A Test of the Theory of DSS Design for User Calibration: The Effects of Expressiveness and Visibility on User Calibration

Brian M. Ashford
U.S. Army Logistics Management College, ashfordb@Lee.Army.mil

George M. Kasper
Virginia Commonwealth University, gmkasper@vcu.edu

A Test of the Theory of DSS Design for User Calibration: The Effects of Expressiveness and Visibility on User Calibration

Brian M. Ashford
U.S. Army Logistics Management College
ashfordb@Lee.Army.mil

George M. Kasper
Virginia Commonwealth University
gmkasper@VCU.edu

ABSTRACT

This paper reports a test of the theory of decision support systems design for user calibration that compares the efficiency of the visual computing paradigm with that of the conventional text paradigm over varied levels of problem novelty. Perfect user calibration exists when a user's confidence in a decision equals the quality of the decision. The laboratory study reported here compared the effects on user calibration of problems depicted either using a text paradigm or visual computing paradigm. The results support the theory. When problems are new and novel, visual depiction improves user calibration. As problems became more familiar and problem novelty decreases, no difference was found in user calibration between subjects exposed to visibility diagrams and those exposed to a traditional text paradigm.

INTRODUCTION

One's belief in the quality of a decision influences the decision selection process (Russo & Schoemaker 1992). Failure to appreciate the quality of a decision can mean that good decisions are not implemented or poor decisions are not properly hedged. Although confidence, as discussed herein, is a subjective prediction, in many situations its accuracy can be objectively assessed. The best-known measure of the accuracy of one's confidence in a decision is calibration, the correspondence between one's prediction of the quality of a decision and the actual quality of the decision (Lichtenstein et al. 1982, Clemen & Murphy 1990, Keren 1991). When this correspondence is equal, and one's decision confidence equals the quality of the decision, calibration is said to be perfect. Perfect calibration is indispensable when selecting a decision from among competing alternatives (Russo & Schoemaker 1992).

The theory of decision support systems (DSS) design for user calibration prescribes requisite DSS design properties needed for users to realize the performance goal of perfect calibration (Kasper 1996). Reviewed below, the theory asserts that a DSS can engender perfect calibration to the extent that it contains requisite properties of **Expressiveness** (expression of words, phases, and audio ranging from, e.g., cryptic to anthropomorphic), **Visibility** (visual icons, images, and animation ranging from, e.g., realistic to abstract), and **Inquirability** (investigative tools and styles ranging from, e.g., data-oriented servile to dialectic contrarian inquiry). The theory further asserts that as problem novelty increases the effective mix of three properties varies from expressiveness to visibility to inquirability.

This paper reports a partial test of the theory of DSS design for user calibration. The effects on user calibration of expressiveness in the form of text and visibility in the form of diagrams were investigated at two levels of problem novelty. Specifically, a laboratory study was conducted in which subjects were exposed to logically identical sets of problems displayed using either expressiveness text or visibility diagrams, and user calibration was computed and compared for higher and lower levels of problem novelty. The results show that the effects of the instantiations of expressiveness and visibility on user calibration varied as prescribed by the theory: visibility resulted in significantly better user calibration when problem novelty was higher, but there was no difference in user calibration between visibility and expressiveness when problem novelty was lower. In other words, visual computing had its greatest impact on user calibration when problems were new and novel.

BACKGROUND

Differentiating confidence, trust, predictability, and decision accuracy, Muir (1994, p. 1915, parenthetics added) states,

> *Predictability* is a basis for trust (and confidence), which in turn, is the basis for an operator (user/decision maker) to make a *prediction* about the future behaviour of a referent. The *accuracy* of that prediction may be assessed by comparing it with the actual behavioural outcome. In addition, an individual who makes a prediction may associate a particular level of *confidence* with the prediction. Thus, *confidence* is a qualifier which is associated with a particular prediction; it is not synonymous with trust.

Realism in confidence is essential for good decisions; the ruinous consequences of unrealistic confidence litter the business decision-making landscape (Russo & Schoemaker 1992). Because action precedes outcome, confidence plays an essential role in both selecting and implementing a decision (Russo & Schoemaker 1992). The confidence ascribed to a predicted outcome when compared to the accuracy of that prediction measures the decision maker's ability to calibrate his or her ascribed confidence

Since its beginning, the primary goal of DSS has been to improve decision quality (Keen & Scott Morton 1978). Unfortunately, evidence suggests that existing DSS can produce "illusory benefits" (Aldag & Powers 1986, Davis et al. 1991), resulting in miscalibration, thereby distorting the decision selection process. Thirty years ago, Chervany and Dickson (1974, p. 1342, parenthetics added) recognized this when they wrote, "Even though the . . .(decision aided) subjects (in their study) did better, their increased average time and reduced average confidence lead to the tentative conclusion that they did not have a 'handle' on the problem." By now, almost everyone can recount from personal experience a situation where computer-generated output produced an aura of exactness and reliance bordering on blind acceptance, even in the presence of compelling evidence to the contrary. In these cases, user calibration may be distorted by the design of the DSS.

Based on and paralleling human problem solving, memory representation, and multiple intelligence theories (Kaufmann 1985; Helstrup 1987; Gardner 1993), Kasper (1996) proposed the notion and detailed a theory of DSS design for user calibration. His design theory prescribes requisite properties of a DSS so the user/decision maker can achieve the goal of perfect calibration. The theory asserts that a user/decision maker can achieve the goal of perfect calibration to the extent that the DSS possesses requisite properties of expressiveness, visibility, and inquirability, and that the effective mix of these properties varies with problem novelty.

The theory of DSS design for user calibration is a design theory (Walls, et al. 1992). It posits a goal, perfect calibration; properties, expressiveness, visibility, and inquirability; and the interaction of these properties to achieve the goal, a mix of expressiveness, visibility, and inquirability that varies systematically with problem novelty.

Expressiveness recognizes that the tone and delivery of words and phrases (written and audio) used in a human-computer interface dialogue (ranging from cryptic to anthropomorphic, from monotone and monotonous to melodic and overly melodramatic) can affect people's beliefs, perceptions, opinions, and predictions. **Visibility** encompasses the icons, symbols, and animation that promote discovery, comprehension, problem solving and engender feelings (Card, MacKinlay, & Shneiderman 1999, Gonzalez & Kasper 1997). **Inquirability** captures the affects produced by actions and interactions with the inquiring system, including scope and nature of dialectics (Churchman 1971) and the restrictiveness and decisional guidance of the system (Silver 1990).

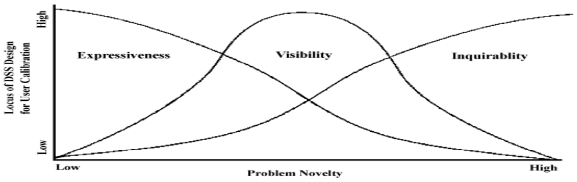

Figure 1 Locus of DSS Design for User Calibration in Relation to Problem Novelty (Kasper 1996)

Depicted in Figure 1, the theory of DSS design for user calibration posits that when problems are somewhat novel and unfamiliar, **Visibility** is the primary contributor to perfect calibration and **Expressiveness** and **Inquirability** play important but lesser, supporting roles. As problems become more familiar and problem novelty decreases, the theory posits that the contribution of **Expressiveness** increases, equals, and eventually exceeds **Visibility** as the primary contributor to user calibration. Stated in the null form, it is hypothesized that:

H_0: There is no difference in user calibration between subjects exposed to Expressiveness and those exposed to Visibility at higher and lower levels of problem novelty.

Larkin and Simon (1987) posited a beneficial role for visibility in search, recognition, and inference processing, and, in response, Bauer and Johnson-Laird (1993) studied the effects of diagrams on inference and found that the use of diagrams improved decision quality. Commenting on their findings, Bauer and Johnson-Laird suggested that in unfamiliar, novel situations, diagrams have a beneficial effect on decision making. Recently, Speier & Morris (2003) found that visual interface users performed better when task complexity was high and their subjective mental workload was less compared to users of a text-based interface. Extending these findings, the study reported here considers the effect of visibility on user calibration and whether this effect, if observed, varies with problem novelty.

EXPERIMENTAL DESIGN, RESEARCH METHOD AND MEASURES

To investigate the hypothesis, a laboratory experiment was conducted. The main effect studied was properties of DSS dialogue design and the dependent variable was user calibration. Specifically, the differential effect of expressiveness and visibility on user calibration was investigated. The experimental design included two different problems to increase the generalizability of the findings and to build upon earlier related research, in particular, that of Bauer and Johnson-Laird. Two calculations of problem novelty, Higher and Lower, were defined by dividing each subject's responses into earlier and later decisions, again based on the work of Bauer and Johnson-Laird. The treatments, measures, formula used to calculate user calibration, and procedures used in the experiment are discussed in detail below.

Treatments

The treatment combinations used in this study were borrowed directly from those developed by Bauer and Johnson-Laird to study deductive reasoning and inference. They developed two logically identical problems presented either as text, a form of expressiveness, or diagrams, a form of visibility. In the interest of space, the reader is directed to Bauer and Johnson-Laird (1993) for detailed descriptions of these treatment conditions. To investigating the hypothesis posited here, subjects also recorded their decision confidence in their selection.

Measuring User Calibration

To measure user calibration requires selecting a method and means for recording both decision quality and the subject's belief in the quality of each decision, a scoring rule and procedure that discourage gaming so that subjects are encouraged to honestly report their beliefs, and a formula for calculating calibration. Each of these requirements is discussed in the next sections.

Recording Beliefs And Decisions

Following convention in calibration research, subjects in this study answered a series of multiple-choice questions by reporting both their decision and confidence in the correctness of each decision. Each subject answered a total of ten multiple-choice questions. The ten questions consisted of the four questions used in the Bauer and Johnson-Laird (1993) study plus six additional questions generated using the same truth table. For each of these ten questions, the subject selected one alternative as his or her choice as the correct alternative and then assigned a confidence value to that alternative and other alternatives as desired. Analysis of pilot study data showed that assigning confidence values to multiple alternatives improved user calibration; a finding consistent with that of Sniezek et al. (1990).

Recording Sales

Confidence is typically recorded on a scale ranging from 0 to 1 or some subset. In this study, this range was divided into increments of five-hundredths (i.e., 0.0, 0.05, 0.10, 0.15,..., 1.0) because research suggests that this is consistent with the respondent's "natural scaling" of decision confidence (Winkler 1971).

Scoring Rules

The purpose of a scoring rule is to encourage respondents to honestly report their confidence in each decision by eliciting values that reflect the respondent's actual belief in the quality of his or her selection. For this to occur, a scoring rule must (1) be understood by the subject so that its implications and the correspondence between beliefs and numerical values can be fully appreciated, and (2) maximize the subject's expected total score only when the subject reports values that correspond to his or her actual beliefs (Stael von Holstein, 1970).

Assume that a subject's true decision confidence is expressed by probability vector $P = (p_1, p_2, ..., p_n)$ for a mutually exclusive and collectively exhaustive set of events, $\{E_1, E_2, ..., E_n\}$. Assume further that the confidence values an assessor reports are represented by $R = (r_1, r_2, ..., r_n)$. A proper scoring rule S exists if S is maximized only when r = p. This requirement is satisfied by only a very few somewhat complex scoring rules that require the respondent to perform high level operations such as exponential, root, or log calculations (Murphy & Winkler 1970). These complex operations make it almost impossible for subjects to quickly compute and fully appreciate the implications of their decisions and the correspondence between their actual beliefs and the values they report. In other words, these scoring rules confuse and may actually interfere with the subject's reporting values reflecting his or her actual beliefs.

A scoring rule that meets the criterion of understandability is the well-known simple linear scoring rule $S_k(r) = r_k$, where k refers to the event that actually occurred and r_k is the confidence probability assigned by the subject to the kth response. Unfortunately, in its simplest form, this scoring rule is not strictly proper because $S(r,p) = \sum p_k r_k$ is maximized by setting one r_i (i.e., the r_i corresponding to the largest p_i) equal to 1.0 and the other r_is equal to 0.0. If $r_{i=k}$, then the subject appears to have complete confidence in the answer that turns out to be correct. On the other hand, if $r_{i\neq k}$, the subject appears totally wrong, but losses nothing because the scoring rule imposes no penalty for being wrong. In other words, a subject maximizes his or her score by assigning a confidence of 1.0 to one answer despite his or her true belief in the quality of any answer.

Despite this limitation, most calibration research has used some variation of this simple linear scoring rule. In fact, comparing three complex proper scoring rules to the simple linear scoring rule, Rippey (1970) reported that the simple linear scoring rule actually produced more reliable results. Likewise, reviewing a number of these studies, Phillips (1970) concluded that the complex proper scoring rules did not yield significantly different values than those collected using a simple linear scoring rule, but, as expected, subjects found simple linear scoring rules more realistic and easier to understand.

Considering these tradeoffs, this study used a variant of the simple linear scoring rule that discouraged gaming and guessing by penalizing wrong answers. The scoring rule used here was:

$$S = r_k - [(\text{largest } r_{i\neq k})/2$$

where S is the score, k refers to the correct alternative, r_k is the confidence probability assigned to that alternative, and $r_{i\neq k}$ are the confidence probabilities assigned to the alternatives that turn out to be incorrect. This variant of the simple scoring rule is easily understood because its implications can be more readily appreciated and the respondent can better understand the correspondence between her beliefs and numerical values she reports. Yet, subjects are encouraged to report numerical values that correspond to their actual beliefs because of the penalty of one-half the largest confidence value assigned to an alternative that is wrong.

Computing Calibration

The most popular calculation for calibration is:

$$calibration = \frac{1}{N}\sum_{t=1}^{T} n_t \left(r_t - c_t \right)^2$$

where N is the total number of responses, n_t is the number of times the confidence value r_t is used, c_t is the proportion correct for all items assigned confidence value r_t, and T is the total number of different response categories used (Lichtenstein & Fischhoff 1977, Clemen & Murphy 1990). Using this formula, perfect calibration is a score of 0.0. The worst possible score, 1.0, can only be obtained when the responses are completely and consistently wrong; that is, $r_t = 1.0$ is always assigned to the wrong answer and $r_t = 0.0$ is always assigned to the answer that turns out to be correct.

Procedures

Subjects were recruited from students enrolled in upper-division, undergraduate courses in information systems and psychology. All participants volunteered for the study and were rewarded course credit as required by American Psychological Association guidelines (1992).

Upon arrival, each subject was randomly assigned to one combination of the two treatment levels, expressiveness or visibility, and the two problems, so as to balance the number of subjects in each cell of the experimental design. The subject then read a two-page handout of instructions that included an example of the expressiveness or visibility display, depending upon the treatment condition assigned, and a description of the navigation procedures and operations the subject would be using to answer the multiple-choice questions. The instructions also included a detailed discussion of the scoring rule, including a table of all possible outcomes that could be referred to throughout the study. The subject was then guided through a demonstration, and questions regarding the procedures and objectives of the study were answered. Each subject then completed a consent form and a short, 11-item questionnaire designed to collect descriptive demographic and background data. To describe the groups' visual acuity, the 16-question Vividness of Visual Imagery Questionnaire (Marks 1972, 1973) was also administered. The subject then began answering the ten questions presented as either visibility diagrams or expressiveness text.

To minimize any question ordering effect, the ten questions in each treatment combination were counterbalanced by order with each question presented in each order position once. This resulted in ten different primary orderings of the ten questions in each treatment. Each question was displayed and data collected using Dell II machines with 17" monitors. The display used in the study was written in ToolBook 5.0 by Asymetrix.

DATA ANALYSIS AND RESULTS

A total of 54 students participated as subjects in the study. Forty subjects, 10 in each group, completed all aspects of the experiment, followed all the instructions and answered all the questions. Although subjects were not given a specific time restriction, on average, they took about 35 minutes to complete all aspects of the study.

Seventy percent of the subjects in the study were information systems majors and the remainder were psychology majors. Most subjects were adult, non-traditional students reporting an average age of 30.3 years. Forty-seven percent of the subjects were female and 80 percent reported that English was their native language. As a group, subjects also reported average to above average (mean = 32.4; s.d. = 9.62) visual acuity as measured by the Vividness of Visual Imagery Questionnaire and self-reported "average" facility with logic and math problems.

Recall that each subject in each treatment answered ten counterbalanced questions. The ten responses from each subject were divided into the first four and the last six responses, again, based on Bauer and Johnson-Laird's research. A calibration score was then computed for each of these two subsets for each subject. These subsets defined the two levels of problem novelty. Calibration based on the first four responses defined the Higher category of Problem Novelty and calibration computed on the subject's last six responses defined the Lower category of Problem Novelty.

Analysis of this data shows that the content of the problem, electrical circuit or people and places, had no effect on either percentage correct (questions 1-4, $F_{(1,36)}$ = .08, p = 0.7 and questions 5-10, $F_{(1,36)}$ = .01, p = 0.9) or user calibration (questions 1-4, $F_{(1,36)}$ = .07, p = 0.7 and questions 5-10, $F_{(1,36)}$ = .01, p = 0.9), so the data was collapsed over the problem content scenarios. In terms of percentage correct, these results are identical to those found by Bauer and Johnson-Laird who also collapsed the data over the same people-and-places and electric circuit scenarios. The mean of user calibration of this pooled data is shown for expressiveness and visibility for the two Problem Novelty categories in Figure 4.

Focusing on the higher category of the Problem Novelty axis shows that subjects using the visibility (V) treatment were much better calibrated, had calibration scores closer to zero, than were those assigned to the expressiveness (E) treatment. Conversely, at the Lower category of Problem Novelty there seems to be little difference between the average calibration of those exposed to expressiveness (E) and those exposed to visibility (V). In other words, over the last six questions, when decisions were more familiar and less novel, exposure to visibility or expressiveness did not differentially affected user calibration.

Figure 4 also shows that average user calibration for the visibility (V) treatment was overall the best, closest to zero, at the Higher category of Problem Novelty (.078). The next best level of user calibration was at the visibility (V) Lower category of Problem Novelty (.100). Comparing these results, the data suggest that the same subjects exposed to the visibility (V) treatment produced better user calibration in the first four tries (.078), when problem novelty was highest, than they did over the last six tries (.100) when problem novelty was lower.

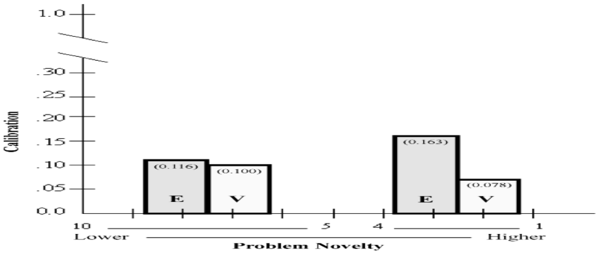

Figure 4 Mean Calibration of Expressiveness (E) and Visibility (V) for Higher and Lower Problem Novelty.

For expressiveness (E), the results in Figure 4 show that subjects exposed to the expressiveness (E) treatment had on average poorer calibration than did those exposed to visibility. Expressiveness produced the poorest average user calibration at both the higher and lower category of Problem Novelty (.163 & .116). However, comparing the two expressiveness (E) bars shows that there was a marked improvement in user calibration from the Higher to the Lower category of Problem Novelty for those exposed to expressiveness (.163 to .116). In this regard, the change in user calibration for those exposed to the expressiveness treatment was as might be expected, user calibration improved as problem novelty decreased.

To assess the statistical significance of the differences in user calibration suggested by the means depicted in Figure 4, a multivariate analysis of variance (MANOVA) was computed using the two dependent variables, user calibration at the Higher and Lower groupings of Problem Novelty, and the independent variable of DSS Locus of Design, either expressiveness or visibility, for each subject. This model produced a Wilks Lambda treatment effect of $F_{(2,37)}$ = 2.8, p = 0.07. Although insignificant at the $\alpha = 0.05$ level, this result does not preclude significant univariate effects. Indeed, in the case of strong positive correlation between the dependent variables (r = 0.45, p = 0.0031), and interaction consistent with that hypothesized in Figure 1, the multivariate test is less powerful than it would be if the data were negatively correlated (Bray & Maxwell 1988, pp. 31-32). In other words, the Wilks Lambda F-value may be confounded by the nature of the interaction between dependent variables.

1.a. ANOVA Results of User Calibration by Expressiveness and Visibility for Higher Problem Novelty (questions 1-4).

Source	df	Type III SS	F-Value	P-Value
E/V	1	.073	5.232	.028*
Error	38	.528		
Total	39	.601		

R^2 = 0.121; * p < .05

1.b. ANOVA Results of User Calibration by Expressiveness and Visibility for Lower Problem Novelty (questions 5-10).

Source	df	Type III SS	F-Value	P-Value
E/V	1	.002	.229	.635
Error	38	.405		
Total	39	.407		

R^2 = 0.006

Table 1: Analysis of Variance of User Calibration for Higher (questions 1-4) and Lower Problem Novelty (questions 5-10).

To clarify the MANOVA results, analysis of variance (ANOVA) was computed for the Higher and Lower groupings of Problem Novelty separately. The results of these analyses are presented in Table 1.

The first ANOVA, Table 1a, shows results for data from the higher category of Problem Novelty. These data show that subjects exposed to visibility produced user calibration that was significantly better than those subjects exposed to expressiveness ($F_{(2,37)}$ = 5.23, p = 0.028). The Bonferroni minimum significant difference of 0.0755 confirms that the difference between 0.163 and 0.078 is significant at the $\alpha = 0.05$ level. For this data, H_0 can be rejected. The evidence shows that for the higher category of Problem Novelty (i.e., when the problems were the most novel), the average calibration of subjects using visibility diagrams was significantly better than it was for those subjects using expressiveness text.

In contrast, results in Table 1b show no significant difference in user calibration as a result of visibility and expressiveness treatment levels ($F_{(2,37)}$ = .229, p = 0.635). The Bonferroni minimum significant difference of 0.0661 exceeds the 0.016 difference in means (0.116 - 0.100). In this case, H_0 cannot be rejected. The data indicate that when problem novelty was Lower and problems were more familiar and less novel, there was no difference in user calibration between subjects using visibility diagrams and those using expressiveness text.

Though not related to the hypothesis, comparisons of visibility (V) or expressiveness (E) across Higher and Lower levels of Problem Novelty resulted in no significant differences. Likewise, comparing visibility (V) at the Higher level of Problem Novelty to visibility and expressiveness at the Lower level of Problem Novelty resulted in no significant differences. Analyses also showed no significant difference in user calibration due to VVIQ subject differences (questions 1-4, $F_{(1,37)}$ = 2.57, p = 0.12; questions 5-10, $F_{(1,37)}$ = .14, p = 0.71) or decision time. These results add to the generalizibility of the main finding that visibility improves user calibration when problems are new and somewhat novel.

SUMMARY AND CONCLUSIONS

The results of a partial test of the theory of DSS design for user calibration are reported. Specifically, a laboratory study was conducted to compare the effects of expressiveness and visibility on user calibration at two levels of problem novelty. The results of this study support the theory. When problems were new and novel, visibility diagrams significantly improved user calibration compared to expressiveness text. Later, when problems became more familiar, less novel, there was no difference in user calibration between visibility and expressiveness.

Bauer and Johnson-Laird (1993) and Speier and Morris (2003) report that diagrams improved decision quality. The results reported here demonstrate that visibility diagrams also improve user calibration. Together, these studies suggest that visibility results in better decisions ***and*** decision makers are better calibrated about their decisions. Specifically, when problems are new and somewhat novel, visibility can both improve performance and help decision makers assess their decision performance, combining to improve user calibration and better outcomes.

For researchers, these findings bode well for the continued development of the DSS design theory for user calibration. To the extent that DSS are applied in novel, one-shot situations, this study demonstrates the importance of visibility in DSS design for user calibration. This study also encourages more research into the effects of different forms of expressiveness, visibility, and inquirability on user calibration at different levels of problem novelty.

For builders and designers of DSS, these results clearly highlight the importance of visibility to decision-making and user performance, especially in new, novel decision environments.

This research also highlights the effects of interface design on user calibration. In particular, the results of this research establish the importance of visibility in DSS design, especially for new and relatively novel decision situations.

REFERENCES (See authors for a more detailed list)

APA, "Ethical Principles of Psychologists and Code of Conduct," *Am Psy*, 47 (1992), 1597-1611.

Bauer, M.I. & P.N. Johnson-Laird, "How Diagrams Can Improve Reasoning," *Psy Sci*, 4, 6 (1993), 372-378.

Bray, J.H. & S.E. Maxwell, *Multivariate Analysis of Variance.* Sage Publications: Beverly Hills, 1985.

Card, S.K., J.D. MacKinlay, & B. Shneiderman, *Readings in Information Visualization: Using Vision to Think.* Morgan Kaufmann Publishers: San Francisco,1999.

Clemen, R.T. & A.H. Murphy, "The Expected Value of Frequency Calibration," *OBHDP*, 46, 1 (1990), 102-117.

Gardner, H. Frames of Mind: *The Theory of Multiple Intelligences.* 10th Ed. Basic Books: New York, 1993.

Gonzalez, C. & G.M. Kasper, "Animation in User Interfaces Designed for Decision Support Systems: The Effects of Image Abstraction, Transition, and Interactivity on Decision Quality," *Dec Sci*, 28, 4 (1997), 793-823.

Helstrup, T., "One, Two, or Three Memories? A Problem-solving Approach to Memory for Performed Acts," *Acta Psy*, 66 (1987), 37-68.

Kasper, G. M., "A Theory of Decision Support Systems Design for User Calibration," *ISR*, 7, 2(1996), 215-232.

Kaufmann, G., "A Theory of Symbolic Representation in Problem Solving" *J Mental Imagery*, 9, 2 (1985), 51-70.

Keren, G., "Calibration and Probability Judgments: Conceptual and Methodological Issues," *Acta Psy*, 77 (1991), 217-273.

Larkin, J. and H. Simon, "Why a Diagram is (sometimes) Worth 10,000 Words," *Cog Sci*, 11 (1987), 65-99.

Lichtenstein, S., B. Fischhoff, & L. Phillips, "Calibration of Probabilities: The State of the Art to 1980," in D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge, England, 1982.

Marks, D.F., "Visual Imagery Differences in the Recall of Pictures," *Brit J Psy*, 64, 1 (1973), 17-24.

Muir, B.M., "Trust Between Humans and Machines, and the Design of Decision Aids," *IJMMS*, 27, 5 (1987), 527-539.

Murphy, A.H. & R.L. Winkler, "Scoring Rules in Probability Assessment and Evaluation," *Acta Psy*, 34 (1970), 273-286.

Phillips, L. D., "The 'True Probability' Problem," *Acta Psy*, 34 (1970), 254-264.

Rippey, R.M., "A Comparison of Five Different Scoring Functions for Confidence Tests," *J Ed. Measure*, 7 (1970), 165-170.

Russo, J.E. & P. J. Schoemaker, " Managing Overconfidence," *Sloan Mgnt. Rev*, (Winter 1992), 7-17.

Shneiderman, B., *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. (Second Edition) Addison-Wesley: Reading, MA, 1992.

Silver, M.S., "Decision Support Systems: Directed and Nondirected Change," *ISR*, 1, 1 (1991), 47-70.

Speier, C & M.G. Morris, " The Influence of Query Interface Design on Decision-Making Performance," *MISQ*, 27, 3 (2003) 397-423.

Stael von Holstein, C.A., "Measurement of Subjective Probability," *Acta Psy.*, 34 (1970), 146-159.

Winkler, R.L., "Probabilistic Predictions: Some Experimental Results," *JASA*, 66, 336 (1971), 675-685.