

IDENTIFYING EXPERTISE WITHOUT A GOLD STANDARD: FOUR APPLICATIONS

James Shanteau, Rickey P. Thomas, Brian Friel
Kansas State University
Manhattan, Kansas

David J. Weiss
California State University
Los Angeles, California

Julia C. Pounds
Civil Aeromedical Institute, Federal Aviation Administration
Oklahoma City, Oklahoma

ABSTRACT

Identification of expertise is vital to any application involving expert human performance. If a gold standard exists, then identification is direct – simply compare individuals against the standard and select whoever is closest. However, such standards are unavailable in many domains where experts work. The purpose of this paper is to present an alternate method for identifying experts in the absence of a gold standard. The method, called CWS for Cochran-Weiss-Shanteau, is illustrated with applications to four domains. In each case, CWS provided new insights into identification of expertise.

INTRODUCTION

The purpose of this paper is to present a new approach (labeled CWS for *Cochran-Weiss-Shanteau*) for identifying experts in the absence of gold standards. The measure is based on examining the behavior of individuals within their domain of expertise. That is, the individual's own decisions are used to validate his or her claims to expertise.

The remainder of the paper is divided into three sections: The first provides background. The second gives four examples of CWS applied to prior studies. The final section offers comments and conclusions.

BACKGROUND

Behavioral studies of experts began 100 years ago (Shanteau, 1999). However, a critical question remains unanswered: How do we know whether someone is an expert? If an external criterion (a “gold standard”) exists, the answer is simple. Just compare individuals to the standard. Whoever is closest is the “expert.”

The problem with this approach is that gold standards are often difficult to find, if they exist

at all, in many domains of expertise. As Gigerenzer et al. (1999) note, experts are needed for precisely those problems where standards are least likely to be found.

Prior Approaches

Prior investigators have used many methods to designate “experts” in their studies. Five of these methods will be briefly described here.

Experience. Frequently, the number of years of on-the-job experience is used as a measure of expertise – the greater the experience, the greater the assumed level of expertise. Of course, there is some correlation since any true expert will necessarily have considerable experience. Unfortunately, the converse does not hold. There are many examples of professionals with years of experience who never achieve the competence required of an expert.

For instance, Goldberg (1968) had clinical psychologists with different amounts of experience diagnose psychiatric cases. He found no relationship between experience and accuracy of diagnosis. However, he did find that confidence increased with experience.

Although there are likely to be many instances where experience is predictive of expertise, this is not universal. Thus, experience has, at best, an uncertain relationship to expertise.

Consensus. Many investigators believe that agreement between individuals can be used as a standard of expertise. Einhorn (1974), for example, argued that such agreement is a necessary condition for experts. Therefore, disagreement is a sign that one (or more) of the individuals is not an expert.

A variety of researchers have used consensus as a basis for defining gold-standard answers (Ashton, 1985). For instance, the between correlation of .95 has been used to argue for the expertise of weather forecasters (Stewart, et al, 1997). Conversely, the lack of agreement has been used to argue against the expertise of stockbrokers (Slovic, 1969).

Intuitively, between-individual agreement seems an obvious characteristic of experts. For example what patient doesn’t feel more comfortable with a second opinion that confirms a diagnosis? The problem is that false consensus can result from GroupThink (Janus, 1972) and other group biases (Weiss & Shanteau, in press). It is not unheard of for many experts to agree – and for most of them to be wrong (Milton, 1994; Shanteau, in press).

Subject Matter Experts (SMEs). One method used by many researchers (including the lead author) has been to rely on the identification of experts by individuals working in the field. Thus, professionals are asked whether someone is, or is not, an expert usually based on “over the shoulder” ratings.

In two of the studies discussed below (Phelps and Nagy), the experts were identified asking those working in the domain to identify whom they considered an expert. When there was agreement, it seemed safe to assume the individual identified was indeed expert.

Identification of experts by SMEs is a reasonable strategy to use. It is not likely that various

professionals will all identify the same person, if they are unqualified. The problem is that there could be a peer-group effect; someone better liked by their peers is more likely to get high SME ratings. In contrast, someone outside the peer group, no matter how capable, is unlikely to receive approval from SMEs. As Milton (1994) argues, pioneering discoveries are often made by individuals who were ignored by their peers.

Discrimination. Hammond (1996) has argued that ability to make subtle distinctions is a critical in experts. That is, experts often detect differences that novices cannot perceive. This means that experts not only know how to combine different sources of information; they also know how to extract that information.

Further, experts characteristically distinguish relevant from irrelevant sources of information, i.e., “separate wheat from the chaff.” The ability to determine relevance is vital for experts (Shanteau, 1992).

It seems clear that discrimination ability is necessary for expertise. However it is not sufficient since non-experts may also be able to make discriminations using some simple, but non-predictive rule. For example, using hairstyle allows for discrimination between job candidates. But it is unlikely to be predictive of worker performance.

Consistency. Einhorn (1972, 1974) argued that intra-individual (within) reliability is for expertise. That is, an expert’s judgments should be internally consistent over time. Conversely, inconsistency would be evidence that the individual is not an expert.

Shanteau (1999) reported partial support for this argument, depending on the presence of decision aids. The average consistency for weather forecasters (a decision-aided task) is quite high at .98 (Stewart, et al, 1997). For stockbrokers (an unaided task), the average consistency is less than .40 (Slovic, 1969).

The difficulty with this approach is that someone can be consistent by always following some simple, but incorrect rule. For example, by always answering “good” or “bad” to alternate questions, one can be perfectly consistent. But such answers would generally be inappropriate. Thus, internal consistency is a necessary condition, but not sufficient for expertise.

Cochran-Weiss-Shanteau (CWS). This recently developed approach assumes that an expert judge must satisfy two basic criteria. These constitute necessary, but not sufficient, conditions for presence of expertise.

The first is that an expert should be able to discriminate among the stimuli within the domain. The ability to differentiate between similar, but not identical stimuli is a hallmark of expertise (Hammond, 1996). Secondly, following Einhorn’s (1974) suggestion, internal consistency is required of an expert. Inconsistency is indicative of novices, not experts.

Both of the criteria are empirical, so that an index of expertise can be constructed purely from data. Using empirical criterion avoids the circularity inherent in approaches that rely on prior expert knowledge to identify experts.

APPLICATIONS

In this section, CWS will be used to reanalyze the results of four prior studies of experts. Each of these will illustrate one or more of the advantages of CWS.

Medical Diagnosis

A recent study by Skånér, Strender, and Bring (1998) in Sweden illustrates how expertise can be evaluated based on a set of judgments. Twenty-seven 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. There were no additional pathological findings based on further examination. Case-specific information was provided for ten cues such as age, gender, history of myocardial infarction, dyspnea, edema, and heart X-ray.

“For each vignette, the doctors were asked to assess the probability that the patient suffered from any degree of heart failure” (Skånér et al., 1998, p. 96). 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 values” (p. 96).

The authors reported: “variation between the GP’s assessments of the probability of heart failure was considerable” (p. 95). After inconclusive analyses of demographic variables, the authors could not explain the large variation between the GPs.

Selected results for two of the GPs (identified by number) are given in Table 1. The five repeated patient cases are represented by letters. The first line gives the judgments for the first presentation and the second line for the second presentation. Thus, the first judgment of Patient A by Doctor #18 is near 100; the second judgment is similar.

Table 1
CWS Calculations for Two Doctors from Skånér, et al

Doctor #18	A	B	C	D	E
<i>Replicate 1:</i>	96	18	94	95	25
<i>Replicate 2:</i>	96	12	91	98	27
Discrimination = $\sum (M_i - GM)^2 / (n-1) = 3365.15$					
Consistency = $\sum d_i^2 / n = 5.80$					
CWS = $3365.15 / 5.80 = 580.20$					
Doctor #16	A	B	C	D	E
<i>Replicate 1:</i>	89	81	85	88	83
<i>Replicate 2:</i>	88	65	79	80	85
Discrimination = $\sum (M_i - GM)^2 / (n-1) = 65.40$					
Consistency = $\sum d_i^2 / n = 36.10$					
CWS = $65.40 / 36.10 = 1.81$					

As can be seen, there is considerable variation between the two GPs. Each GP shows a distinctive pattern in terms of discrimination and consistency. Doctor #18 is highly discriminating (sizable differences between patients) and highly consistent (little difference between first and second presentations). Doctor #16 is consistent, but treats all patients similarly – all are seen as having a moderately high chance of heart failure.

Based on their data alone, we can gain considerable insight into the judgment strategies and abilities of the GPs. Doctors #18 and #16 are both consistent, but one discriminates and the other does not. We believe that without knowing anything further, most clients would prefer someone like Doctor #18, who can make clear discriminations in a consistent way. In effect, the CWS index quantifies this intuition.

CWS Analysis

CWS is illustrated in Table 1 for the two doctors selected from the Skånér et al. (1998) study. Doctor #18 has high discrimination (3,365.15) and low inconsistency (5.80), yielding a CWS value of 580.20. In isolation, we cannot say whether this is good or bad. Therefore we need to consider the CWS values for the other doctors. In comparison, Doctor #16, with low discrimination but low inconsistency, has a CWS value of 1.81.

An obtained CWS index depends upon both the individual's expertise and the particular set of stimuli used. The more the stimuli differ from each other, the easier they are to discriminate. It is therefore not meaningful to compare CWS scores for individuals who judge different stimulus sets, just as it is not meaningful to compare across different domains (i.e., expertise of doctors cannot be compared to lawyers).

For the two doctors from the Skånér et al. (1998) study, it is apparent that CWS has identified Doctor #18 as superior. It is notable that this identification occurred without any other knowledge about the doctors, the patients, or the correct standards for medical diagnosis. Thus, CWS avoids the circularity of having to know the answer before an expert can be identified.

Audit Judgment

Ettenson (1984) asked two groups of auditors to evaluate 24 financial cases described by a common set of cues. One group of 15 "expert" auditors was recruited from Big Six accounting firms in Omaha, Nebraska. For comparison, 15 "novice" accounting students were obtained from two Midwestern universities.

Every financial case was described using 16 cues, each of which was given either a high or low value. For example, *Net Income* was set at either a high or low number. For each case, participants were asked to make a *Going Concern* assessment. A fractional factorial design was used to generate 16 cases. Eight of these cases were then replicated to produce a total of 24 stimuli.

Based on feedback from an auditor collaborator, the cues were classified as "diagnostic" (e.g., *Net Income*), "partially diagnostic" (e.g., *Aging of Receivables*), or "nondiagnostic" (e.g.,

Prior Audit Results). From analysis of the fractional design, discrimination was estimated from the mean square values for each cue – high variance implies high discrimination. Inconsistency was estimated from the average of within-cell variances – low variance implies high consistency. The ratio of discrimination variance divided by inconsistency variance was computed to form separate CWS values for diagnostic, partially diagnostic, and nondiagnostic cues.

The results in Table 2 show that average CWS values decline systematically as the diagnosticity of the cues declines. For the expert group, the differences are notable, especially between diagnostic and partially diagnostic cues. For the novice group, there is a similar but less pronounced decline. More important, there is a sizable difference between experts and novices for diagnostic cues. The size of this difference is less for partially diagnostic cues, and nonexistent for nondiagnostic cues.

Table 2
CWS Values for 3 Groups of Auditors from Ettenson

Group	Relevance		
	<i>High</i>	<i>Medium</i>	<i>Low</i>
<i>Partners:</i>	5.00	2.55	1.75
<i>Seniors:</i>	4.88	2.11	1.81
<i>Grad Students:</i>	2.95	1.23	2.22

For diagnostic cues, CWS clearly distinguishes between experts and novices. Moreover, the size of difference between the groups declines for less diagnostic cues. These results show that CWS can distinguish between expert and novice groups.

Livestock Judgment

Phelps (1977) had four professional livestock judges evaluate 27 drawings of gilts – female pigs. These drawings were created by an artist to yield a 3-x-3-x-3, size x breed x meat quality, factorial design. The judges independently evaluated each gilt for *breeding quality* (how good is the animal for reproduction) and *slaughter quality* (how good is the meat from the animal.) All stimuli were presented three times, although judges were not told.

Two of the judges were internationally recognized experts in assessment of swine and were very familiar with gilts of the sort shown. The other two were internationally recognized cattle experts; although they were knowledgeable about swine, they lacked day-to-day familiarity and experience.

For breeding judgments in Table 3, swine experts produced the largest CWS values for breed and meat cues. In comparison, cattle experts produced large CWS values only for the meat cue. This apparently reflects the unfamiliarity of breed characteristics of swine by cattle judges; the more familiar meat quality characteristics, however, were readily emphasized by cattle judges.

Table 3
CWS Values for Cattle and Swine Experts from Phelps

	Cattle Experts		Swine Experts	
	<i>Breeding</i>	<i>Slaughter</i>	<i>Breeding</i>	<i>Slaughter</i>
<i>Size</i>	.00	.00	1.20	.00
<i>Breed</i>	.53	.87	1.73	.51
<i>Meat</i>	1.90	1.99	1.82	2.33

In slaughter judgments, the meat cue dominates for both swine and cattle judges. For cattle experts, there is little difference in CWS between breeding and slaughter judgments. For swine experts, however, there is a considerable difference between breeding and slaughter judgments. Thus, it appears that swine judges are more sensitive to task changes. This study therefore highlights the role that task plays in expertise.

Personnel Selection

Nagy (1981) used summary descriptions of job candidates for the position of computer programmer at a large company in the state of Washington. She asked four professional personnel selectors (experts) and 20 management (MBA) students (novices) to evaluate these candidates. Each candidate was described by two legally relevant attributes (*recommendations from prior employers* and *amount of job-relevant experience*) and three legally irrelevant attributes (*age*, *gender*, and *physical attractiveness*). Filler information from local phone books was used to supply background information, such as phone number and home address, on the application summaries.

Each participant evaluated 32 applicants (generated from a 2-x-2-x-2-x-2 factorial design) twice. Before the evaluations, participants were reminded about the legal requirements for hiring, i.e., what information should and should not be used. The importance of the five attributes was determined for each participant on a 0-100 normalized scale; average CWS values are reported for each group.

As shown for the relevant attributes (upper part of Table 4), average CWS values are nearly identical for the two groups. This is not surprising given that individuals were told immediately before the study about hiring guidelines. In contrast, CWS values for irrelevant attributes (lower part) reveal a different pattern. For professionals, CWS is near zero. In contrast, CWS values are considerably larger for students. Despite being reminded that age, gender, attractiveness, and gender are not legally allowed, MBA students had sizable CWS values for these irrelevant attributes.

Table 4
CWS Values for Personnel Selection from Nagy

Relevant Attributes			
	<i>Recommendations</i>	<i>Experience</i>	
<i>Professionals</i>	4.48	4.46	
<i>Students</i>	4.49	4.46	
Irrelevant Attributes			
	<i>Age</i>	<i>Attractiveness</i>	<i>Gender</i>
<i>Professionals</i>	.00	.46	.00
<i>Students</i>	3.35	3.26	2.59

Certainly, it is not easy to ignore something as obvious as age or gender, although that is what the legal guidelines require. Experts apparently have developed strategies to do just that. But MBA students have yet to acquire this skill. As this study illustrates, there may be tasks where CWS values for irrelevant attributes may be more diagnostic of expertise than relevant attributes.

CAVEATS AND CONCLUSIONS

Caveats

There are four caveats that deserve mention. First, the application of CWS to these prior studies is encouraging as far as it goes. However, more evidence is needed before CWS can be used by itself to identify experts. For now, it seems clear that CWS can be used as a useful supplement to other approaches, e.g., SME ratings.

Second, the stimuli used in these studies were abstractions of real-world problems. Specifically, cases were presented in static (non-changing) environments, with no feedback or dynamic changes. We are now applying CWS to complex, real-time environments.

Third, CWS was applied here to individuals whose results were combined to produce group averages. However, most experts work in teams. If teams are treated as a decision-making unit, then it is possible to apply CWS in the same way as with individuals. Preliminary efforts to apply CWS to team decision making have been encouraging.

Finally, it is possible for CWS to yield high values for non-experts who use consistent, but incorrect rules. Suppose all job candidates with short names (e.g., *Ann*) get high recommendations while all job candidates with long names (e.g., *Georgette*) get low recommendations. Because of high consistency, such an inappropriate rule would produce high CWS values.

One way around this “catch” is to ask judges to evaluate the same cases in different contexts, e.g., recommendations for a different job. If judgments are the same for both jobs, then the participant is not likely to be an expert – despite having a high CWS value.

Conclusions

The present application of CWS leads to four conclusions: First, in the analyses above, CWS proved superior to any previous approach for identifying experts. If CWS continues to be successful, it may provide an answer to the longstanding question of how to identify expertise when there is no gold standard.

Second, the success of CWS across different domains is noteworthy. In addition to medical diagnosis, auditing, livestock judging, and personnel selection, we have applied CWS to wine judging, soil judging, microworld simulations, sensory food evaluations, and air traffic control. Thus far, CWS has worked well in every domain.

Third, in addition to identifying experts, CWS has provided new insights into interpretation of previous research. In the Phelps study of livestock judges, for example, CWS clarified a longstanding question about how to distinguish between experts from closely related specialty areas.

Finally, by focusing on discrimination and consistency, CWS may have important implications for selection and training of novices to become experts. It is unclear, for example, whether discrimination and consistency can be learned, or whether novices should be preselected for these skills. Either way, CWS offers new perspectives on what it means to be an expert.

ACKNOWLEDGMENTS

Preparation of this manuscript was supported by grant 98-G-026 from the Federal Aviation Administration in the Department of Transportation. We wish to thank Ward Edwards and Alice Isen for valuable discussions regarding the substantive issues involved in the implementation of the CWS index. We also wish to thank Dr. Ylva Skånér for sharing the data from her study with us.

Correspondence concerning this project should be addressed to James Shanteau, Department of Psychology, Bluemont Hall 492, 1100 Mid-Campus Drive, Kansas State University, Manhattan, KS 66506-5302 USA. E-mail: <shanteau@ksu.edu>

REFERENCES

- Ashton, A. H. (1985). Does consensus imply accuracy in accounting studies of decision making? *Accounting Review*, *60*, 173-185.
- 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*, Unpublished doctoral dissertation, Kansas State University, Manhattan, KS.

- Gigerenzer, G., Todd, P., & the ABC group. (1999). *Simple heuristics that make us smart*. London: Oxford University Press.
- Goldberg, L. R. (1968). Simple models or simple processes? Some research on clinical judgments, *American Psychologist* 23 482-496.
- Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. NY: Oxford University Press.
- Janis, I. L. (1972). *Victims of groupthink*. Boston: Houghton-Mifflin.
- Milton, R. (1994). *Forbidden science: Suppressed research that could change our lives*. London: Fourth estate.
- Nagy, G. F. (1981). *How are personnel selection decisions made? An analysis of decision strategies in a simulated personnel selection*. Unpublished doctoral dissertation. Kansas State University: Manhattan, KS.
- Phelps, R. H. (1977). *Expert livestock judgment: A descriptive analysis of the development of expertise*, Unpublished doctoral dissertation. Kansas State University, Manhattan, KS.
- Skånér, Y., Strender, L., & Bring, J. (1998). How do GPs use clinical information in the judgments of heart failure? *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.
- Shanteau, J. (1992). Competence in experts: The role of task characteristics. *Organizational Behavior and Human Decision Processes*, 53, 252-266.
- Shanteau, J. (1999). Decision making by experts: The GNAHM effect. In: Shanteau, J., Mellers, B. A., Schum, D. A. (Eds.), *Decision science and technology: Reflections on the contributions of Ward Edwards*. Boston: Kluwer Academic Publishers, pp. 105-130.
- Shanteau, J. (in press). What does it mean when experts disagree? In E. Salas and G. Klein (Eds.), *Linking expertise and naturalistic decision making*. Hillsdale, NJ: Erlbaum.
- 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.
- Weiss, D. J., & Shanteau, J. (in press). The vice of consensus and the virtue of consistency. In J. Shanteau, P. Johnson, & K. Smith (Eds.), *Psychological explorations of competent decision making*. NY: Cambridge University Press.