertise is a cornerstone of that assurance.

APPENDIX A

We propose the following construction for the index with nominal data. The numerator should capture the discrimination in the judge's assignments. Because responses are nominal, one can measure only whether the responses agree; the amount of discrepancy is not defined. The fewer matches among the responses to different stimuli, the better the discrimination. We compare the number of obtained nonmatches among all pairs of responses to different stimuli with the possible number of matches to get the numerator for the CWS index.

For the denominator, we look for inconsistency among the responses to the same stimuli. Scoring over replications, the more matches among responses to the same stimulus, the more consistent the judgments. We compare the number of nonmatches among all pairs of responses to the same stimulus with the number of possible matches. We illustrate the computations for some artificial data in Table A1.

For both numerator and denominator, we utilize the proportion of obtained pairwise nonmatches to possible matches. In measuring discrimination, a match is evidence of failure to discriminate, so the greater the proportion of observed nonmatches, the greater the discrimination. In measuring inconsistency, a *match* means the response was consistent, so the greater the proportion of observed nonmatches, the greater the inconsistency. Expert performance is marked by few matches across columns (stimuli) and many matches within columns (replications). If there are no matches within columns – no consistency at all – the CWS ratio is undefined, but that outcome unambiguously connotes a lack of expertise.

CWS numerator (discrimination) =

$$\sum_{\text{Matrix}} \frac{\text{Nonmatches across columns}}{\text{Possible matches across columns}}.$$

The number of possible matches across columns is most easily calculated by subtracting the number of possible within-column matches from the total number of possible matches. Each response may be matched to any other, so the total number of possible matches is ${}_{15}C_2$ (= 105). There are ${}_3C_2$ (= 3) possible matches within each column, so the number of possible within-column matches is $5 \times {}_3C_2$ (= 15). Therefore, there are 90 possible matches across columns and 15 possible matches within columns.

In the example given in Table A1, there were 7 pairs of “A” responses in different columns. “B” was matched 18 times, “C” once, and “D” was not matched at all.

TABLE A1: Illustration of CWS Index for Nominal Data (4 Response Alternatives)

	Stimulus				
	1	2	3	4	5
Replicate 1	A	D	B	C	C
Replicate 2	A	B	B	B	B
Replicate 3	A	B	A	B	A
Matches	3	1	1	1	0

$$\text{Numerator} = \frac{90 - 26}{90} = 0.711.$$

CWS denominator (inconsistency) =

$$\sum_{\text{Columns}} \frac{\text{Nonmatches within columns}}{\text{Possible matches within columns}}.$$

$$\text{Denominator} = \frac{15 - 6}{15} = 0.60.$$

$$\text{CWS Index} = \frac{\text{CWS numerator}}{\text{CWS denominator}} = \frac{0.l711}{0.60} = 1.185.$$

ACKNOWLEDGMENTS

Preparation of this manuscript was supported by Grant 98-G-026 from the Federal Aviation Administration in the Department of Transportation. We thank Julia Pounds and Rickey Thomas for valuable discussions regarding the substantive issues involved in the implementation of the CWS index. We are especially grateful to Ward Edwards for the insights he provided after critically reviewing the manuscript. We received very helpful suggestions from the journal’s anonymous reviewers. Further information about the index and applications can be found at the CWS Web site: http://www.ksu.edu/psych/cws.

REFERENCES

Ashby, F. G., & Maddox, W. T. (1992). Complex decision rules in categorization: Contrasting novice and experienced performance. *Journal of Experimental Psychology: Human Perception & Performance, 18,* 50–71.

Ashton, A. H. (1985). Does consensus imply accuracy in accounting studies of decision making? *Accounting Review, 60,* 173–185.

Ashton, R. H. (2000). A review and analysis of research on the test-retest reliability of professional judgment. *Journal of Behavioral Decision Making, 13,* 277–294.

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.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78,* 1–3.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

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

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.

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology, 47,* 273–305.

Ericsson, K. A., & Smith, J. (1991). Prospects and limits in the empirical study of expertise: An introduction. In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 1–38). Cambridge, UK: Cambridge University Press.

Ettenson, R. (1984). *A schematic approach to the examination of the search for and use of information in expert decision making.* Unpublished doctoral dissertation, Kansas State University, Manhattan, KS.

Ettenson, R., Shanteau, J., & Krogstad, J. (1987). Expert judgment: Is more information better? *Psychological Reports, 60,* 227–238.

Friel, B. M., Thomas, R. P., Shanteau, J., & Raacke, J. (2001). CWS applied to an air traffic control simulation task (CTEAM). In *Proceedings of the 2001 International Symposium on Aviation Psychology* [CD-ROM]. Columbus: Ohio State University.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103,* 650–669.

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.

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine, 73,* 538–540.

Lunney, G. H. (1970). Using analysis of variance with a dichotomous dependent variable: An empirical study. *Journal of Educational Measurement, 7,* 263–269.

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, KS.

O'Grady, K. E. (1982). Measures of explained variance: Cautions and limitations. *Psychological Bulletin, 92,* 766–777.

Phelps, R. H., & Shanteau, J. (1978). Livestock judges: How much information can an expert use? *Organizational Behavior and Human Performance, 21,* 209–219.

Russell, B. (1993). *Sceptical essays.* London: Routledge.

Schumann, D. E. W., & Bradley, R. A. (1959). The comparison of the sensitivities of similar experiments: Model II of the analysis of variance. *Biometrics, 15,* 405–416.

Shanteau, J. (1995). Expert judgment and financial decision making. In B. Green (Ed.), *Risky business* (pp. 16–32). Stockholm: University of Stockholm School of Business.

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.

Swets, J. A. (1986). Indices of discrimination or diagnostic accuracy: Their ROC's and implied models. *Psychological Bulletin, 99,* 110–117.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1,* 1–26.

Taylor, L. W. (1959). *Physics: The pioneer science* (Vol. 2). New York: Dover.

Thomas, R. P., & Pounds, J. (2002). *Identifying performance in complex dynamic environments.* Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Thomas, R. P., Willems, B., Shanteau, J., Raacke, J., & Friel, B. (2001). CWS applied to controllers in a high-fidelity simulation of air traffic control. In *Proceedings of the 2001 International Symposium on Aviation Psychology* [CD-ROM]. Columbus: Ohio State University.

Uebersax, J. S. (1993). Statistical modeling of expert ratings on medical treatment appropriateness. *Journal of the American Statistical Association, 88,* 421–427.

Weiss, D. J. (1985). SCHUBRAD: The comparison of the sensitivities of similar experiments. *Behavior Research Methods, Instrumentation, and Computers, 17,* 572.

Weiss, D. J. (1986). The discriminating power of ordinal data. *Journal of Social Behavior and Personality, 1,* 381–389.

Weiss, D. J., & Shanteau, J. (in press). The vice of consensus and the virtue of consistency. In K. Smith, J. Shanteau, & P. Johnson (Eds.), *Psychological explorations of competent decision making.* New York: Cambridge University Press.

David J. Weiss received his Ph.D. 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 Ph.D. in experimental psychology in 1970 from the University of California, San Diego. He is a professor of psychology at Kansas State University.

Date received: October 2, 2001
Date accepted: November 11, 2002