

Running head: VICE OF CONSENSUS

The Vice of Consensus and the Virtue of Consistency

David J. Weiss

California State University, Los Angeles

James Shanteau

Kansas State University

4/24/01

Abstract

Agreement among professionals is often considered as evidence that a decision is correct.

The reasoning behind this principle is that it is unlikely that independent experts would all choose a wrong alternative. Concurring opinions in medicine, consensus on faculty committees, and unanimous appeals court decisions exemplify how the principle makes us confident. The expertise of someone who disagrees with the consensual answer is deemed questionable. We challenge this view, arguing that agreement with other experts is neither necessary nor sufficient for expertise.

The Vice of Consensus and the Virtue of Consistency

Our title is adapted from that of an essay by G. B. Shaw (1956), in which he compared two closely related practices, gambling and insurance, and reached opposing conclusions about their merits. In Shaw's view, gambling is fundamentally ruinous, whereas insurance protects the citizenry. Here we criticize the use of consensus as a criterion in science, arguing that it is a poor surrogate for consistency. Consistency, on the other hand, is a necessity.

Consensus is viewed as a pathway to truth. When we receive a disturbing medical evaluation, we rush to seek a second opinion. If the second opinion differs, then we feel justified in questioning the accuracy of the first. If that second judgment is consistent with the first, we are more likely to accept the unhappy situation. If the prognosis is really terrible, we may seek a third opinion, again hoping to disconfirm the opinion. The rationale for this is that at an intuitive level, we are intuitive statisticians employing the binomial distribution. The probability that k independent judges reach the same wrong conclusion is q^k , where q is the probability of an individual answer being wrong. If we have confidence in the medical profession, so that we think q is small, then q^k becomes very small as k increases. When the experts agree, they are likely to be right. Meehl (1999) has argued similarly regarding the opinions of different scientists who have given their opinions on the correctness of a theory.

On the other hand, Meehl (1999) has also stressed the necessity of knowing that those expressing the opinions are indeed expert: "on a disputed point in quantum mechanics I would rely on Dirac's judgment rather than on the pooled judgments of ten psychologists." Similarly, the noted physicist Feynman (1985) recounts his service on a

committee to choose a science text for elementary schools. A book he considered worthless had been “approved by sixty-five engineers at the Such-and such Aircraft Company!” Feynman did not want to claim that he was “smarter than sixty-five other guys – but the average of sixty-five other guys, certainly!”

Both quotations illustrate our basic argument: the opinion of one highly qualified expert can be far more valuable than the opinion of 100 novices. This issue is fundamental in the identification of the qualified expert (Weiss & Shanteau, 2001).

At the turn of the twentieth century, the consensus among physicists was that neo-Newtonian physics offered the best account of the universe. Einstein did not share this belief. Had physics relied on consensus to determine correctness, Einstein’s views would have been rejected and many of the conceptual and technological advances of the last 100 years would not have occurred. In psychology, we continue to view research in political (i.e., popularity) terms. That means that our Einstein, should he or she burst onto the scene, will be ignored because of lack of consensus.

Independence

It is important to note that the binomial argument presumes the judgments to be independent. In practice, independence may be violated in several ways. The most blatant is that judges may reach their decisions in concert. Sharing opinions prior to reaching a decision, as is done in a jury setting, clearly reduces the effective number of independent voices. In some situations, discussion is disallowed. For example, during figure skating competitions, judges are expressly forbidden from interacting.

A more subtle form of nonindependence is that decision makers may follow the same rules. For example, figure skating judges are carefully trained to follow specific

performance guidelines. Special schools for judges stress the importance of applying uniform criteria. Skating judges are taught how much to value a particular maneuver, and how to recognize when it has been carried out properly. Common training thus reduces the independence of evaluations, yet we would consider the scoring chaotic if judges were not looking for the same performance characteristics.

An inherent contradiction in applying the binomial logic is that one kind of independence violation, collusion, is considered inappropriate, yet others are generally deemed desirable. Perhaps the reason is that collusion generally occurs immediately prior to the judgment and is thereby an obvious violation, while training typically takes place long before the judgment. We feel that this is a difference of degree rather than kind.

Collusion generates highly correlated judgments. Routine training is likely to yield judgments that are only moderately correlated, because the training is imperfect and the judges forget some of it. People are willing to ignore these moderate correlations in order to justify the use of consensus.

Expert Judgment

The binomial perspective is the basis for Einhorn's (1974) suggestion that agreement with other experts is a necessary characteristic of an expert. That is, the experts in a given field should agree with each other (Ashton, 1985). Einhorn argued that if opinions disagree, then some of the members of the proposed set of experts must not be functioning at the appropriate level. He used this reasoning to disparage several professions, among them clinical psychology and stock brokerage, by showing that

agreement among practitioners (as measured by inter-individual correlations) was unexpectedly low.

Lack of agreement among peer reviewers for grants and manuscripts has come under a good deal of scrutiny (Cicchetti, 1991). Poor inter-reviewer reliability is the norm across a variety of disciplines. Mixed reviews usually lead to negative decisions. This means disappointment for the submitter. Those whose academic or economic fortunes depend upon the luck of the draw are likely to lose confidence in both their colleagues and the process.

Most respondents to Cicchetti's (1991) target article agreed that reviewer disagreement is undesirable. Rather lonely among the reactions were those of two experienced journal editors (Bailar, 1991; Kiesler, 1991), who argued that reviewers are selected for their complementary perspectives in order to inform the editor's decision. In their view, discrepancy is a healthy sign that reviewers are attending to different aspects of the manuscript, thereby enhancing the validity of the evaluation.

When high consensus does occur, it is because the expert community has largely solved the problems of the domain. Because each individual expert is getting the correct answer, usually with the aid of well-developed technology, their answers agree. It is not agreement that makes the answers correct, rather it is that answers agree when they are correct.

Laypersons are disturbed when experts do not agree because they overestimate the scientific success achieved by the professional community. Noting that experts often disagree, Shanteau (1999) characterized the "Experts Should Converge" argument as a fundamental misunderstanding of the way experts think. Because the real-world

problems that experts tackle seldom have single, stable answers, disagreement is inevitable and even useful. Indeed, one might argue that too much inter-individual agreement is a signal that the problem is trivial, and scarcely worthy of an expert. Expert judgment may have been replaced by a mechanical device (Weiss & Shanteau, 2001). Certainly, there are manuscripts on which the reviewers agree; but those manuscripts tend to be the ones rated as poor (Cicchetti, 1991).

Structural Reasons for Disagreement among Experts

Analysis of the context in which most experts work provides five structural factors why experts may disagree. These factors reflect the situational constraints under which most experts work.

(1) In the domains where experts work, the “ground truth” is often a fiction. Single-point optimal solutions do not exist. Despite the tremendous analytic ability of master players and the incredible computation speed of computer programs such as Deep Blue, for example, the game of chess still does not yield optimal solutions. If this is true for a well-structured game such as chess, how can it be possible to find a “correct answer” in an ill-structured setting? The reason we need experts in the first place is that they offer us answers that we could not obtain any other way (Shanteau, 1999). When there is no single best answer, it should not be surprising that different experts choose different solutions.

(2) A distinction can be made between the different levels of decisions made by experts. Using terminology from medicine, it is possible to distinguish between three levels: The first is diagnosis (what is it?) based on categorization and/or classification. The second is prognosis (what is the likely outcome?) based on forecasting future

scenarios. And the third is treatment (what to do about it?) involving selection of a course of action. There are thousands of diagnoses and hundreds of prognoses, but relatively few treatments. As pointed out by medical researchers (e.g., Schwartz & Griffin, 1986), experts might disagree at one level (diagnosis), but agree at another (treatment).

(3) Despite the assumption made by many researchers, experts are seldom asked to make single-outcome decisions. The concept of a “point prediction” is largely a fiction created for the convenience of the researcher and is not descriptive of the tasks that experts do. As Golde (1969) noted, although “an expert does sometimes make decisions, his (her) role is usually much more of an advisor . . . (they) let me know the kinds of decisions or actions that I must take.” In other words, the job of the expert is to clarify alternatives and describe possible outcomes for clients.

(4) As Klein, Orasanu, Calderwood, & Zsombok (1993) emphasized, experts generally work in dynamic situations with frequent updating. Thus, the problems faced by experts are unpredictable, with evolving constraints. In such situations, there are rarely ideal answers. Therefore, while outsiders assume a stationary target, the reality faced by experts is generally more like a moving target.

(5) A long-term perspective reveals that experts work in realms where the basic science is still evolving. For instance, the rapid changes in medicine mean that the current “best answers” are soon obsolete. Why should we expect experts to agree on a single “correct answer,” say for the treatment of AIDS, when new knowledge will likely provide better solutions tomorrow?

Functional Reasons for Disagreement

Five functional factors underlie disagreement among experts. These factors have to do with how experts think about the decisions and judgments they make.

(1) Most experts operate as if they have flat loss functions for deviations from optimality. They see small deviations as having minor consequences. In comparison, von Winterfeldt and Edwards (1986) have observed that researchers often operate as if experts have steep loss functions. That is, researchers view any deviation from optimality, no matter how slight, as having large consequences. Similarly, they see any disagreement between experts, no matter how small, as reason for concern.

(2) While those in the “heuristics and biases” tradition (e.g., Kahneman, Slovic, & Tversky, 1982) view any deviation between behavior and the “correct answer” as an “error”, experts have a different definition of error. As noted above, experts are usually more concerned about avoiding big mistakes, whereas researchers are looking for perfection. Thus, the same outcome could well be called an “error” by the researcher and a “success” by the expert. In the same situation, experts may see agreement where investigators see disagreement.

(3) In many, perhaps most, settings, experts expect to disagree with each other. In a discussion between any two academics, for instance, we know that they invariably will find something about which to argue. Even when they agree on 99% of the issues, they will quickly find the last 1% and disagree about that. Similarly, experts in almost any field bypass points of agreement to focus instead on disagreements. Thus, experts view disagreements as a normal part of doing their job.

(4) Disagreements are often the route by which experts increase understanding of their field. By seeking out areas of disagreement between one another, experts explore

the limits of their own knowledge and stretch their range of competency. Therefore, experts see disagreements as a key step in increasing their grasp of the field.

(5) Once a domain has advanced to the point where all issues are resolved, there will be few disagreements among experts because there is nothing left to argue about. When a field has developed to that extent, however, the answers are known and agreed upon. Thus, total agreement among experts is an indication that there is no longer much of a role for experts to play in that domain.

Domain Differences

We all know that experts in different domains perform different tasks. Yet it is common to treat all experts alike, so that the term “expert” is used generically. For instance, Kahneman (1991) concluded “there is much evidence that experts are not immune to the cognitive illusions that affect other people.” On the other hand, it is widely known that at least some experts, such as weather forecasters (Murphy & Winkler, 1977), show little sign of biases or “cognitive illusions” in their professional capacities. Thus, despite the generalizations drawn about experts in general, we know there are many exceptions to the rule.

In an effort to account for these domain differences, Shanteau (2000) constructed Table 1 to differentiate between those domains where experts do well and those where experts do not. The table is based on a continuum from high to low competence. In the left column are those domains where experts make aided decisions using Decision Support Systems (DSS) or other computerized tools, e.g., weather forecasters. The next column contains domains where experts make skilled but largely unaided decisions, e.g., livestock judges. The third column lists domains where experts show limited

competence, e.g., clinical psychologists. The behavior of experts in the last column is little better than random, e.g., stockbrokers. Note: Assignment of domains within the table was based on a review of the literature, i.e., the assessment of competence is based on the opinions of researchers who study each domain.

 Insert Table 1 here

There are many ways to describe the differences in this table (see Shanteau, 1992a,b). For present purposes, it makes most sense to note that domains to the left side possess more stable (static) properties. That is, the stimuli and the problem “hold still” for experts to evaluate. The domains to the right side, however, involve more changeable (dynamic) properties. Thus, the stimuli are less stable, harder to specify, and more like “moving targets.” It makes sense, therefore, that expert agreement will be higher on the left side and lower on the right side.

To test this idea, Table 2 summarizes results from studies of domain experts in the four categories of Table 1. Two domains are listed under each category, with the between-expert agreement (consensus) given as average correlations. As can be seen, the average consensus (r) value for weather forecasters is .95, whereas average values for livestock judges, clinical psychologists, and stock forecasters are .50, .40, and .32, respectively. Comparable results appear for other domains on the second line. The values support the trend outlined above – better structured domains lead to high consensus and less structured domains to low consensus.

 Insert Table 2 here

For comparison, the average within-expert correlations (consistency) for these same domains are listed in Table 3. The trends are similar, with better structured domains leading to higher internal consistency. As expected, consistency (\bar{r}) values (except for pathologists) are higher than corresponding consensus values in Table 2. In two domains (livestock judges and polygraphers), there are notable discrepancies between the consensus and consistency correlations. In these domains, there appear to be “schools of thought” that have produced sizable disagreements among various experts (Shanteau, 2000)

 Insert Table 3 here

Latent Trait Analysis

A mathematically sophisticated extension of the agreement perspective has been presented by Uebersax (1988, 1992, 1993) using latent class analysis. Acknowledging that while experts might disagree for example, in diagnoses of individual patients, his proposal is that generally experts will agree in what they are doing. Experts will consider similar aspects of the situation and employ similar processes, although they may vary in terms of biases and inconsistencies. Uebersax proposed that interobserver agreement at the level of the latent structure underlying the judgments confers validity. Uebersax and

Grove (1990) do acknowledge the possibility that the latent trait may not be the trait of interest and may reflect shared, but inappropriate, criteria.

Our view is that high degrees of consensus do not necessarily connote expertise. The history of science is replete with examples of the false consensus. Some well-known examples of premature agreement within the scientific community are flat earth, phlogiston, phrenology, and the Rorschach test. Martin Gardner's (1957) compendium gives other illustrations. Groupthink in the Cuban missile crisis (Janis, 1972) is a classic example of how premature consensus retards effective decision making.

A far more important characteristic of expertise, also proposed by Einhorn (1974; see also Weiss & Shanteau, 2001), is consistency. Imagine your consternation if a physician's reports on your medical condition were variable. Inconsistency is virtually unimaginable, and it would be a brave patient who asked the doctor to do an independent re-evaluation. In fact, when we do observe inconsistency in an acknowledged expert's opinions, we typically ascribe the variation to changes in the situation. We do not allow for the possibility that the judgments reflect sampled observations from a distribution characterized by high variance. In order to be considered expert, one must be consistent.

In principle, consistency is easy to measure; we need merely ask for repeated responses. But those replications must be independent. This requirement is not merely a statistical nicety. An expert presented with the same stimulus situation will strive to appear consistent, and so will repeat the previous response because of self-presentation considerations (Goffman, 1959). If the stimuli are memorable, as would certainly be the case with a medical patient, the response is likely to be identical. In a research setting, it might be possible to disguise the stimuli to render the judgments more independent.

Usually, though, independence is sought by spacing the judgments over time. In the medical setting, delay is not only inconvenient; the passage of time is often accompanied by real changes, and so evaluating consistency is problematic.

Because consistency is hard to ascertain, people rely upon consensus as a surrogate. Presenting the same problem to several experts independently is easy, and therefore seeking consensus is attractive. It is difficult to explore intra-individual agreement, but it is easy to look at inter-individual agreement.

The logical error of relying upon consensus remains unexposed because empirically, consensus is usually associated with consistency. In a review of experts across eight domains, Shanteau, Weiss, Thomas, and Pounds (2000) found that intra-individual agreement and inter-individual agreement yielded almost identical measures. Reported correlations ranged from .98 for consistency and .95 for consensus for weather forecasters and .40 and .32, respectively, for stockbrokers.

It seems unlikely that consensus can be higher than consistency (except for chance fluctuations), i.e., that you can agree with others more than you agree with yourself. The limitation parallels the ceiling placed on the validity of a psychometric test by its reliability. In the fields cited by Shanteau et al. (2000), consensus is about as high as it could be, given the degree of consistency shown by the experts in the domain. The correspondence probably means that the judgments are well prescribed by coherent training across the community or by equipment. However, for two of the expert domains reviewed by Shanteau et al. (2000), livestock judging and polygraphy, consistency was much greater than consensus. Apparently, those experts are using rules that are simple enough that they can be consistent, but the rules reflect different schools of thought.

Cultural Consensus Analysis

The cultural consensus approach has been used by anthropologists (Romney, Weller, & Batchelder, 1986) to determine which informants contribute trustworthy information. The technique is mathematically sophisticated and quite elegant, but at its heart a simple idea: majority answers to questions are likely to be true¹. From this premise, the authors derive both estimates of the relative competencies of the informants and assessments of the likelihood that a particular answer is correct. They provide a striking demonstration that it is possible to reconstruct the answer key to an objective test (Batchelder & Romney, 1988).

Generalization of this approach to domains of expertise is problematic. The approach plays on the idea that response clustering implies expertise, because experts must agree on the right way to carry out the task and non-experts behave idiosyncratically. We consider the premise to be dubious. As we have suggested, for many problems that experts face, there may not be a single right answer. Even if it is granted that experts should be required to agree, it does not follow logically that those who do agree must be experts. Experimentally, the approach suffers from the limitation that it requires the subject population to have sizable numbers of experts as well as non-experts. The difficulty is that experts tend to be rare. Clustering cannot be demonstrated when only one or two experts are in the sample.

Meta-analysis

Another arena in which consensus supplants consistency is in the evaluation of scientific hypotheses. Though not without its critics (e.g., Chow, 1987), the statistical combination of results using meta-analysis (Cooper & Rosenthal, 1980) has in recent

years come to be accepted as the standard method of inquiry. The reviewer exercises judgment in deciding which studies enter the compilation, but assiduously tries not to be influenced by data in doing so. Meta-analysis attempts to bring the virtue of quantification to the difficult task of integrating disparate research results. Although minor procedural differences across studies are the norm, effect sizes are combined and essentially averaged. Such combinations are often misleading, especially in studies where there is experimental control, because effect size depends on the researcher's choice of the levels of the independent variable².

Our deeper objection is on philosophical grounds. Meta-analysts argue that their approach avoids disparaging researchers whose results are outliers, since there is no attempt to invalidate, or even examine, the experiments that yield anomalous results. We believe that such reviews induce the profession to reach premature closure. The outliers are simply overwhelmed by the majority. The following thought-experiment illustrates the potential problem. Let us suppose that researchers are trying to decide which way water flows when it is going down the drain. The researchers in Los Angeles, in Manhattan, in Miami, in Boston, as well as those in Paris and Beijing all report that the water circles in a clockwise manner. These results are cross-cultural and they agree, so the scientific community happily accepts the result. One lone voice from Melbourne offers a contradictory report, that the water flows in a counter-clockwise direction in his laboratory. Although no one says anything uncomplimentary, the Antipodean result scarcely affects the consensually certified answer. The dissenting report is, pardon the expression, washed away.

The point of the thought-experiment is that an unknown, powerful effect may be overlooked when the field decides the question is settled. Indeed, it seems likely that our understanding of behavioral phenomena is often comparable to that of the fictitious hydrologists. Focus on consensus within the scientific community threatens to stifle inquiry. Meta-analysis may be little more than a sophisticated vote-counting scheme (Light & Smith, 1971), one that takes into account sample sizes and effect sizes, but the imprimatur of quantification reifies its results. A large mean effect size is likely to be interpreted as a final statement about the problem.

Instead, we argue that the important determinant of whether a laboratory result should be accepted is its replicability. Significance statements are poor substitutes for repeated results (Fisher, 1925). This is hardly a novel idea, but the current infrastructure of the scientific community does not reward a researcher who seeks to examine the consistency of experimental results.

Our recurrent concerns with independence apply here as well. In order for a replication to strengthen the evidence base for a phenomenon, it must provide independent confirmation of the results. Clearly, an experiment that mimics another will yield results that are constrained by the common design.

Conclusion

Reliance upon consensus is ultimately a democratic notion, and democracy is a valued political ideal. We are not arguing that a consensus-based measure is always inappropriate. To determine the most desirable of a set of alternatives, such as the best political leader or most attractive building decor, it is sensible to use a polling system.

The mean response reflects how a typical member of the sampled population views the stimulus. It is implicit that everyone's opinion is equally valued.

It is understandable that consensus of opinion and consensus of results have come to be means to scientific validation. However, what works in the world of politics need not work in science. Because consensus and consistency often coincide, the scientific community has been fooled into relying upon consensus, when it is reproducibility that is the cornerstone of good science.

References

- Ashton, A. H. (1985). Does consensus imply accuracy in accounting studies of decision making? Accounting Review, 60, 173-185.
- Bailar, J. C. (1991). Reliability, fairness, objectivity and other inappropriate goals in peer review. Behavioral and Brain Sciences, 14, 137-138.
- Batchelder, W. H., & Romney, A. K. (1988). Test theory without an answer key. Psychometrika, 53, 71-92.
- Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. Behavioral and Brain Sciences, 14, 119-186.
- Chow, S. L. (1987). Meta-analysis of pragmatic and theoretical research: A critique. Journal of Psychology, 121, 259-271.
- Cooper, H. M., & Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. Psychological Bulletin, 87, 442-449.
- Einhorn, H. J. (1974). Expert judgment: some necessary conditions and an example. Journal of Applied Psychology, 59, 562-571.
- Feynman, R. (1985). Surely you're joking, Mr. Feynman! New York: Norton.
- Fisher, R. A. (1925). Statistical methods for research workers. London: Oliver & Boyd.
- Goffman, E. (1959). The presentation of self in everyday life. Garden City, NY: Doubleday/Anchor Books.
- Goldberg, L. R., & Werts, C. E. (1966). The reliability of clinicians' judgments: A multitrait-multimethod approach. Journal of Consulting Psychology, 30, 199-206.
- Golde, R. A. (1969). Can you be sure of your experts? NY: Award Books.

- Gardner, M. (1957). Fads and fallacies in the name of science. New York: Dover.
- Janis, I. L. (1972). Victims of groupthink. Boston: Houghton-Mifflin.
- Kahneman, D. (1991). Judgment and decision making: A personal view. Psychological Science, 2, 142-145.
- Kahneman, D., Slovic, P., & Tversky, A. (1982). Judgment under uncertainty: Heuristics and biases. Cambridge: Cambridge University Press.
- Kida, T. (1980). An investigation into auditor's continuity and related qualification judgments. Journal of Accounting Research, 22, 145-152.
- Kiesler, C. A. (1991). Confusion between reviewer reliability and wise editorial and funding decisions. Behavioral and Brain Sciences, 14, 151-152.
- Klein, G. A., Orasanu, J. T., Calderwood, R., & Zsombok, C. E. (1993). Decision making in action: Models and methods. Norwood, NJ: Ablex Publishing Corporation.
- Light, R. J., & Smith, P. V. (1971). Accumulating evidence: Procedures for resolving contradictions among different research studies. Harvard Educational Review, 41, 429-471.
- Lykken, D. T. (1979). The detection of deception. Psychological Bulletin, 80, 47-53.
- Meehl, P. E. (1999). How to weight scientists' probabilities is not a big problem: Comment on Barnes. British Journal for the Philosophy of Science, 50, 283-295.
- Murphy, A. H., & Winkler, R. L. (1977). Can weather forecasters formulate reliable forecasts of precipitation and temperature? National Weather Digest, 2, 2-9.
- O'Grady, K. E. (1982). Measures of explained variance: Cautions and limitations. Psychological Bulletin, 92, 766-777.

- Phelps, R. H., & Shanteau, J. (1978). Livestock judges: How much information can an expert use? Organizational Behavior and Human Performance, 21, 209-219.
- Raskin, D. C., & Podlesny, J. A. (1979). Truth and deception: A reply to Lykken. Psychological Bulletin, 86, 54-59.
- Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: a theory of culture and informant accuracy. American Anthropologist, 88, 313-338.
- Schwartz, S., & Griffin, T. (1986). Medical thinking: The psychology of medical judgment and decision making. New York: Springer-Verlag.
- Shanteau, J. (1992a). Competence in experts: The role of task characteristics. Organizational Behavior and Human Decision Processes, 53, 252-266.
- Shanteau, J. (1992b). How much information does an expert use? Is it relevant? Acta Psychologica, 81, 75-86.
- 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).
- Shanteau, J. (2000). What does it mean when experts disagree? In G. Klein & E. Salas (Eds.), (pp.). Hillsdale, NJ: Erlbaum.
- Shanteau, J., Weiss, D. J., Thomas, R., & Pounds, J. (2000). Performance-based assessment of expertise: How can you tell if someone is an expert? European Journal of Operations Research, in press.
- Shaw, G. B. (1956). The vice of gambling and the virtue of insurance. In J. R. Newman (Ed.), The world of mathematics, Vol. 3. (pp. 1524- 1533). New York: Simon & Schuster.

- 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, 62-71.
- Uebersax, J. S. (1988). Validity inferences from interobserver agreement. Psychological Bulletin, 103, 405-416.
- Uebersax, J. S. (1992). Modeling approaches for the analysis of observer agreement. Investigative Radiology, 27, 738-743.
- 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.
- Von Winterfeldt, D., & Edwards, W. (1986). Decision analysis and behavioral research. Cambridge: Cambridge University Press.
- Wainer, H. (1983). Pyramid power: Searching for an error in test scoring with 830,000 helpers. American Statistician, 37, 87-91.
- Weiss, D. J., & Shanteau, J. (2001). Empirical assessment of expertise. Manuscript submitted for publication.

Authors' Note

David J. Weiss, Department of Psychology, California State University, Los Angeles.

James Shanteau, Department of Psychology, Kansas State University.

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 Alice Isen, Gary McClelland, Julia Pounds and Rickey Thomas for valuable discussions.

Correspondence concerning this article should be directed to David J. Weiss, Department of Psychology, California State University, Los Angeles, 5151 State University Drive. Los Angeles, CA 90032. email: dweiss@calstatela.edu.

Footnotes

1. Wainer (1983) presents an example of the risk in relying upon consensus to define correctness. The Education Testing Service inadvertently designated an incorrect option to a difficult geometry problem on the PSAT as the correct one. Subsequent investigation found that students who scored highest on similar questions were much more likely to have selected the official (but incorrect) answer than the correct one.
2. In general, the farther apart the levels are spaced, the greater the apparent effect size (O'Grady, 1982).

Table 1

Progression of Domains from High to Low Performance

Stability of Domain Stimuli

High Levels of PerformanceLow Levels of Performance

<u>Aided Decisions</u>	<u>Competent</u>	<u>Restricted</u>	<u>Random</u>
Weather Forecasters	Chess Masters	Clinical Psychologists	Polygraphers
Astronomers	Livestock Judges	Parole Officers	Managers
Test pilots	Grain Inspectors	Psychiatrists	Stock Forecasters
Insurance Analysts	Photo Interpreters	Student Admissions	Parole Officers
Physicists	Soil Judges	Intelligence Analysts	Court Judges

Table 2

Consensus Values for Experts in Different Domains

Stability of Domain Stimuli

High Levels of PerformanceLow Levels of Performance

<u>Aided Decisions</u>	<u>Competent</u>	<u>Restricted</u>	<u>Random</u>
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: Values cited in this table were drawn from the following studies (from left to right): 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 3Intra-Individual (Consistency) Values in Different Domains

Stability of Domain Stimuli

High Levels of PerformanceLow Levels of Performance

<u>Aided Decisions</u>	<u>Competent</u>	<u>Restricted</u>	<u>Random</u>
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: Values cited in this table were drawn from the following studies (from left to right): 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).