

DO JUDGMENTS ALONE PROVIDE SUFFICIENT INFORMATION TO DETERMINE THE EXPERTISE OF THE JUDGE WHO MADE THEM?

David J. Weiss
California State University, Los Angeles
Los Angeles, California

James Shanteau
Kansas State University
Manhattan, Kansas

ABSTRACT

Our goal is to derive an empirical measure of expert judgment. In the absence of accuracy standards, it is difficult to determine whether an expert judge is performing expertly. We partition the tasks that experts do into four categories, with evaluation being the primary function underlying all expertise. Viewing the evaluator as a measuring instrument, we propose that two necessary characteristics of an expert are the ability to discriminate among different stimuli in the domain and to be consistent in judgments of the same stimulus. We combine measures of those characteristics to form a ratio, the CWS index of expertise.

INTRODUCTION

Suppose the International Olympic Committee needs to select judges for a sport new to Olympic competition. How should the selection be carried out? For a truly new event, there would be no standards, no history, and no established group of potential judges. Yet, the competition cannot take place without judges. How can we tell if a candidate for the position of judge is a qualified expert?

The challenge is that we generally have no objective criterion for excellence. How can we decide whether the judge is accurate when we cannot independently decide on the best painting, wine, or figure skating routine? How do we determine whether a prospective employee was indeed the best choice? If there were an external criterion, a *gold standard*, then the problem would be straightforward. All we would have to do is to compare the candidate's answers or actions to the gold standard. Some details remain – how shall the scoring system penalize errors of various magnitudes – but basically, the problem is straightforward. The closer to the standard, the more expert is the candidate.

Unfortunately, in most important areas calling for expert judgment, this attractive solution is not unavailable. The gold standard is unknown or inaccessible. Experts are most likely to be needed precisely in those domains where correct answers do not exist (Hoffrage & Gigerenzer, 1998; Shanteau, 1995). In such situations, various pragmatic approaches to identifying those who are among the best in their domain are used. In practice, experts are often self-identified, acknowledged by virtue of a title or rank, or chosen based on their years of experience. Thus, the identification of experts is often a subjective, even a political, process.

Categories of Expertise

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

We argue that *evaluative skill* is the basic cognitive ability that characterizes all these areas of expertise. Whatever the task, therefore, the expert must attend to relevant aspects of the situation and decide what needs to be done. It is this common element, evaluation, that our index is designed to capture. As shown in Table 1, what distinguishes the categories is what the expert must do after the evaluation has been carried out.

Table 1

Evaluation + Qualitative or quantitative expression	= Expert Judgment
Evaluation + Projection	= Expert Prediction
Evaluation + Communication	= Expert Instruction
Evaluation + Execution	= Expert Performance

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

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

The expert instructor needs to be able to *communicate* judgment strategies to novices. The skills required include to the ability to break down the process into comprehensible sub-units, to explain the requisite steps, to illustrate the appropriate behavior, to observe student performance and provide feedback, and to motivate students. An expert instructor, then, must be able both to evaluate and to communicate. An expert critic requires similar skills.

The performance expert must add *execution* to the requisite evaluation. In general, motor skills such as strength, coordination, dexterity, and stamina as well as evaluation are required to exhibit performance expertise. Basketball is a good illustrative domain. Michael Jordan's greatness as a basketball player (Halberstam, 1999) starts with the ability to evaluate quickly the complex situation on the court, thereby leading to a correct decision whether (and where) to shoot, pass, or dribble. In addition, he has the ability to carry out the selected maneuver. Many professional players have athletic gifts similar to Jordan's, but few seem to have the accompanying cognitive skills of evaluation.

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

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

Identifying experts in the performance category has been seen as straightforward; the runner wins the race, the basketball goes in the basket, and the winning golfer has the lowest score. An obvious criterion of successful performance is available. Similarly, expert predictors may be certified according to the accuracy of their predictions. Expert weather forecasters stand out because their predictions match actual weather (Stewart, Roebber, & Bosart, 1997). Assessment of violence risk among incarcerated offenders can be compared to post-release violent behavior (Swets, Dawes, & Monahan, 2000). Financial predictions may be compared with actual market outcomes (Camerer & Johnson, 1991). Thus, identifying expert predictors would seem simple: compare the candidate's responses with the outcome measures, and select those with the highest correspondence. When there exists a criterion measure that has high face validity and is easily obtained, predictive skill is quantifiable.

However, outcome measures may not always reflect the underlying degree of expertise. The fastest runner may not win because a shoe came off. The ball may not go in the basket because a defender arrives suddenly to block the shot. The best golfer may not win because there is an unexpected clump of grass. A market projection may fail because a surprise governmental policy changes economic conditions. Observation periods are generally not long enough that such "minor" aberrations can be overcome by averaging. In general, outcome measures are hazardous indices for specific behaviors (Weiss, Walker, & Hill, 1988).

Of the four categories, expert judgment is likely to be the easiest to assess, because there are fewer perturbations that might mask the expertise when the candidate is asked merely to express an evaluation. Therefore, we begin the task of empirically measuring expertise with a proposal for an index of demonstrated judgmental proficiency.

Previous Proposals for Quantitative Assessment

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

One limitation of this approach is that high consistency can be obtained by someone following a simple, but incorrect, rule. As long as the rule is followed precisely, the person's behavior will exhibit high consistency. For example, by always answering "yes" and "no" to alternate questions, one can be perfectly reliable. Obviously, such answers would generally be inappropriate. Thus, internal consistency is a necessary condition – an expert could hardly behave randomly – but it is not sufficient for defining expertise.

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

On the surface, consensus appears to be a compelling property for experts. After all, patients feel comfortable when two or more experts (such as medical doctors) agree about which procedure to follow. When the experts disagree, on the other hand, patients feel uncomfortable committing to a course of action.

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

Our view is that consensus in any guise is an inappropriate criterion for expertise (Shanteau, in press; Weiss & Shanteau, in press). The problem with consensus is that the agreement can result from premature closure, e.g., groupthink (Janis, 1972) or other group biases. There are many illustrations where the best answer was not the one identified by a group of experts because they collectively focused on an inferior alternative.

There is a deeper concern. As Bertrand Russell put it, "Even when all the experts agree, they may be wrong." This is the problem of *false consensus*. Such historical examples as flat-earth, ether in physics, phrenology in medicine, and the Rorschach test highlight the danger of reliance upon consensus as a basis for determining expertise (Gardner, 1957).

The Core of Expertise

We propose that an expert judge must satisfy two essential criteria. These constitute necessary, but not sufficient, conditions for expertise. The first is that an expert should be able to discriminate among the stimuli within the domain. The expert must see distinctions that others may miss. The ability to differentiate between similar, but not identical stimuli is a hallmark of expertise (Hammond, 1996). Secondly, we follow Einhorn's (1974) suggestion that internal consistency is required of an expert. Inconsistency is indicative of novices, not experts.

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

The CWS Index

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

$$\text{CWS} = \frac{\text{Discrimination}}{\text{Inconsistency}}$$

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

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

Like a good measuring instrument, the expert must be both discriminating and consistent. It is easy to display either quality by using a simple response strategy, one that requires little knowledge of the stimulus objects. One can show discrimination simply by generating a wide variety of responses; one can exhibit consistency by making the same response for all cases. But adopting either of these strategies alone guarantees that the other quality will be lost. To be able to incorporate both qualities simultaneously, on the other hand, requires accurate and consistent assessment of stimuli, the essence of expert judgment.

The CWS index tries to capture what physicists (Taylor, 1959) call the "resolving power" of the expert. The definition does not incorporate preference. A wine judge who rates the white wines in the set over the red wines can score as well as one who gives higher ratings to the reds (Weiss, 1980).

Calculating CWS

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

Effective discrimination implies that as the stimulus changes, the evaluation changes accordingly. High inconsistency implies that repeated evaluations of the same stimulus differ considerably. Both of these constructs may be viewed as statistical dispersion, since it is the extent of differences that is crucial. We have used the variance among mean responses to different stimuli (MS_{Stimuli}) as the estimate of discrimination and the variance among responses to the same stimulus ($MS_{\text{Replications}}$) as the measure of inconsistency. Variances, with their heavy weighting of large discrepancies, have traditionally been used by statisticians to capture precision of measurement (Grubbs, 1973).

Implementation Issues

The basic task for our expert is to appraise each of a set of stimulus objects repeatedly. For ephemeral stimuli such as an athletic performance, we could make a recording that preserves the information an expert uses. This allows the same objects to be presented more than once to each individual. A factorial design is often convenient as it automatically allows the estimate of inconsistency to be based upon variability sampled across the stimulus range, but it is not required.

Repetition within an individual may require special care on the part of the examiner. If the candidate remembers the previous response given to a particular stimulus, succeeding responses will not be independent; thus the measure of inconsistency may be misestimated. When stimuli are identifiable and memorable, spacing trials over time or relabeling should be considered. Independence of observations is a customary assumption in experimental work.

It is also necessary that stimuli vary in perceptible ways. This may not be trivial to achieve, since determination of the extent of variation might require an expert and we may not wish a priori to assume that we have identified one. These details of stimulus variation are crucial to our ability to establish expertise. If the objects vary too little, then no one will be able to discriminate among them. On the other hand, if the objects vary too much, then all candidates will discriminate perfectly, and no one will appear to be any better than anyone else. Therefore, we want a fine-grained, but differentiable, set of stimuli.

As a practical matter, the best course of action may be simply to begin with a wide ranging set of stimuli, planning to refine the selection with further research. Our experience has been that this is not a severe problem. Most researchers routinely choose a wide set of stimulus objects in their investigations of experts, although there is a school that favors a "difficult case" approach and thereby uses only a few extreme stimuli (Hoffman, Shadbolt, Burton, & Klein, 1995).

These requirements highlight the value of controlled testing for expertise. An expert may not show to advantage in everyday situations, because insufficient opportunities present themselves. The mediocre physician can do an adequate job of routine diagnoses, because most patients who come to the clinic have common ailments. Expertise might only be exhibited when rare problems come along. The “good pickup” by a diagnostician is the identification of an obscure condition that others might miss. Without a sufficient range of cases, therefore, the identification of expertise may be difficult or impossible for any approach.

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

Trade-off Relationship

Configuring discrimination and inconsistency in a ratio format highlights our perspective that the two quantities trade off. Consider an expert who is urged to emphasize consistency, as might be the case if an internist were asked to triage patients in an emergency room. We expect to see less discrimination compared to the diagnoses made in the internist’s normal practice. Conversely, an expert who is asked to be more discriminating, as might happen if a university were to switch from letter grades to numerical grades, will show less consistency.

It can be informative to look at discrimination and inconsistency separately. Suppose that for identical stimuli, a group of trainees yields similar discrimination measures and dissimilar inconsistency measures. Remediation designed to improve discrimination might be undertaken; of course, one would have to ensure that the desired improvements did not come at the expense of consistency.

Cautions and Limitations

As stated previously, a CWS index can only be interpreted relatively, not absolutely. That is, CWS is meaningful only in a comparative sense, i.e., it can be used to say which of two candidate experts is performing better. The distribution of expertise within the population is likely to vary across domains. If true expertise is rare for the judgments we request, we may not include any experts in the study. Hence, the identified “experts” may not really be very expert.

We need to be careful about certification. Expertise is measured in a specific setting, with specific stimuli and a specific task. Someone who excels in one context may not excel in others that seem similar. Applying the term “expert” to a person is a shorthand description of a set of results rather than a characterization of the person. Talent and training may combine to yield a person we label as expert, but it must be kept in mind that the label is a generalization. It is the behavior that is or is not expert. If a baseball umpire forgets his glasses, or has a hangover, and thereby miscalculates the pitches, is he still an “expert”?

There is a structural danger inherent in the CWS approach. A “correct” judgment is a weighted combination of assessed values on the observed features of the stimulus. Relevant aspects should receive high weights; irrelevant aspects should receive zero weights. Because in general we do not presume domain knowledge, we can be misled if a candidate attends consistently to inappropriate stimulus features. A figure skating judge who evaluates the contenders primarily on the basis of, say, appearance (weighting costume and hair style heavily) would be deemed an expert according to our index - if those attributes were used to discriminate consistently among the athletes. Clearly, this is not real expertise for the task of judging athletic performance.

CONCLUSION

The approach outlined in this paper provides a widely useful approach to selection of experts that is based solely on their performance. CWS does *not* require any *a priori* knowledge of right (or wrong) answers. Instead, our approach is based on combining two necessary conditions, discrimination between similar stimuli and consistency in evaluating the same stimuli on repeated occasions.

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 Julia Pounds and Rickey Thomas for valuable discussions regarding the substantive issues involved in the implementation of the CWS index. We are especially grateful to Ward Edwards for his insights.

REFERENCES

- Ashton, A. H. (1985). Does consensus imply accuracy in accounting studies of decision making? *Accounting Review*, *60*, 173-185.
- Bolger, F., & Wright, G. (1992). Reliability and validity in expert judgment. In G. Wright & F. Bolger (Eds.), *Expertise and decision support* (pp. 47-76). New York: Plenum Press.
- Camerer, C. F., & Johnson, E. J. (1991). The process-performance paradox in expert judgment: How can experts know so much and predict so badly? In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 195-217). New York: Cambridge University Press.
- 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.
- Grubbs, F. E. (1973). Errors of measurement, precision, accuracy and the statistical comparison of measuring instruments. *Technometrics*, *15*, 53-66.
- Halberstam, D. (1999). *Playing for keeps: Michael Jordan and the world he made*. New York: Random House.

- Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. New York: Oxford University Press.
- Hoffman, R. R., Shadbolt, N. R., Burton, A. M., & Klein, G. (1995). Eliciting knowledge from experts: A methodological analysis. *Organizational Behavior and Human Decision Processes*, 62, 129-158.
- Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, 73, 538-540.
- Janis, I. L. (1972). *Victims of groupthink*. Boston: Houghton-Mifflin.
- 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. (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.
- Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, 1, 1-26.
- Taylor, L. W. (1959). *Physics: The pioneer science* (Vol. 2). New York: Dover.
- Uebersax, J. S. (1993). Statistical modeling of expert ratings on medical treatment appropriateness. *Journal of the American Statistical Association*, 88, 421-427.
- Uebersax, J. S., & Grove, W. M. (1990). Latent class analysis of diagnostic agreement. *Statistics in Medicine*, 9, 559-572.
- Uebersax, J. S., & Grove, W. M. (1993). A latent trait finite mixture model for the analysis of rating agreement. *Biometrics*, 49, 823-835.
- Weiss, D. J. (1980, August). *Training the expert judge*. Paper presented at the Mathematical Psychology Meeting, Madison, WI.
- 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*. New York: Cambridge University Press.
- Weiss, D. J., Walker, D. L., & Hill, D. (1988). The choice of a measure in a health-promotion study. *Health Education Research: Theory and Practice*, 3, 381-386.