ELSEVIER

European Journal of Operational Research 136 (2002) 253–263

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

www.elsevier.com/locate/dsw

Performance-based assessment of expertise: How to decide if someone is an expert or not

James Shanteau [a,*], David J. Weiss [b], Rickey P. Thomas [a], Julia C. Pounds [c]

[a] *Department of Psychology, Kansas State University, Bluemont Hall 492, 1100 Mid-Campus Drive, Manhattan, KS 66506-5302, USA*
[b] *California State University, Los Angeles, CA, USA*
[c] *Federal Aviation Administration, Oklahoma City, OK, USA*

Received 15 December 1999; accepted 15 April 2000

Abstract

The identification of an expert is vital to any study or application involving expertise. If external criterion (a "gold standard") exists, then identification is straightforward: Simply compare people against the standard and select whoever is closest. However, such criteria are seldom available for domains where experts work; that's why experts are needed in the first place. The purpose here is to explore various methods for identifying experts in the absence of a gold standard. One particularly promising approach (labeled CWS for *Cochran–Weiss–Shanteau*) is explored in detail. We illustrate CWS through reanalyses of three previous studies of experts. In each case, CWS provided new insights into identifying experts. When applied to auditors, CWS correctly detected group differences in expertise. For agricultural judges, CWS revealed subtle distinctions between subspecialties of experts. In personnel selection, CWS showed that irrelevant attributes were more informative than relevant attributes. We believe CWS provides a valuable tool for identification and evaluation of experts. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Psychology; Expert systems; Auditing; Management; Agriculture

1. Introduction

Although experts have been studied for over a century (Shanteau, 1999), there remains a critical unanswered question – how can we describe who is, and who is not, an expert? If there is an external criterion (a "gold standard"), the answer is straightforward. All we have to do is compare a would-be expert's judgments to the correct answer. If a person's answers are close to correct, then he or she is an "expert." If not, not.

This *validity*-based approach is compelling in its simplicity. Unfortunately, it is problematic in application. The difficulty is that experts are needed precisely in domains where correct answers seldom exist (Gigerenzer et al., 1999; Shanteau, 1995). Indeed, if we could compute (or look up) correct answers, why would we need an expert at all?

[*] Corresponding author. Tel.: +1-785-532-0618; fax: +1-785-532-5401.
E-mail address: shanteau@ksu.edu (J. Shanteau).

PII: S0377-2217(01)00113-8

The purpose of this paper is to explore the application of a new measure of expertise (labeled CWS for *Cochran–Weiss–Shanteau*) for identifying expertise in the absence of external criteria. The measure is based on the behavior of would-be experts by using their performance in the domain. In effect, this is a bootstrap approach in which the individual's own decisions are used to validate (or invalidate) his/her claim to expertise.

The remainder of this paper is organized into five sections: In Section 2, we review approaches used in prior research to identify would-be experts. In Section 3, we introduce our proposed approach to identification of expertise. In Section 4, we apply this approach to several previously conducted studies of experts. In Section 5, we consider caveats and restrictions that should be considered when applying CWS. Finally, we offer our conclusions about the future of CWS.

2. Prior approaches

Many approaches have been used by previous investigators to identify experts. Nine of these traditional approaches will be summarized here. We also consider the advantages and, more importantly, the disadvantages of each approach.

2.1. Experience

In many studies, the number of years of job-relevant experience is used as a surrogate for expertise. Participants with many years of experience are classified as ‘‘experts,'' while others with little experience are labeled ‘‘novices.'' On the surface, this approach appears convincing. After all, no one can function as an ‘‘expert'' for any length of time if they are totally incompetent.

Although the argument can be made that experts almost always have considerable experience, the converse does not necessarily follow. There are many examples of professionals with considerable experience who never become experts. Such individuals may even work with top experts, but they seldom rise to the performance levels required for true expertise.

In a study of grain judges, for instance, Trumbo et al. (1962) found that number of years of experience did not correlate with accuracy of wheat grading. Instead, their results showed a different trend: judges with more experience systematically overrated grain quality (an interesting form of ‘‘grade inflation''). Similarly, Goldberg (1968) asked clinical psychologists with varying degrees of experience to diagnose psychiatric patients. He found no relation between experience and accuracy of the diagnoses; however, the confidence of clinicians in their diagnoses did increase with experience.

Although there are undoubtedly instances where a positive relationship exists between experience and performance, there is little reason to expect this to apply universally. At best, experience is an uncertain predictor of degree of expertise. At worst, experience reflects seniority – and little more.

2.2. Certification

In many professions, individuals receive some form of accreditation or title as a reflection of their skill. For instance, doctors may be ‘‘board certified'' and university faculty may be ‘‘full professor.'' Generally, it is safe to say that a certified individual is more likely to be an expert than someone who is uncertified.

The problem with certification is that it is more often tied to years on the job than it is to professional performance. This can be particularly true in bureaucracies. In military photo interpretation, for instance, the rank of the individuals can vary from Sergeant to Major. Yet performance is unrelated to rank (Tod Levitt, personal communication).

Another example occurs in the Israeli Air Force, where the lead pilot in a battle is identified by skill rather than rank – that means a General may follow a Captain. This has been cited as one reason for superiority of the Israelis in air combat against Arab Air Forces (where lead pilots are usually determined by rank). The Israelis recognized that talent is not always reflected by formal certification (R. Lipshitz, personal communication).

Another problem with certification is the ‘‘ratchet up effect'' – people generally move up the certification ladder, but seldom down. Once certified, the recipient is accredited for life. Even if the skill level of the individual suffers a serious decline, the title or rank remains. (Just ask students about the teaching ability of some senior professors.)

2.3. Social acclamation

One method used by many researchers (including the present authors) has been to rely on identification of experts by people working in the field. That is, professionals are asked whom they consider to be an expert. When there is some agreement about the identification of such an individual, that person is then labeled an expert by "social acclamation."

In her analysis of livestock judges, for example, Phelps (1977) asked professionals in agriculture whom they considered the best. From their answers, she identified four top livestock judges to be the experts in her investigation (for further details on this study, see below).

Absent other means of identifying experts, acclamation is a reasonable strategy to follow. It is unlikely that multiple professionals working in a field would identify the same unqualified person as an expert. If they agree, it seems safe to assume that the agreed-upon person is an expert. The problem with this approach is a "popularity effect" – someone better known by his or her peers is more likely to be identified as an expert. Meanwhile, another person outside the peer group is unlikely to be seen as an expert – even though that person may be on the cutting edge of new knowledge. Indeed, those who make new discoveries in a field are frequently unpopular in the eyes of their peers at the time of their breakthroughs.

2.4. Consistency (within) reliability

Einhorn (1972, 1974) argued that intra-person (within) reliability is a *necessary condition* for expertise. That is, an expert's judgments should be internally consistent. Conversely, inconsistency would be prima facie evidence that the person is not an expert.

Table 1 lists within-person consistency values from eight prior studies of experts. The four vertical categories correspond to a classification of task difficulty proposed by Shanteau (1999). There are two domains listed for each category, with internal consistency correlations. For example, the average consistency for weather forecasters (a decision-aided task) is quite high at 0.98. For stockbrokers (an unaided task), the average consistency is less than 0.40.

As might be expected, aided tasks produce higher internal consistency values than unaided tasks. To a first approximation, therefore, it appears that intra-person reliability corresponds closely to the performance level of experts in different domains.

The difficulty with this approach is that someone can be consistent by following some simple, but incorrect rule. As long as the rule is followed routinely, the person's behavior will exhibit high consistency. For example, by always answering "yes" and "no" to alternate questions, one can be perfectly repeatable. But such answers would generally be inappropriate. Thus, internal consistency

Table 1
Reliability (consistency) values across levels of expert performance[a]

Highest levels of performance		Lowest levels of performance	
Aided decisions	Competent	Restricted	Unaided decisions
Weather forecasters $r = 0.98$	Livestock judges $r = 0.96$	Clinical psychologists $r = 0.44$	Stockbrokers $r = <0.40$
Auditors $r = 0.90$	Grain inspectors $r = 0.62$	Pathologists $r = 0.50$	Polygraphers $r = 0.91$

[a] The values cited in this table (left–right and top–bottom) were drawn from the following: Stewart et al. (1997), Phelps and Shanteau (1978), Goldberg and Werts (1966), Slovic (1969), Kida (1980), Trumbo et al. (1962), Einhorn (1974), and Raskin and Podlesny (1979).

is a necessary condition – an expert could hardly behave randomly – but not sufficient for expertise.

2.5. Consensus (between) reliability

Einhorn (1972, 1974) argued that agreement between individuals is a necessary condition for expertise. That is, he believed that experts in a given field should agree with each other (also see Ashton, 1985). If there is disagreement, then it suggests that one, some, or all of the would-be experts are not really what they claim to be.

Table 2 lists average between-expert correlations for the same studies listed in Table 1. For instance, the consensus correlations for weather forecasters and stockbrokers are 0.95 and 0.32, respectively. Except for pathologists, the consensus values are similar to, but lower than, the corresponding consistency values in Table 1.

Livestock judges and polygraphers display quite different consistency and consensus results. Further analysis reveals that there are several schools of thought in these domains about how to make decisions. Thus, experts from each school may be internally consistent, but show sizable disagreement with experts from another school. This could explain the discrepancy between the high consistency values and the low consensus values in these two domains.

On the surface, consensus appears to be a compelling property for experts. After all, we feel quite uncomfortable when two or more experts (such as doctors) argue about which is the correct procedure to follow. When the experts agree, on the other hand, we feel more comfortable with the mutually agreed-upon course of action.

The problem with consensus is that agreement can result from premature closure, e.g., *groupthink* (Janis, 1972). There are many illustrations where the best answer was not the one identified by a group of experts because they focused initially on an inferior alternative. Thus, they become blind to better options. Therefore, many experts may agree – but they may all be wrong (Shanteau, in press; Weiss and Shanteau, in press).

2.6. Discrimination ability

Hammond (1996) and others have pointed out that the ability to make fine discriminations between similar, but not equivalent, cases is a defining skill of experts. That is, an expert must be able to perceive and act on subtle differences that a non-expert may often overlook. In the study of livestock judges by Phelps described below, the researchers were able to develop quantitative models of the experts' judgments. However, it proved impossible for these researchers to apply the models to actual livestock due to the difficulty of perceiving the appropriate characteristics of animals. Thus, knowing *how* to combine information is of no value without knowing *what* information to combine.

Although it seems clear that discrimination is a necessary condition for expertise, there is a catch. A non-expert may well differentiate between cases using some easily identifiable, but irrelevant attribute. For instance, it is easy to distinguish between livestock based on the length or curliness of

Table 2
Reliability (consensus) values across levels of expert performance[a]

Highest levels of performance		Lowest levels of performance	
Aided decisions	Competent	Restricted	Unaided decisions
Weather forecasters $r = 0.95$	Livestock judges $r = 0.50$	Clinical psychologists $r = 0.40$	Stockbrokers $r = 0.32$
Auditors $r = 0.76$	Grain inspectors $r = 0.60$	Pathologists $r = 0.55$	Polygraphers $r = 0.33$

[a] The values cited in this table (left–right and top–bottom) were drawn from the following: Stewart et al. (1997), Phelps and Shanteau (1978), Goldberg and Werts (1966), Slovic (1969), Kida (1980), Trumbo et al. (1962), Einhorn (1974), and Lykken (1979).

their tails. However, tail characteristics play no role in the meat quality of farm animals (Bill Able, personal communication). Thus, discrimination ability is a necessary, but not sufficient, condition for identifying experts.

2.7. Behavioral characteristics

Research by (Abdolmohammadi and Shanteau, 1992; also see Shanteau, 1989) found that expert auditors share many common behavioral characteristics. Some examples are *self-confidence, creativity, perceptiveness, communication skills*, and *stress tolerance*. A complete list of characteristics (along with their definitions) appears in the original paper.

Because many experts exhibit such traits, Abdolmohammadi and Shanteau proposed that behavioral characteristics might be used to develop a "trait profile" of experts. If appropriate tests can be identified or constructed, then would-be experts would take such tests. Those that score closest to the profile of established experts would then become potential experts.

Although this approach has considerable potential, there are three critical problems. First, the required tests for several of these characteristics do not exist, e.g., Communication Skills or Tolerance of Stress. Second, even if they did, the tests would have to be normalized for a domain (e.g., auditors). Third, the extent to which non-experts may also share these same characteristics is unclear. Thus, although this approach holds promise, more work is needed before experts can be identified using their behavioral characteristics.

2.8. Knowledge tests

In studies of problem solving or game-playing experts are often identified based on tests of factual knowledge. For example, Chi (1978) used knowledge about dinosaurs to separate children into experts and novices.

Knowledge of relevant facts is clearly a prerequisite for expertise. Someone who knows nothing about a domain will be unable to make competent decisions. Yet, knowledge alone is not sufficient to establish that someone is an expert. In the Chi study, for example, knowledge about different types of dinosaurs is not enough to know what they ate, where they lived, how long they survived, or why they died out.

The problem is that it takes more than knowledge of facts for expertise. It is also necessary to see which facts to apply in a given situation. In most domains, that is the hard part.

2.9. Creation of experts

In certain contexts, it is possible for experts to be "created" through extensive training by researchers. This approach has significant advantages, including the fact that development of expertise can be studied longitudinally. Moreover, the skills learned are under direct control of researchers.

One notable example of this approach is a student who worked with William Chase at Carnegie-Mellon University to enhance his short-term memory span (Chase and Ericsson, 1981). Because the student was a track athlete, he learned to translate groups of digits into times for various running distances. When asked to retrieve the digits, he recalled the times in clusters tied to running. Using this strategy, the student broke the old record for short-term memory span of 18 digits established by a German mathematician. The new record – over 80! (Other students since have extended the record beyond 100.)

Experts can be created in this way for certain narrow tasks, e.g., to play computer games or work in a simulated microworld environment. In most realms of expertise, however, a broad range of skills is required based on years of training and experience. For instance, becoming a medical doctor can take a dozen years just to get started. Obviously, training students for a few months cannot simulate such expertise.

3. A new approach

As the preceding survey shows, many approaches have been advanced for identifying experts. Each of these approaches, however, has one or more serious flaws. No generally acceptable approach exists at the present time. To fill this gap, the two senior authors (Weiss and Shanteau, submitted) proposed a new approach for defining expertise. They combined two necessary, but not sufficient, measures, into a single index.

First, they agreed with Hammond (1996) that *discrimination* is critical for an expert. The ability to differentiate between similar, but not identical, cases is a hallmark of expertise. That is, experts perceive and act on subtle distinctions that others miss. Second, they followed Einhorn's (1974) suggestion that *consistency*, or within-person reliability, is necessary in an expert. If someone cannot repeat their judgment in a similar situation, then they are unlikely to be an expert.

Discrimination refers to a judge's differential evaluation of different stimulus cases. Consistency refers to a judge's evaluation of the same stimuli over time; inconsistency is its complement.

3.1. CWS ratio

As shown in Eq. (1), Weiss and Shanteau combine discrimination and consistency into a ratio. The CWS ratio will be large when a judge discriminates consistently, but will be small if the judge either discriminates less or has lower consistency.

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

Our construction of this index parallels Cochran's (1943) suggestion to use a ratio of variances to assess the quality of a response instrument. (Another reason for using variance ratios is that they are asymptotically efficient (I.R. Goodman, personal communication).) Cochran argued that an effective instrument should allow participants to express perceived differences among stimuli in a consistent way. We view an effective expert in the same way. We acknowledge our intellectual debt to Cochran by referring to our performance-based index as CWS.

The intuition underlying the index is that a good measuring tool necessarily has a high CWS ratio. That is, a proper instrument yields different measures for different objects, and gives the same measure whenever it is applied to the same object. A ruler, for example, discriminates among objects of varying length, and produces identical scores for the same objects. Thus, a proper measuring instrument will produce a high CWS value as defined in Eq. (1).

Similarly, an expert must be both discriminating and consistent. It is easy to display one or the other, but hard to do both. One can show discrimination by generating a wide variety of responses over stimuli; one can exhibit consistency by repeating the same response to all stimuli. But adopting either of these strategies alone means that the other entity will be lost. To display both properties simultaneously requires careful assessment of the stimuli, the essence of expert judgment.

3.2. Using CWS

CWS can be estimated by asking would-be experts to make judgments of a series of stimulus cases; this allows for assessment of their discrimination ability. In addition, at least some of the cases should be repeated; this allows for assessment of their consistency.

Discrimination and inconsistency values can be estimated using a variety of analytic procedures, such as analysis of variance or multiple regression. It is important to emphasize that the use of ratios is descriptive, not inferential. That is, CWS is more of a qualitative tool than a quantitative tool. There are no comparisons to statistical tables and no determinations of significance. Rather, CWS is used to establish that someone behaves more (high value) or less (low value) like an expert.

To rank-order two (or more) would-be experts, CWS ratios can be compared using a procedure developed by Schumann and Bradley (1959). This allows the researcher to determine whether one individual is performing better than another (Weiss, 1985).

4. Reanalyses of prior studies

In this section, we apply CWS to three previous studies of experts. By reanalyzing these results, we hope to show the utility of CWS in a variety of contexts.

4.1. 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. The expert group included audit seniors and partners, with 4–25 years of audit experience. For comparison, 15 novice accounting students were obtained from two large 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; participants were not told that some cases were identical. The order of presentation of cases was randomized.

Based on feedback from an auditor collaborator, the cues were classified as either "diagnostic" (e.g., *net income*), "partially diagnostic" (e.g., *aging of receivables*), or "non-diagnostic" (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 non-diagnostic cues.

The results in Table 3 show that average CWS values decline systematically as the diagnosticity of the cues declines. For the expert group (first row in Table 3), the differences are notable, especially between diagnostic and partially diagnostic cues. For the novice group (second row in the table), 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 non-existent for non-diagnostic cues.

Table 3
Average CWS values for two groups of auditors with three categories of cues[a]

	Diagnostic	Partially diagnostic	Non-diagnostic
Experts	13.10	6.42	3.32
Novices	8.08	5.13	3.03

[a] Results based on a reanalysis of Ettenson (1984).

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.

4.2. 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 × 3 × 3, size × breeding × 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 that they were being shown the same drawings.

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

For breeding judgments (upper panel in Table 4), swine experts produced the largest CWS values for breeding and meat cues. In comparison, cattle experts produced large CWS values only for the meat cue. This apparently reflects the unfamiliarity of breeding characteristics of swine by

Table 4
Average CWS values for swine judgments for two types of livestock experts[a]

	Size	Breeding	Meat
Breeding judgments			
Swine experts	15.9	53.8	65.6
Cattle experts	< 1.0	3.4	79.2
Slaughter judgments			
Swine experts	< 1.0	3.2	212.7
Cattle experts	< 1.0	7.5	98.0

[a] Results based on a reanalysis of Phelps (1977).

cattle judges; meat quality characteristics, however, were readily emphasized by cattle judges.

For slaughter judgments (lower panel in Table 4), the meat cue dominates for both swine and cattle judges. However, there is over a 2-to-1 difference in the magnitude of CWS for meat between swine and cattle judges. Breeding and size dimensions were small for both types of judges.

Interestingly, for cattle judges, there is little difference in CWS between breeding and slaughter judgments. For swine judges, however, there is a considerable difference between breeding and slaughter judgments, especially for the breeding cue. Thus, it appears that swine judges are more sensitive to changes in the task. In all, CWS provides a revealing picture of the difference between these two highly skilled types of experts. This study also highlights the role that specific tasks play in expertise.

4.3. Personnel hiring

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 students (novices) to evaluate these candidates. Each candidate was described by legally relevant attributes (*recommendations from prior employers* and *amount of job-relevant experience*) and 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 $2 \times 2 \times 2 \times 2 \times 2$ 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 can be seen for the relevant attributes (upper panel in Table 5), average CWS values are nearly identical for the two groups. This is not surprising given that participants were told immediately be-

Table 5
Average CWS values for two groups of personnel selectors[a]

	Recommendations	Experience	
Relevant attributes			
Professionals	88.25	86.17	
Students	88.81	86.88	
	Age	Attractiveness	Gender
Irrelevant attributes			
Professionals	0.99	1.58	0.00
Students	28.12	25.19	13.32

[a] Results based on reanalysis of Nagy (1981).

fore the study about hiring guidelines. In contrast, CWS values for irrelevant attributes (lower panel) reveal a different pattern. For professionals, CWS approaches zero (as it should). In contrast, CWS values are considerably larger for students. Despite being reminded that age, gender, and attractiveness are not legally allowed, business students had sizable CWS values for these irrelevant attributes. Certainly, it is not easy to ignore something as obvious as age or gender, although that is what the legal guidelines require. Experts, however, apparently have developed strategies to do just that. Thus, there are tasks where CWS values for irrelevant attributes may be more diagnostic of expertise than relevant attributes.

5. Caveats

There are five caveats and precautions that deserve mention. First, the application of CWS to these three 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 is clear that CWS can be used as a useful supplement to other approaches, e.g., social acclamation.

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/temporal changes. We are now applying CWS in 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.

Fourth, CWS assumes that there are real differences in the stimuli to be judged. If the stimuli are not different, then there is nothing to discriminate. If multiple patients have the same disease, for instance, then there will be no differential diagnoses. Therefore, there must be a range of stimuli before CWS can be used to identify experts.

Finally, it is possible for CWS to yield high values for non-experts who use a consistent, but incorrect rule. 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 as before, then the participant is not likely to be an expert – despite having a high CWS value.

6. Conclusions

The present application of CWS leads to five conclusions: First, in the analyses above, CWS proved superior to any previously proposed approach for identifying experts. If CWS continues to be successful, it may provide an answer to the long-standing question of how to identify expertise in the absence of external criteria.

Second, the success of CWS across different domains is noteworthy. In addition to auditing, livestock judging, and personnel selection, we have applied CWS to wine judging, medical decision making, 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 long-standing question about how to distinguish between experts from closely related specialty areas.

Fourth, 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.

Finally, we are now applying CWS to data sets where there is no prior information about the relevance of attributes. The question is whether CWS can identify experts in the absence of any knowledge of what is relevant and what is irrelevant. In preliminary analyses, the differences do not appear to be as large as shown in the present tables. However, CWS does consistently separate experts from non-experts. In all, the future for CWS looks hopeful.

Acknowledgements

Preparation of this manuscript was supported, in part, by grant 96-12126 from the *National Science Foundation* and by grant 98-G-026 from the *Federal Aviation Administration* in the *Department of Transportation* (in the USA).

References

Abdolmohammadi, M.J., Shanteau, J., 1992. Personal characteristics of expert auditors. Organizational Behavior and Human Decision Processes 58, 158–172.

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: Anderson, J.R. (Ed.), Cognitive Skills and Their Acquisition. Erlbaum Associates, Hillsdale, NJ, pp. 141–189.

Chi, M.T.H., 1978. Knowledge structures and memory development. In: Siegler, R.S. (Ed.), Children's Thinking: What Develops? Erlbaum Associates, Hillsdale, NJ, pp. 73–96.

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. Oxford University Press, London.

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.

Hammond, K.R., 1996. Human Judgment and Social Policy. Oxford University Press, New York.

Janis, I.L., 1972. Victims of Groupthink. Houghton-Mifflin, Boston.

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. 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.

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: Rohrmann, B., Beach, L.R., Vlek, C., Watson, S.R. (Eds.), Advances in Decision Research. North-Holland, Amsterdam, pp. 203–215.

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

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. Kluwer Academic Publishers, Boston, pp. 105–130.

Shanteau, J., in press. What does it mean when experts disagree? In: Salas, E., Klein, G. (Ed.), Linking Expertise and Naturalistic Decision Making. Erlbaum Associates, Hillsdale, NJ.

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.

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., in press. The Vice of Consensus and the Virtue of Consistency. In: Shanteau, J., Johnson, P., Smith, C. (Eds.), Psychological Explorations of Competent Decision Making. Cambridge University Press, New York.

Weiss, D.J., Shanteau, J., submitted. Empirical assessment of expertise.