

**How Can You Tell if Someone is an Expert?**

**Empirical Assessment of Expertise**

James Shanteau  
Kansas State University

David J. Weiss  
California State University, Los Angeles

Rickey P. Thomas  
Kansas State University

Julia Pounds  
Federal Aviation Administration, Oklahoma City

Please mail all correspondence to:

James Shanteau  
Department of Psychology  
Bluemont Hall 492  
1100 Mid-Campus Drive  
Kansas State University  
Manhattan, KS 66506-5302 USA  
(Phone: 785/532-0618)  
(Fax: 785/532-5401)  
(e-mail: <[shanteau@ksu.edu](mailto:shanteau@ksu.edu)>)

To appear in: Sandra L. Schneider, S. L., & Shanteau, J. (In Press). *Emerging perspectives on decision research*. Cambridge, U.K.: Cambridge University Press.

## Abstract

The definition of “expert performance” is obviously vital to any analysis of expertise. If external standards exist, then the definition is straightforward: the correct answers determine expert performance. Unfortunately, external standards seldom exist in domains requiring expertise; that is why experts are needed in the first place. The purposes of this paper are (1) to review traditional methods for defining expert performance in the absence of external standards and (2) to present a promising new approach for assessing expert performance (labeled CWS for *Cochran-Weiss-Shanteau*). Eight traditional procedures are reviewed, along with discussions of their advantages and disadvantages. Then CWS is presented along with a technical elaboration. The CWS approach is illustrated through reanalyses of three previous studies of experts. A study of physicians revealed that expert diagnostic skill was related to both discrimination and consistency – the two components of CWS. In a study of agricultural judges, the sensitivity of CWS was demonstrated by its ability to distinguish between livestock judges with different subspecialties. In a study of auditors, CWS correctly linked group differences in expertise (experts vs. novices) to information relevance. These reanalyses demonstrate that CWS offers new insights into the nature of expertise.

## Background

Although behavioral studies of expertise have been conducted for over a century (Shanteau, 1999), there remains an unanswered question – what is the definition of “expert performance?” If there are correct answers (a “gold standard”), the solution is straightforward – the correct answers define expert performance.

This validity-based approach is compelling in its simplicity. Unfortunately, it is problematic in its application. The problem is that experts are needed precisely in those domains where correct answers are least likely to exist (Gigerenzer et al., 1999; Shanteau, 1995). Indeed, if we have the correct answers, why do we need an expert?

The purpose of this paper is to explore the application of a novel approach to assessing expert performance labeled CWS for *Cochran-Weiss-Shanteau*. The measure is based on the behavior of would-be experts, i.e., their performance in the domain. In effect, the judgments of an individual are used to evaluate their expertise.

The chapter is organized into five sections: First, we review approaches used in previous research to define expert performance. Second, we develop our proposed approach to evaluation of expert performance. Third, we apply this approach to several previous studies involving experts. Fourth, we will describe some conditions that should be considered when applying CWS. Finally, we offer conclusions and final comments.

## Previous Approaches

Investigators have used many approaches to define expert performance in previous studies. Eight of these traditional approaches are described here. We also consider the advantages and, more important, the disadvantages of each approach.

### **Experience.**

Many prior studies used number of years of job-relevant experience as a surrogate for expertise. That is, participants with more years of experience are classified as “experts,” while others with less experience were labeled as “novices.” In general, this approach seems reasonable. Presumably, no one can work as a professional for any length of time if they are incompetent.

Unfortunately, while experts normally have considerable experience, the converse may not be true. Many professionals with years of experience never become experts. Such individuals may work along side top experts, but they never meet the standards required of true expertise.

In a study of grain judges, for instance, Trumbo, Adams, Milner, and Schipper (1962) reported that experience was uncorrelated with accuracy of wheat grading. Instead, they found a different trend: judges with more experience overrated the quality of the grains – an early and interesting form of “grade inflation.” Similarly, Goldberg (1968) asked clinical psychologists with varying amounts of experience to diagnose psychiatric patients. He found no relation between years of experience and accuracy; however, the confidence of clinicians in their diagnoses did increase with experience.

Although there are certainly instances of positive correlations between experience and expertise, there is little reason to expect this relation to apply universally. At best, experience is an uncertain predictor of degree of expertise. At worst, experience may reflect years on the job – and little more.

### **Accreditation.**

In many professions, individuals receive some form of accreditation or title as a certification of their skill. For instance, doctors may be “board-certified specialists” and university faculty may become “full professors.” It is safe to say that a certified specialist is more likely perform as an expert than someone who is not certified.

The problem with accreditation is that it is often tied more to time on the job than it is to professional performance. This can be particularly true in structured bureaucracies, such as the

military. The rank of photo interpreters (PI), for instance, may range from Sergeant to Major. Yet, the performance of PI's is often unrelated to their rank (T. Levitt, personal communication, 1992).

Another example comes from the Israeli Air Force, where lead pilots in combat are identified by skill not rank, i.e., a General may be following a Captain. This has been cited as one reason for the superiority of the Israelis against Arab Air Forces (where lead pilots are usually determined by rank, as it is in most countries). The Israelis recognized that skill is not always reflected by formal certification (R. Lipshitz, personal communication, 1995).

Another problem with accreditation is a ‘ratchet -up effect’ – people often move up the certification ladder, but seldom down. That is, once certified, the person is accredited for life. Even if an individual’s skill level suffers a serious decline, the title or rank remains. (Just ask students about the teaching ability of some full professors.)

### **Peer Identification.**

A method commonly used by many researchers (including the present authors) is to rely on identification of expert performers by those working in the field. That is, professionals are asked whom they would consider to be an ‘expert.’ When there is some agreement on the identity of such individuals, then they are labeled to have expertise.

In a study of livestock judges, for example, Phelps (1977) asked professional animal scientists whom they considered to be the best. From their responses, she identified four livestock judges to be the ‘experts’ in her investigation (this study is described further below). Similarly, Nagy (1981) relied on peer identification to identify experts in her study of personnel selectors.

Absent other means of assessing expertise, peer identification is a good strategy to follow. It is unlikely that others working in a field will all identify the same ‘wrong person’ as an expert. If they agree, it seems safe to assume that the agreed-upon person probably is an expert.

The problem with this approach is a “popularity effect” – someone who is better known or more popular with their peers is 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 on the cutting edge of new insights. Indeed, those who make new discoveries in a domain are frequently out-of-step with their peers at the time of their breakthroughs (Gardner, 1957). Thus, peer identification is more likely to identify “yesterday’s expertise” than “tomorrow’s expertise.”

### **Between-Expert Reliability.**

In seminal research on expertise, Einhorn (1972, 1974) argued that agreement between experts is a necessary condition. That is, experts should agree with each other (also see Ashton, 1985). If they do not, then it suggests that some of the would-be experts are not what they claim to be.

To examine this argument, Table 1 lists between-expert correlations from eight published studies. The four categories correspond to a classification of task difficulty proposed by Shanteau (1999). There are two domains listed for each category, with the reliability values given as correlations. For example, the average correlation for weather forecasters (a decision aided task) is quite high at .95, whereas the average  $r$  for stockbrokers (an unaided task) is low at .32. As predicted from Shanteau’s classification, the correlations increase with more structured tasks.

To be sure, between-person reliability appears to be a compelling property for expertise. After all, we feel confused when two (or more) medical doctors disagree about a medical diagnosis. When doctors agree, on the other hand, we feel more much comfortable with the agreed-upon diagnosis.

The problem with between-expert reliability is that agreement can result from artificial consensus, e.g., groupthink (Janis, 1972). There are many historic cases where the best course of action was not identified by a group of experts because they focused initially on an inferior alternative, e.g., Bay of Pigs. Thus, a group of experts may agree – and they all may be wrong.

### **Within-Expert Reliability.**

Einhorn (1972, 1974) also argued that intra-person reliability is necessary for expertise. That is, an expert's judgments should be consistent from time to time. In contrast, inconsistency would be *prima facie* evidence that the person is not behaving as an expert.

To examine consistency, Table 2 lists within-person consistency values for the eight domains listed in Table 1. The average consistency  $r$  for weather forecasters is quite high at .98, whereas the average consistency for stockbrokers is less than .40.

As expected from the Shanteau (1988) classification, more structured tasks produce higher consistency values than unstructured tasks. To a first approximation, therefore, it appears that within-expert reliability provides a good correspondence to the performance levels of experts.

The shortcoming in this approach is that high consistency can be obtained by following a simple, but incorrect rule. As long as the rule is followed precisely, a person's decisions will exhibit high consistency. By answering "yes" and "no" to alternating questions, for instance, one can be perfectly reliable. But such answers would generally be inappropriate. Thus, within-expert reliability is necessary – an expert could hardly behave randomly – but it is not sufficient.

### **Subject Matter Experts (SMEs).**

In many fields, the decisions of one or more "super experts" become(s) recognized as the equivalent of a gold standard. That is, the answers of these pre-identified top experts become the *de facto* standard for evaluating the performance of others. In effect, the decisions of subject matter experts (SMEs) become the correct answer.

Not surprisingly, this approach is commonly used when no certifiable correct answers exist. For example, performance of air traffic controllers (ATC) is evaluated by SMEs, i.e., an ATC operator is certified based on how close he/she matches the judgments of an SME. Similarly in accounting, "standard practice" is established by a committee of SMEs.

Clearly, SMEs are often needed. For example, our empirical studies of experts have depended on input from SMEs in the planning and interpretation stages of research. Similarly, SMEs are essential for feedback on new procedures, equipment, etc. There is an obvious circularity, however, in having one expert certify the performance of another expert. Among other problems, this can lead to gatekeeping whereby senior SMEs keep out (demote) ‘young turks’ who may have new ideas. Thus, SMEs can retard progress in a field, e.g., by slowing adoption of new ideas.

Another difficulty is that reliance on SMEs confuses subjective and objective standards. One example occurs in food tasting panels where trained tasters are often viewed as being ‘just as good as a machine.’ When the senior author questioned such claims, he was told that unreliability and lack of consensus are not an issue with trained panelists. In effect, defining SMEs’ judgments as the gold standard solved the problem of evaluating expertise. Clearly, such claims need to be tested.

### **Factual Knowledge.**

In many studies of skilled problem solving or game playing, expert performance has been identified using tests of factual knowledge. For example, Chi (1978) used knowledge about dinosaurs to separate children into ‘experts’ and ‘novices.’ Similarly, baseball ‘experts’ have been identified using a test of knowledge about the rules of the game.

Knowledge of relevant facts is clearly necessary for expertise. Someone who does not know the facts about a domain will be unable to make competent decisions. Yet knowledge alone is seldom sufficient to establish that someone has expertise. For example, knowledge of the rules of baseball is not enough to know how to play a particular position, or to pitch a curve ball, or to manage a team in the bottom of the ninth inning. Each of these skills requires more than knowing what the facts are. They also require an understanding of what to do with the facts.

Another problem is knowing which facts to apply in a given situation. As one expert livestock judge put it, “expertise lies in knowing when to follow the rules – and when not to.” In most domains, that is the hard part for experts.

### **Creation of Experts.**

In certain special contexts, it is possible to create experts through extended training and practice. This approach has significant advantages, including the ability to study the development of expertise longitudinally. Moreover, the skills learned are under direct control of the researchers.

Chase used this approach to create a short-term memory “digit -span expert” (Chase & Ericsson, 1981). A student, who was a track athlete, learned to translate groups of digits into running times for various distances. When asked to retrieve the digits, the student recalled the digits in clusters tied to running. Based on such strategies, the student was able to break the old record for digit span of 18. His new world record – over 80! (Other students have since pushed the record beyond 100.)

Expert performance can be created in this way for specific tasks, e.g., playing computer games or running a simulated microworld. Most realms of expertise, however, require a broad range of skills based on years of training and experience. For instance, it takes many 15 to 20 years of practice to become a livestock judge. Obviously, this level of expertise cannot be simulated by working with students for a few months. Still, the creation of experts offers a promising approach that should be explored further.

### **The CWS Approach**

As the preceding survey shows, many approaches have been used previously to define expertise. However, each of these approaches has one (or more) serious flaws; no generally acceptable technique exists at this time. To address this need, the two senior authors developed a novel approach for assessing expert performance. They combined two measures, each of which is a necessary but not sufficient condition for expertise, into a single index.

First, Weiss and Shanteau argued that *discrimination* is necessary for expertise (also see Hammond, 1996). The ability to differentiate between similar, but not identical, cases is a hallmark of expert performance; top experts make distinctions that novices miss. Second, they followed Einhorn' s (1974) suggestion that within-person *consistency* is necessary for expertise; if someone cannot repeat his/her judgment of a case, then he/she is not behaving as an expert.

Following a formulation used by Cochran (1943) in a different context, these two concepts are combined into an index using a ratio. The measure for *discrimination* (more is better) is divided by a measure for *inconsistency* (less is better). This ratio thus provides a descriptive index of degree of expertise, where bigger is better. The ratio has been labeled CWS (for *Cochran-Weiss-Shanteau*).

The CWS approach can be applied by asking would-be experts to judge a series of stimulus cases – this allows for assessment of their discrimination ability. In addition, at least some cases are repeated – this allows for assessment of their consistency. These two measures are combined by a ratio to yield the CWS value. The ratio can then be used to assess whether someone is behaving more (high value) or less (low value) like an expert.

### **Technical Issues**

There are five technical issues involving CWS that deserve elaboration. For a more detailed discussion of these issues, see Weiss and Shanteau (submitted).

First, the CWS index could be computed in a variety of ways in addition to a ratio, e.g., as a difference score. Alternatively, discrimination and consistency could be reported separately, without combining them into an index. The goal behind our approach was to summarize the tradeoff between two inconsistent entities, discrimination and consistency, in a rigorous, quantitative fashion. Ratios are commonly used in such situations, e.g., price/quality ratio, speed/accuracy tradeoff, and cost/benefit ratio. We view CWS in a similar light. However, we also acknowledge that other analytic rules may have value in some settings.

Second, any measure of dispersion can be used to measure discrimination and consistency. There are three common measures of dispersion: variance (mean squares), standard deviation (square root of variance), and mean absolute deviation (MAD). While any of these might work, we use variances as our default option. That is, we estimate discrimination as the variance between responses to different stimuli – larger variances imply greater discrimination. Similarly, we estimate consistency as the variance between responses to repeated stimuli – smaller variances imply greater consistency. Statisticians have traditionally used variances to evaluate the precision of measures (Grubbs, 1973). Further, a ratio of variances is an asymptotically efficient estimator (Goodman, personal communication, 1999). Another consideration is that the Schumann-Bradley (1959) procedure can be used to determine whether two CWS ratios are significantly different (see the example below).

Third, computing CWS as a ratio of variances, of course, is parallel to the F-ratio commonly used in Analysis of Variance (Anova) for statistical tests. However, our use of a variance ratio is different in at least four respects: (1) CWS is a measurement instrument to evaluate individuals or conditions. We have no interest in using the variance ratio to test significance, e.g., between treatments in an experiment. (2) The application of CWS is restricted to a particular set of stimuli in a specific domain. Thus, CWS cannot be used to compare responses of different stimuli sets or across different domains. In contrast, the F-test is designed to do both. (3) CWS makes no assumptions about the underlying distributions of responses or their properties. Anova, in contrast, rests on assumptions such as normality and independence of errors. (4) The concept of a CWS ratio (Discrimination/Consistency) can be extended to any level of response measure. For instance, CWS can be computed for categorical, classification, or nominal cases (Weiss & Shanteau, submitted).

Fourth, CWS is defined as a ratio with no other scaling required. There are conditions, however, where a transformation of the ratio might be appropriate. When we applied CWS to response times

for air traffic controllers (Friel, Thomas, Raacke, & Shanteau, 2001), there was a strong correlation between means and variances. In such situations, the solution to the problem of homoscedasticity is to apply a log transformation (Winer, 1971). This of course leaves the rank orders of the CWS scores unchanged. But it improves the distributional properties of the ratio.

Finally, although the correct ‘gold standard’ answers are traditionally held up as the ultimate criteria, there may be circumstances where CWS can actually outperform a gold standard. One example is air traffic control where preventing airplanes from getting too close together (‘separation errors’) is recognized as a gold standard. Although professional controllers rarely commit such errors, there are still differences in performance in terms of efficiency, time, etc. CWS has proved to be sensitive to these differences in performance, even when there were no errors. In other words, CWS is capable of assessing levels of expert performance beyond the point at which (because of ceiling effects) gold standards are no longer useful (Friel et al., 2001).

### **Reanalyses of Previous Research**

In this section, we illustrate CWS by applying the approach to three previous studies. The research in each case was designed to study a substantive problem involving experts. The use of CWS provides important supplementary information that advances the goals of each project.

#### **Medical Diagnosis.**

Data from a study of medical diagnosis by Skånér, Strender, and Bring (1998) illustrates how CWS can be applied to evaluate expert performance. Twenty-seven Swedish General Practitioners (GPs) judged the probability of heart failure in 45 case vignettes. Unknown to the GPs, five of the cases were presented twice. Based on real cases of chronic fatigue, the provided patient-specific information for age, gender, lung sounds, cardiac rhythm, heart rate, and heart/lung X-rays, etc.

For each vignette, GPs rated the probability of heart failure using a graphic rating scale ranging from ‘totally unlikely’ to ‘certain;’ the responses were later converted to 0 -100 values. The results

of the original study revealed wide, individual differences across GPs. Despite extensive analyses, the authors were unable to explain the large variation between GPs.

**Graphical Analysis.** The results for three GPs (identified by number) are graphed in Figure 1. The letters along the horizontal axis represent the five repeated cases. The open and filled points are the judgments for the two presentations. For instance, the first judgment of Case A by Doctor #12 is near 100; the second judgment is similar.

As can be seen, there is considerable variation between the three GPs. Still, each GP shows a distinctive pattern in terms of discrimination and reliability. Doctor #12 is highly discriminating (there are sizable differences between patients) and consistent (there is little difference between first and second presentations). Doctor #6 shows some discrimination, but lacks consistency (especially for patient E). Doctor #26 is more consistent, but treats patients rather similarly – they are all viewed as having moderately high chances of heart failure.

Based on these plots alone, we can gain considerable insight into the judgment abilities of the GPs. Doctors #12 and #26 are consistent, but one discriminates and the other does not. Doctors #12 and #6 show discrimination, but one is consistent and the other is not. We believe that without knowing anything further, most clients would prefer someone like Doctor #12, who makes clear discriminations in a consistent way. In effect, our proposed CWS measure quantifies this intuition.

**CWS Analysis.** The CWS ratios for the three doctors are given in Figure 1. To estimate discrimination, the variance between the means for the five patients was calculated; discrimination variances were 3377.60, 914.40, and 65.35 for Doctors #12, #6, and #26, respectively. To estimate consistency, the variance between the two responses for each case was calculated; consistency variances were 86.50, 218.80, and 68.30, respectively. When combined into a ratio, this led to CWS values of 39.05, 4.18, and .96, for the three GPs.

To compare two (or more) experts, CWS measures can be compared using a procedure developed by Schumann and Bradley (1959). This allows the researcher to determine whether one variance ratio is different from another (Weiss, 1985). This may be useful, for example, when CWS values vary widely and the goal is to determine if they are significantly different. For the three doctors in Figure 1, the Schumann-Bradley test revealed the three GPs are significantly different from one another in 2-tailed tests at the .05 level.

From these analyses, we can see that CWS analyses confirm the graphical analyses. That is, Doctor #12 appears to be a better diagnostician than Doctor #6, who in turn is better than Doctor #26. Moreover, we make these statements without any knowledge of the correct answers; in medicine, it is rare that gold standards exist at the time of initial diagnoses. That is why CWS is valuable: It provides a means of quantifying the observed patterns of behavior.

The results for the other 24 GPs can be categorized into three groups. Nine of the GPs had high discrimination and high consistency – as illustrated by Doctor #12. Thirteen of the GPs were discriminating but inconsistent – as illustrated by Doctor #6. The other five GPs revealed little discrimination, although they were fairly consistent – as illustrated by Doctor #26.

The latter pattern is particularly interesting as it may illustrate a strategy of “defensive medicine.” By classifying all patients as needing further attention, GPs such as #26 may be trying to reduce false negative diagnoses. The cost, of course, is an increase in the rate of false positives. The implications of this finding are presently being explored in collaboration with the original authors.

**Caveats.** Two caveats are necessary. First, these results are only meaningful if the five cases selected by Skånér et al. (1998) are representative real patients. We have no way of assessing this, although we have been assured that the cases are not atypical (Skånér, personal communication, 2001). Thus, the usefulness of CWS depends on the selection of stimuli. Of course, the same applies to any other analytic procedure that might be used to examine expert performance.

Second, the CWS ratios are informative about rankings, but should not be interpreted further. Specifically, we cannot say that Doctor #6 has four times the expertise as Doctor #26, or that the difference between Doctors #12 and #6 is greater than the difference between Doctors #6 and #26. What we can say is that the ordering of diagnostic ability for the three doctors is #12, #6, and #26.

### **Livestock Judging.**

The sensitivity of CWS was demonstrated in a reanalysis of a study of livestock judging. Phelps (1977) had four professional livestock judges evaluate 27 drawings of gilts (female pigs). An artist created the drawings to reflect a 3-x-3-x-3, size x breeding potential x meat quality, factorial design. The judges independently made slaughter judgments (how good is the meat from the animal) and breeding judgments (how good is the animal for breeding) for each gilt. All stimuli were presented three times, although judges were not told about the repetition.

Two of the judges were nationally recognized swine experts and were quite familiar with gilts of the type shown in the drawings. The other two judges were nationally recognized cattle experts; they knew about swine judging, but lacked day-to-day experience with gilts. The CWS ratio for each judge was computed separately, and then averaged with the other similar judge. Based on the factorial structure of the stimuli, it was possible to compute CWS ratios for each of the three dimensions. The CWS values were computed separately for slaughter and breeding judgments.

The results appear in Figure 2. As can be seen, there is a substantial difference between the two sets of judges. For cattle judges, there is little difference in slaughter and breeding judgments – for both cases, meat quality dominated. For swine judges, meat quality also dominated for slaughter, but CWS values for breeding potential and meat quality are sizable for breeding judgments. This may reflect the unfamiliarity of breeding in swine by cattle judges. Swine judges, in contrast, focused on meat quality for slaughter judgments and breeding potential for breeding judgments.

It is worth emphasizing that the four judges were all highly skilled professionals. Nonetheless, the CWS approach proved sensitive to the differences between these two types of experts. This study also highlights the importance of examining task differences when evaluating expertise. In too many studies, the term “expert” is used generically without reference either to the domain or the task within the domain. As illustrated in this study, even highly qualified (and highly paid) professionals have skills that vary from task to task. Thus, we should not expect an expert on one task in one domain necessarily to behave expertly on another task.

### **Auditing Judgment.**

CWS has also proved sensitive to information relevance in distinguishing aggregate difference in expertise. Ettenson (1984; also see Ettenson et al., 1987) had two groups of auditors evaluate a series of accounting cases described by a set of financial cues. One group of 15 “expert auditors” was recruited from Big Six accounting firms; they included audit seniors and partners, with 4 to 25 years of experience. The other 15 “novice auditors” were advanced students in accounting.

For each case, participants were asked to make a *Going Concern* judgment. Based on feedback from a senior auditing associate, the financial cues were classified as either *relevant* (e.g., Net Income), *partially relevant* (Aging of Receivables), or *irrelevant* (Prior Audit Results) for the *Going Concern* judgment. Discrimination was estimated from the mean-square variance for each financial cue. Consistency was estimated from the average within-cell variance for each participant. In this way, separate CWS values could be estimated for the relevant, partially relevant, and irrelevant cues. The CWS values were then averaged for the two groups.

The results in Figure 3 show that CWS values for experts drop systematically as cue relevance declines. For novices, there is a similar, but smaller decline. More important, the difference between experts and novices decreases as relevance drops. That is, the size of the difference between groups

is smaller for less relevant cues. These results show that CWS can distinguish between levels of expertise for different groups – especially when cue information is highly relevant.

### **Comments**

Six comments are appropriate on our use of CWS in these reanalyses. First, the stimuli in these studies were abstractions of real-world problems. In particular, the stimulus cues in each case were presented in static (non-changing) formats, i.e., there was no feedback or other dynamic changes over time. We are now applying CWS in real-time, dynamic environments. Thus far, the results are encouraging (see Friel et al., 2001).

Second, CWS was applied to individual performance in these studies. However, experts often work in teams. If teams are treated as decision-making units, then CWS can be used to evaluate each team. Preliminary efforts to apply CWS to team decision making have been positive.

Third, CWS assumes there are real differences in the stimulus cases to be judged. If there are no such differences, then all judgments should be the same. If patients all have the same disease, for example, they presumably will all have the same diagnosis. The problem then is that it becomes impossible to distinguish between doctors with different diagnostic ability. Therefore, there must be differences in cases before CWS can evaluate expertise.

Fourth, it is possible for CWS to produce high values for a non-expert who uses a consistent, but incorrect rule. Suppose all patients over 70 receive one diagnosis and all patients under 70 receive another. Such a strategy would produce high CWS values, i.e., consistent discriminations. But such a diagnostic strategy would clearly be inappropriate. One way around this “catch” is to ask judges to evaluate the same cases for different purposes, e.g., diagnosis (what is it?) vs. prognosis (what to do about it?). If the judgments in these two cases are the same, then the candidate is not behaving expertly – despite having a high CWS value.

Fifth, because CWS relies on repeated cases (to estimate consistency), it is necessary to take steps to prevent participants from memorizing responses and thus artificially inflating their consistency. This is potentially a problem if names or other unique identifiers are attached to the case descriptions. The solution, therefore, is to change such identifiers from replicate to replicate. Thus, a patient might be “Gerald” the first time, “William” the second time, and so forth. In our experience, such simple precautions effectively eliminate recall of previous responses.

Finally, it is important to emphasize that CWS does not determine which cues or responses to use in a study. As in any research, the dictates of good experimental design and sound methodology should guide the selection of independent and dependent variables. Once those decisions have been made, then CWS can be applied to the data. However, the adage “garbage in, garbage out” applies.

### **Discussion**

The present discussion of CWS leads to five conclusions: First, in our analyses, CWS has proved useful in evaluating expert performance. If CWS continues to be successful, it may provide the answer to the longstanding question of how to evaluate expertise in the absence of an external standard.

Second, the success of CWS across different domains is noteworthy. Beyond medical diagnosis, livestock judging, and auditing, we have applied CWS to wine judging, personnel selection, food tasting, and air traffic control. To date, CWS has worked well in each domain.

Third, beyond evaluating expert performance, CWS has provided insights into the equipment and methods used by experts. In a simulation of air traffic control, for instance, we used CWS to track changes in operator performance as a function of monitor display sizes. In such cases, the purpose is to compare performance across conditions, not to evaluate expertise.

Fourth, by focusing on discrimination and consistency, CWS may have important implications for the selection and training of novices. It is an open question whether discrimination and

consistency can be learned, or whether novices should be preselected to already have these skills. Either way, CWS offers new perspectives on skill acquisition.

Finally, CWS outperforms existing approaches to evaluating expert performance. For instance, CWS and SME ratings were compared in a reanalysis of data from an air traffic control study conducted by the Federal Aviation Administration. SME ratings of ATC behavior were moderately sensitive to changes in the airspace environment. The effect sizes were considerably larger, however, when CWS was applied to the data (Thomas, Willems, Shanteau, Raacke, & Friel, 2001).

### **Conclusions**

The present chapter has shown that prior approaches to evaluating expertise have substantial flaws. The alternate CWS approach avoids most of these pitfalls. Reanalyses of previous studies demonstrated that (1) CWS can be applied to analyze performance of experts in variety of domains, (2) expert performance is tied to the two components of CWS – discrimination and consistency, (3) CWS is sensitive to task differences and to differences in subspeciality skills, and (4) differences between expert and novice performance can be distinguished by CWS – especially when cue relevance is taken into account. Some potential applications of CWS include evaluation of expert performance, selection and training of experts, and assessment of new methods/technologies. In summary, CWS is a promising new tool for studying expert behavior.

**References**

- Ashton, A. H. (1985). Does consensus imply accuracy in accounting studies of decision making? *Accounting Review*, *60*, 173-185.
- Chase, W. G., & Ericsson, K. A. (1981). Skilled memory. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition*. Hillsdale, NJ: Erlbaum Associates.
- Chi, M. T. H. (1978). Knowledge structures and memory development. In R. S. Siegler (Ed.), *Children's thinking: What develops?* (Pp. 73-96). Hillsdale, NJ: Erlbaum.
- 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.
- Ettenson, R., (1984). *A schematic approach to the examination of the search for and use of information in expert decision making*. Doctoral dissertation, Kansas State University.
- Ettenson, R., Shanteau, J., & Krogstad, J. (1987). Expert judgment: Is more information better? *Psychological Reports*, *60*, 227-238.
- Friel, B. M., Thomas, R. P., Raacke, J., & Shanteau, J. (2001). Utilizing CWS to track the longitudinal development of expertise. In, *2001 Proceedings of the Human Factors Society*. Minneapolis, MN.
- Gardner, M. (1957). *Fads and fallacies in the name of science*. NY: Dover.
- Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart*. NY: Oxford University Press.

- 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 Consulting 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*. NY: Oxford University Press.
- Janis, I. L. (1972). *Victims of groupthink*. Boston: Houghton-Mifflin.
- Kida, T. (1980). An investigation into auditor's continuity and related qualification judgments. *Journal of Accounting Research*, 22, 145-152.
- Lykken, D. T. (1979). The detection of deception. *Psychological Bulletin*, 80, 47-53.
- Nagy, G. F. (1981). *How are personnel selection decisions made? An analysis of decision strategies in a simulated personnel selection task*. Doctoral dissertation, Kansas State University.
- Phelps, R. H. (1977). *Expert livestock judgment: A descriptive analysis of the development of expertise*. Doctoral dissertation, Kansas State University.
- Phelps, R. H., & Shanteau, J. (1978). Livestock judges: How much information can an expert use? *Organizational Behavior and Human Performance*, 21, 209-219.
- Raskin, D. C., & Podlesny, J. A. (1979). Truth and deception: A reply to Lykken. *Psychological Bulletin*, 86, 54-59.
- 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. (1989). Psychological characteristics and strategies of expert decision makers. In B. Rohrman, L. R. Beach, C. Vlek, & S. R. Watson (Eds.), *Advances in Decision Research* (pp. 203-215). Amsterdam: North Holland.

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. (1999). Decision making by experts: The GNAHM effect. In J. Shanteau, B. A. Mellers, & D. A. Schum (Eds.), *Decision science and technology: Reflections on the contributions of Ward Edwards* (pp. 105-130).

Skånér, Y., Strender, L. E., & Bring, J. (1998). How do GPs use clinical information in their judgements of heart failure? A Clinical Judgment Analysis study. *Scandinavian Journal of Primary Health Care, 16*, 95-100.

Slovic, P. (1969). Analyzing the expert judge: A descriptive study of a stockbroker's decision processes. *Journal of Applied Psychology, 53*, 255-263.

Stewart, T. R., Roebber, P. J., & Bosart, L. F. (1997). The importance of the task in analyzing expert judgment. *Organizational Behavior and Human Decision Processes, 69*, 205-219.

Thomas, R. P., Willems, B., Shanteau, J., Raacke, J., & Friel, B. (2001). Measuring the performance of experts: An application to air traffic control. In, *2001 Proceedings of Aviation Psychology*. Columbus, OH.

Trumbo, D., Adams, C., Milner, M., & Schipper, L. (1962). Reliability and accuracy in the inspection of hard red winter wheat. *Cereal Science Today, 7*.

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

Weiss, D. J., & Shanteau, J. (submitted). Performance-based assessment of expertise. Manuscript under review for publication.

**Author Notes**

Preparation of this manuscript was supported, in part, by grant 96-12126 from the *National Science Foundation*, by grant N00014-00-1-0769 from the *Office of Naval Research*, and by grant 98-G-026 from the *Federal Aviation Administration* in the Department of Transportation.

Further information about the studies and analytic procedures described in this chapter can be found at the CWS Website: <[www.ksu.edu/psych/edu](http://www.ksu.edu/psych/edu)>

We wish to thank Ward Edwards, Brian Friel, Alice Isen, and Gary McClelland for their insightful comments on prior versions of the manuscript. We also wish to acknowledge the feedback from various anonymous reviewers who have helped us clarify a number of concepts.

Correspondence concerning this project should be addressed to James Shanteau, Department of Psychology, Bluemont Hall 492, Kansas State University, Manhattan, KS 66506-5302 USA. E-mail: <[shanteau@ksu.edu](mailto:shanteau@ksu.edu)>

**Table 1**

Between-Expert Reliability Values

Higher Levels of Performance.....Lower Levels of Performance

Aided Decisions	Competent	Restricted	Random
Weather Forecasters r = .95	Livestock Judges r = .50	Clinical Psychologists r = .40	Stockbrokers r = .32
Auditors r = .76	Grain Inspectors r = .60	Pathologists r = .55	Polygraphers r = .33

Note: The values cited in this table (left to right) were drawn from the following studies: Stewart, Roebber & Bosart (1997), Phelps & Shanteau (1978), Goldberg & Werts (1966), Slovic (1969), Kida (1980), Trumbo, Adams, Milner & Schipper (1962), Einhorn (1974), and Lykken (1979).

**Table 2**

Within-Expert Reliability Values

Higher Levels of Performance.....Lower Levels of Performance

Aided Decisions	Competent	Restricted	Random
Weather Forecasters r = .98	Livestock Judges r = .96	Clinical Psychologists r = .44	Stockbrokers r = < .40
Auditors r = .90	Grain Inspectors r = .62	Pathologists r = .50	Polygraphers r = .91

Note: The values cited in this table (left to right) were drawn from the following studies: Stewart, Roebber & Bosart (1997), Phelps & Shanteau (1978), Goldberg & Werts (1966), Slovic (1969), Kida (1980), Trumbo, Adams, Milner & Schipper (1962), Einhorn (1974), and Raskin & Podlesny (1979).

**Figure Captions**

Figure 1. Judgments of probability of heart failure for five patients (listed as letters) by three GPs from Skånér et al. (1998). Open and close points show judgments for two presentations. CWS values listed for each GP.

Figure 2. Mean CWS values for breeding and slaughter judgments of gilts (female pigs) by cattle specialists and swine specialists. Values shown as a function of size, breeding potential, and meat quality (from Phelps, 1977).

Figure 3. Mean CWS values for *Going Concern* judgments by expert and novice auditors as a function of three levels of relevance for financial information cues (from Ettenson, 1984).