Journal of Applied Research in Memory and Cognition 4 (2015) 169–175

Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

journal homepage: www.elsevier.com/locate/jarmac

Original Article

# Why task domains (still) matter for understanding expertise

James Shanteau *

*Department of Psychology, Kansas State University, Bluemont Hall 492, Manhattan, KS 66506, USA*

ARTICLE INFO

*Article history:*
Received 13 July 2015
Accepted 13 July 2015
Available online 20 July 2015

ABSTRACT

Researchers often find it unsettling that domain experts sometimes agree and sometimes do not. This paper argues that previous investigators may have lacked sufficient appreciation of domain differences. That is, task characteristics have a major impact on the behavior of experts.

The purposes of this paper are: (1) to review empirical evidence on agreement/disagreement by experts in various domains, (2) to outline a commonly-accepted conceptualization that questions the competence of experts when there are disagreements, (3) to relate domain differences to the degree of agreement between experts, (4) to suggest an alternate conceptualization of expertise that views such differences in agreement/disagreement as inevitable, (5) to look at implications of this conceptualization for future directions of research on expertise.



## 1. Introduction

Several of the scholars in this special issue have devoted much, if not all, of the careers to understanding the part that intuition plays in judgment and decision making (JDM) generally and expertise specifically. For instance, Klein conceptualizes intuition as *"the way we translate our experience into action"* (Klein, 2003, p. iv; see also Klein, 2015, and Hammond, 2015, in this collection of papers). Almost by definition, experts are assumed to be superior decision makers, in large part, because of their experience. (It is worth noting that expert and experience have the same Latin root.)

To the extent that experts in a particular domain have a common background and shared experiences, e.g., in medicine or weather forecasting, it seems likely that they would have developed similar intuitions and thus similar decisions. This suggests that empirical studies of experts in a domain should reveal high levels of agreement between experts.

The goal here is to examine this conclusion. My contention is that agreement between experts is, in fact, expected in some domains. However, I will also show that in other domains, disagreement between experts is not only common, but also predictable from characteristics of those domains. The position taken here is that disagreement between experts should be expected depending on the domains of expertise.

The purpose of this paper is to explore the connection between the extent to which experts agree or disagree and their area of domain specialty. The paper is organized into five sections. First, there is a review of the literature on agreement/disagreement between experts. Second, a commonly held view – that experts must agree or be considered incompetent – is outlined. Third, the role that domain differences play in agreement/disagreement between expertise is considered. Fourth, an alternate perspective is offered that posits that experts in many domains are likely to disagree. Finally, the paper concludes with implications for future research directions.[1]

### 1.1. Background

Since the start of systematic research on decision-making expertise in the 1940's, investigators have often expressed dismay at the extent to which experts disagree. Moreover, such disagreements have led to doubts about the claimed competence of experts. If two experts are asked to assess the viability of a company, the expectation of most researchers is that they should make the same decision. If they make different decisions, then we wonder whether they are as skilled as they claim.

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

* Tel.: +1 785 556 4891; fax: +1 785 532 5401.
*E-mail address:* shanteau@ksu.edu

[1] This paper is an extension of an earlier article on task characteristics (Shanteau, 1992) that has become my most widely cited publication according to *ResearchGate* (30 Dec 2013 7:59), an online academic resource site.

http://dx.doi.org/10.1016/j.jarmac.2015.07.003

examined consistency or within-expert reliability over time, i.e., the extent to which one expert says the same thing in similar situations on different occasions. For pathologists, Einhorn reported a within-expert consistency *r* of .50.

In comparison, weather forecasters have been reported to have high consensus values, $r = .95$, for short-term predictions. Moreover, their internal consistency is near perfect, $r = .98$ (Stewart, Roebber, & Bosart, 1997). Thus, in two domains of expertise that have been frequently studied over the years – medical diagnosis and weather forecasting – there is conflicting evidence in the literature as to whether experts do, or do not, agree both with other experts and with themselves over time.

Initial research in my laboratory focused on agricultural judgments. In a study of four professional livestock judges, for instance, experts were asked to evaluate overall breeding quality of swine (Phelps, 1977). Despite a high level of internal consistency (average $r = .96$), the consensus agreement was much lower, $r = .50$. Apparently, livestock experts have internally consistent strategies, but they do not agree with each other about what those strategies should be.

Similar analyses have been conducted for other types of agricultural judgments. For instance, grain inspectors were found to have a moderate consensus value between judges of $r = .60$, with comparable internal consistency average of $r = .62$ (Trumbo, Adams, Milner, & Schipper, 1962).

In non-agricultural domains, the values are often lower. For example, Hoffman, Slovic, and Rorer (1968) and Goldberg and Werts (1966) reported consensus values of less than .40 for judgments by professional stockbrokers and clinical psychologists. The internal consistency values were similar with correlations of just over .40.[2]

Several studies have explored whether between-expert consensus increases with experience. Ettenson, Shanteau, and Krogstad (1987) found that between-auditor correlations increased from .66 to .76 to .83, for students, audit seniors (mid-level), and full partners, respectively. Messier (1983) reported comparable results – audit partners with more than 15 years experience had greater consensus than partners with less experience (also, see Hamilton & Wright, 1982).

These results suggest three conclusions: First, experts in various domains sometimes agree, but often disagree; the consensus correlations range from .40 to .60. Second, some experts in domains such as weather forecasters show higher levels of agreement, with *r* values up to .95. Finally, for nearly all domains, the internal consistency values are higher than the between-expert consensus values.

Based on this background, a contrast can be made between two perspectives on disagreements between experts: At one extreme is the view that experts should agree and any disagreement suggests something wrong about the qualifications of one or more experts (or all). At the other extreme is the view that disagreements between experts should not be surprising; instead, they are a reflection of the normal state of affairs in many domains. The paper will now turn to a more detailed examination of each perspective.

## 2. Experts-should-agree perspective

The less-than-impressive consensus correlations in most studies of expertise led many researchers to question the abilities of experts in general. Following Einhorn's logic, these investigators assumed that agreement is a necessary condition for expertise. The lack of agreement, therefore, suggests *"experts are no damn good"* (Gettys, personal communication, 1980). This interpretation of reliability data apparently derived from an implicit five-part argument about experts:

(1) For tasks performed by experts, there is assumed to be a single *"gold standard"* or unique *"ground truth."* When this truth is readily accessible, anyone can obtain it directly, e.g., from books or computer programs. For expert tasks, however, the truth is outside the realm of common knowledge. That is why we need experts.
(2) Because of their special skills and experience, experts are uniquely qualified to tell us about this *"ground truth."* That is, experts can determine what others cannot determine.
(3) Since by definition there can be only one *"ground truth,"* all experts should give us a single correct answer. The special abilities of experts thus allow them to make the same decision.
(4) If experts disagree, then someone is wrong—they cannot all be correct. Some (or all) of them must not real experts. Thus, disagreement is a reflection of incompetence.
(5) Since lay people do not know which so-called *"experts"* are correct when there is disagreement, the only safe course of action is to distrust all (or most) of them. Thus, lack of consensus between experts implies that we should be suspicious of their claimed special abilities.

This argument, of course, is not a formal chain of logic. However, it is implicit in the way that many researchers have interpreted evidence of disagreements between experts.

## 3. Domain differences

It is common knowledge that experts in different domains perform different tasks. Yet, decision researchers persist in treating all experts alike, so that the term *"expert"* is used generically. For instance, Kahneman (1991, p. 1165) concluded, *"there is much evidence that experts in general are not immune to the cognitive illusions that affect other people."* This may be true in some domains, e.g., medical diagnosis (Jacavone & Dostal, 1992). But in other domains, such as weather forecasters, e.g., Murphy & Winkler (1977), experts show little sign of biases or *"cognitive illusions."* Thus, despite the generalizations drawn about experts, there are well-known exceptions to the rule.

In an earlier effort to account for these domain differences, I constructed a table to differentiate between those domains where experts do well and those where experts do not (Shanteau, 1992). The left and right columns in the earlier table represented domains where good or superior performances have been reported in the literature. It has since become clear that a dichotomy is overly simplistic and that a more refined view is needed. Therefore, an updated version of the table is presented in Table 1, with four columns.

The table is based on a continuum from high to low competence. In the left column are those domains e.g., weather forecasting, where experts make aided decisions using Decision Support Systems (DSS) or other computerized tools. The next column contains domains, e.g., livestock judges, where experts make skilled, but largely unaided decisions. The third column lists domains where experts, e.g., clinical psychologists, show limited competence. In the last column, the behavior of experts, e.g., stockbrokers, is close to random.

There are many ways to describe the differences in this table. For present purposes, it is sufficient to observe that domains to the left (more competent) side involve stable (static) properties. That is, the stimuli and the problem *"hold still"* for experts to evaluate. The

[2] Most of the studies reporting consensus and consistency values were conducted between 1970 and 2000. Relatively few studies of experts in the new millennium have reported such values. Hence, the literature cited in this paper is, of necessity, somewhat dated. Nonetheless, the values reported for specific domains have remained more-or-less constant over the years, e.g., see the studies of auditing expertise (Ashton, 1974; Messier, 1983; Shanteau, 1993).