

In, Green, B., Cressy, R., Delmar, F., Eisenberg, T., Howcroft, B., Lewis, M., Schoenmaker, D., Shanteau, J., & Vivian, R. (eds.). (2000). *Risk behaviour and risk management in business life*. Dordrecht, The Netherlands: Kluwer Academic Press. (pp. 186-196).

Why Do Experts Disagree?

James Shanteau

*Professor/Director, Institute for Social and Behavioral Research
Kansas State University*

Researchers are often perplexed when they observe sizable and consistent disagreements between subjects. This is particularly true when the subjects are experts. As shown by the following quotations, however, disagreements have been viewed historically as necessary.

By different methods different men excel (Churchill, 1764)

The history of scholarship is a record of disagreements (Hughes, 1936)

The tough-minded . . . respect differences (Benedict, 1940)

The position taken here is that disagreements between experts are a normal part of their jobs.

The purpose of this paper is to explore why it is natural for experts to disagree with each other. The material is organized as follows. First, there is a review of the literature on disagreement between experts and a discussion of a commonly held hypothesis about expertise. Second, 10 structural and functional factors behind why experts often disagree are presented. Third, domain differences in the extent to which experts agree are considered. Fourth, an alternate hypothesis is offered that more closely corresponds to how experts view tasks. Finally, the paper concludes with a discussion of research implications and future directions for analyses of experts.

Background

Since the start of systematic research on decision making in the 1950's, researchers have expressed dismay at the extent to which experts disagree. For example, if we ask two financial experts to assess an investment, the expectation of most investigators is that they will make the same recommendation. If they arrive at different recommendations, we wonder whether they are as skilled as they claim.

In a seminal paper, Einhorn (1974) argued that *consensus* (between-expert reliability) is a necessary condition for expertise. He reported, however, sizable differences in the diagnoses by three expert medical pathologists. The average between-expert correlation (r) was .55 (where .0 is chance and 1.0 is perfect).

Similar evidence was presented in a study of expert livestock judges evaluating overall breeding quality of swine. Despite a high level of internal consistency (average $r = .96$), the consensus was much lower, $r = .50$. Apparently, livestock experts have internally consistent strategies, but do not agree with each other about what those strategies should be. Comparable results have been reported by researchers for other domains, eg, consensus values of less than .40 were found for judgments by professional stockbrokers and clinical psychologists.

Several studies of professional auditors have reported greater between-expert consensus. Three studies of audit judgment revealed correlations of .76, .70 and .68 for assessments of financial profiles, internal controls and control reliance, respectively. Several studies have explored the question of whether agreement increases with experience. One of my students compared 11 accounting students with 10 audit seniors and 10 audit partners; the mean between-judge correlations increased from .66 to .76 to .83, respectively. Other studies have generally reported similar trends. (For more information, see Shanteau, 1992).

These results suggest three conclusions: First, experts in a variety of domains frequently disagree; the consensus correlations range from .40 to .55. Second, the agreement between financial experts is higher, with correlations ranging from .68 to .83. Third, there is some evidence that increased experience may lead to greater consensus. Before making too much of these findings, however, it is important to note the wide range of tasks and skill levels studied.

Experts-Should-Converge Hypothesis

The unimpressive consensus correlations for non-financial experts and somewhat better results for auditors lead many researchers to question the abilities of experts in general. Following Einhorn's logic, these investigators assume that agreement is a necessary condition for expertise. The lack of agreement, therefore, suggests that "experts are no damn good." This analysis and interpretation of reliability data are based on an implicit hypothesis about experts. The hypothesis, labeled the *Experts-Should-Converge* (ESC), is based on the following five arguments:

- (1) For most tasks performed by an expert, there is assumed to be a 'gold standard' or unique 'ground truth' that can be identified. If this truth was easy to access, we could get it directly. For expert tasks, however, it is outside the realm of common knowledge or direct sensory experience of most people. Thus, unique correct answers exist, at least in theory, but they are difficult to obtain.
- (2) Because of their special skills and experience, experts should be able to figure out and tell the rest of us about this ground truth. That is, experts can access what others cannot.
- (3) Since by definition there can be only one ground truth, all experts should give us the same answer. The special abilities of experts should allow them to obtain a single truth.
- (4) If experts disagree, then someone is wrong – they cannot all be correct. Some or all of them must not be experts, ie, disagreements are a reflection of ignorance or incompetence.
- (5) Since non-experts do not know which of the so-called 'experts' are correct, the only safe course of action is distrust all of them. That is, disagreement between experts implies that we should be suspicious of their claimed special abilities.

This hypothesis, of course, is not a formal chain of logic. But, it is implicit in the way that most researchers reason about findings reporting disagreements between domain experts.

Contributing Factors

The ESC hypothesis is supported by two disconnected lines of thought, one of which affects theoretical reasoning and the other influences empirical research. First, research on decision making has been linked historically to theories from economics (e.g., expected utility theory) and statistics (e.g., Bayes theorem). When applied to well-specified problems, such theories lead to unique solutions. Given a specified set of antecedent conditions, these formalisms provide point estimates of the optimal decision strategy. Interestingly, a great deal of laboratory work has been devoted to showing the inadequacy of such theories. Nevertheless, economics and statistics continue to provide a point of reference for much of decision research.

This reasoning has also been applied to analyses of domain experts. Although most experts work on problems that are not well specified, researchers nonetheless assume that unique solutions exist – just as they do for simpler cases. Many investigators have recently questioned the relevance of economic/statistical theories as a basis for making real-world decisions. While economic and statistical theories often cannot be applied to the domains where experts work, it is nonetheless assumed that point solutions exist.

Second, nearly a century of experimental psychology (particularly in America) has been focused on analysis of the *Generalized Normal Human Adult Mind* - GNHAM. This emphasis dates to the beginnings of experimental psychology, particularly Wilhem Wundt and Edward Titchener: “Titchener’s interest lay in the generalized, normal, human adult mind that had also been Wundt’s main concern.” (Heidbreder, 1933).

In fact, there is no evidence that Wundt ever used this term. References to the concept, however, are common in Titchener’s writing, e.g., “psychology is concerned with the normal, human, adult mind.” Modern historians of psychology now believe that Boring created a “myth of origin” to provide a justification for Titchener’s (who was his mentor) place in psychology.

According to GNHAM, the goal of behavioural researchers should be to investigate commonalities among humans, not differences between them. That is, the focus of psychology belongs on the generalized mind, not on individual minds. Because the Titchener-Boring view dominated (at least in America), research was directed to the search for ‘universal truths’ of behaviour. As a result, psychology has not developed paradigms to deal with outlier behaviour. And expertise, by definition, is outlier behaviour. (See Shanteau, 1988 for background.)

Research, therefore, has been influenced by two research streams that assumed (1) that decision problems should have unique correct answers and (2) that differences between individuals are not importance to empirical investigations. The persistence of observed differences between experts provided evidence inconsistent with these views. Thus, disagreements between experts lead to the conclusion that something must be wrong. In the next two sections, I explore 10 factors behind why researchers should not be surprised when experts disagree.

Structural Factors

Analysis of the context in which most experts work provides five structural factors behind why experts may disagree.

(1) In most contexts where experts work, the ‘ground truth’ is a fiction. Single-point optimal solutions do not exist. Despite the tremendous analytic ability of master players and the incredi-

ble computation speed of computer programs such as *Deep Blue*, the game of chess still does yield optimal solutions. If this is true for a well-structure game such as chess, how can it be possible to find a ‘correct answer’ in an ill-structured setting? In fact, the reason we need experts in the first place is because they offer us answers that we could not obtain any other way.

(2) A distinction should 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. 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. It should not be surprising to find that experts may disagree at one level (diagnosis), but agree at another (treatment).

(3) Despite the assumption behind the ESC hypothesis, 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 (1970) noted, although “an expert does sometimes make decisions, his (her) role is usually much more of an advisor.” In other words, the job of the expert is to clarify alternatives and describe possible outcomes.

(4) Experts generally work in dynamic situations where there is frequent updating, ie, the problems faced by experts are unpredictable, with evolving constraints. In such situations there are rarely any best or correct answers. Therefore, while the ESC model assumes a stationary target, the reality faced by experts is generally more like a moving target.

(5) A long-term perspective reveals that experts frequently 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 Factors

An analysis of the strategies used by experts to make decisions reveals at least five functional or process factors behind why experts may disagree. These factors have to do with how experts think about the choices and judgments that make.

(1) While most researchers view an ‘error’ as any deviation between behaviour and the ‘correct answer’, experts have a different definition of error. 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.

(2) Most experts operate as if they have flat loss functions for deviations from optimality. They see small deviations as having minor consequences. But researchers often operate as if they have steep loss functions. That is, they view any deviation from optimality, no matter how slight, as having large consequences. A similar argument can be made for how disagreements between experts are viewed.

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

(4) Disagreements are often a route by which experts increase understanding of their field. By seeking out areas of disagreement, experts examine the limits of their own knowledge and stretch their range of competency. Therefore, experts see disagreements as a key step for maintaining and increasing their grasp of their field.

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

Domain Differences

We all know that experts in different domains perform different tasks. Yet researchers persist in treating all experts alike, so that the term ‘expert’ is used generically. Nearly all researchers are aware that at least some experts, e.g., weather forecasters (see Stewart, et al, 1997), show little sign of bias or ‘cognitive illusions.’ Thus, despite the generalizations often drawn about experts, we know there are (many) exceptions to the rule.

In an effort to account for these domain differences, I constructed Table 3.7.1 (see Shanteau, 1992) 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 expert make aided judgments using Decision Support Systems (DSS) or other computerized tools, e.g., in weather forecasting. The next column contains domains where experts make skilled but unaided decisions, e.g., livestock judges. The third column lists domains where experts show limited competence, e.g., clinical psychologists. The behaviour of ‘experts’ in the last column is close to random, e.g., stockbrokers.

Table 3.7.1

Progression of Domains from High to Low Performance

Highest Levels of Performance.....Lowest Levels of Performance			
<i><u>Aided Decisions</u></i>	<i><u>Competent</u></i>	<i><u>Restricted</u></i>	<i><u>Quasi-Random</u></i>
Weather Forecasters	Chess Masters	Clinical Psychologists	Astrologers

Astronomers	Livestock Judges	Parole Officers	Polygraphers
Test Pilots	Grain Inspectors	Psychiatrists	Stock Forecasters
Insurance Analysts	Photo Interpreters	Student Admissions	Parole Officers
Physicists	Soil Judges	Intelligence Analysts	Court Judges

There are many ways to describe the differences in this table. 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 and problem are less stable, harder to specify, and more like ‘moving targets.’ It makes sense, therefore, that expert agreement will be higher for the left side and lower for the right side.

Table 3.7.2

Reliability (Consensus) Values for Experts

Highest Levels of Performance.....Lowest Levels of Performance

<u>Aided Decisions</u>	<u>Competent</u>	<u>Restricted</u>	<u>Quasi-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	Astrologers r => .0

To test this idea, Table 3.7.2 lists reliability values obtained from studies of domain experts fitting the four categories in Table 3.7.1. Two domains are listed under each category, with the between-expert agreement (consensus) as correlations. As can be seen, the average consensus *r* value for weather forecasters is .95, whereas the average values for livestock judges, clinical psychologists, and stock forecasters are .50, .40, and .32, respectively. Comparable results appear for other domains in the second line. The trend supports the prediction outlined above – better structured domains lead to higher consensus and less structured domains to lower consensus.

For comparison, the within-expert reliability (consistency) correlations for these domains are listed in Table 3.7.3. Again, the trends are similar. With one exception, the consistency values are higher than corresponding consensus values in Table 3.7.2.

Table 3.7.3

Reliability (Internal Consistency) Values

Highest Levels of Performance.....Lowest Levels of Performance

<u>Aided Decisions</u>	<u>Competent</u>	<u>Restricted</u>	<u>Quasi-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
 $r = .50$

Astrologers
 $r = ?$

Researchers' View of Experts

Compared to other fields of inquiry, decision researchers have taken an idiosyncratic view of the abilities of experts. Investigators in artificial intelligence, expert system design, cognitive science, and computer science have all concluded that experts are superior problem solvers worthy of emulation. This is why knowledge engineers build computer simulations around what experts know and do. Further, most domain-specific researchers (such as in medicine, weather forecasting, and financial assessment) view experts as possessing unique information essential for making good decisions. In short, investigators in these fields see human expertise as something to be emulated.

In contrast, researchers, especially in America, have concluded that experts are flawed and prone to making simple errors. Moreover, experts and novices are viewed as sharing the same shortcomings. For instance, Tversky (quoted in Gardner, 1985) stated, "whenever there is a simple error that most laymen fall for, there is always a slightly more sophisticated version of the same problem that experts fall for."

Investigators have overlooked the fact that in most real-world problems, unique solutions do not exist. Instead, there are multiple solution paths. For instance, in medicine there can be many ways to treat an illness – when one approach does not work, a physician seeks out another. It should not be surprising, therefore, to find experts disagreeing about which is the recommended approach to take. Other inquiry systems, therefore, accept multiple points of view across experts as inevitable. In contrast, researchers with their simplified, single-answer view of the world seem to find a multiple-solution perspective difficult to understand.

Multiple-Solution Hypothesis

The position taken here is that previous researchers have unknowingly adopted the ESC view of expertise. According to ESC, disagreement between experts is a sign that something is wrong. This in turn leads to the conclusion that experts are not as skilled or as competent as they claim to be. In this section, I will propose an alternate hypothesis that is closer to the view of how experts see themselves. This is labeled the *Multiple Solution Model* – MSM. There are five arguments behind this hypothesis:

(1) To begin, the primary job of an expert is not to make decisions but to help clients reach a broadly defined goal state. For example, the goal of the client may be to design a better investment portfolio or to find a better loan strategy. These goals do not involve single answers, but instead require something more elaborate from the expert, such as a strategic plan.

(2) To reach the client's goal state requires dealing with multiple, constantly changing factors. The situation faced by experts is different and more complex than the simplified settings studied in research laboratories. Thus, experts work on problems that are considerable more complex than those studied in lab settings.

(3) Using their knowledge and experience, the role of the expert is to recognize patterns and find consistencies in a dynamic problem space. The expert's job is to clarify the issues for the

client. In other words, the challenge for an expert is 'to make sense out of chaos.'

(4) Based on their experience and insights into the nature of the problem, experts help clients clarify their thinking. For instance, an expert will often identify several alternate paths to the client's desired goal states. The expert's role is to lay out the options and consequences in a clear and comprehensible fashion for the client.

(5) In the end, it is the client, not the expert, who actually makes most choices. The expert offers insights and observations, but the client makes and implements the final choice(s). Thus, the responsibility generally rests on the client, not the expert.

The issue is nicely summarized by the management consultant Golde (1970): "We seem to expect too much and the wrong things of our experts." As expressed by MSM, experts generally act more like knowledgeable consultants. Rarely do they function as the 'all-knowing' single-answer decision makers envisioned by most researchers. Instead, experts help clients by giving them the insights and information needed to make their own judgments.

Conclusions

By relying on comparisons to economics and statistical analysis and by incorporating GNAHM assumptions into their research designs, investigators have unknowingly adopted a restricted view of what experts do. By drawing a parallel to economic/statistical theory, researchers have adopted a 'single correct answer' approach to assessing expertise. When an expert (or anyone else) gives an answer different from the 'correct answer', he/she is said to have a "bias" (Tversky & Kahneman, 1974). And when two or more experts give different answers, their claim to special competence is questioned (Einhorn, 1974).

The position here is that researchers have relied on an incorrect view of how experts function. For instance, the environment in which experts work is much different from that incorporated into traditional laboratory research. The complex, changeable environment that experts operate in is considerably more complicated than the small, stable environment constructed by investigators. In reality, problems rarely are simple enough to lead to a single correct answer. Instead, there are multiple answers (or at least multiple routes to answers). If so, it should not be surprising to find that experts often take different approaches (as envisioned by MSM) to making recommendations.

The underlying problem is that researchers misunderstand what experts do and what is expected of them. Investigators seem to think that experts should see the world as they do. However, experts generally have another view of the world – with many complexities and contingencies, but few optimal solutions. From the MSM perspective, disagreements between experts are to be expected. Although researchers view disagreements as evidence of incompetence, experts see disagreements as a more-or-less inevitable part of the job.

By adopting the GNAHM approach to research, most investigators have largely ignored individual differences. Average results are emphasized, variances are not. More important, researchers almost never look at the distribution of responses over subjects. Consequently, exceptions are overlooked and statements are made about results that imply that all people follow the same pattern.

The dangers of reliance on GNAHM thinking for researchers have been recognized for some

time. However, the warnings have not been heard. What should be done? Edwards (1983) offered two messages: "One is that psychologists have failed to heed the urging of Egon Brunswik that generalizations from laboratory tasks should consider the degree to which the task . . . resembles or represents the context to which the generalizations is made." The other message is that "experts can in fact do a remarkably good job of assessing and working with probabilities."

Therefore, disagreement between experts should not be viewed as a source of concern about the competence of experts. It is time for researchers to rethink their views about the lack of consensus leading to a supposed incompetence of experts. Persistence in such beliefs at this time says more about the biases of investigators than it does about experts.

References

- Edwards, W. (1983). Human cognitive capacities, representativeness, and ground rules for research. In P. Humphreys, O. Svenson, and A. Vari, (Eds.), *Analyzing and aiding decision processes*. Budapest: Akademiai Kiado.
- Einhorn, J. (1974). Expert judgment: Some necessary conditions and an example'. *Journal of Applied Psychology*, 59, 562-571.
- Gardner, H. (1985). *The mind's new science: The history of the cognitive revolution*. NY: Basic.
- Golde, R. A. (1969). *Can you be sure of your experts?* NY: Award Books.
- Heidbreder, E. (1933). *Seven psychologies*. NY: Appleton-Century Co.
- Shanteau, J. (1992). Competence in experts: The role of task characteristics. *Organizational Behavior and Human Decision Processes*, 53, 252-266.
- Shanteau, J. (1998). Decision making by experts: The GNAHM effect. In J. Shanteau, B. Mellers, & D. Schum (Eds.), *Decision science and technology: Reflections on the contributions of Ward Edwards*, Norwell, MA: Kluwer Academic Publishers.
- 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.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.

Acknowledgement

The preparation of this chapter was supported, in part, by *National Science Foundation* grant DMII 96-12126 from the Program on Management of Technological Innovation and, in part, by support from the *Institute for Social and Behavioral Research*, Kansas State University.