

## Rule and Instance Based Strategies in Expert Judgment

Jennifer Tsai, Alex Kirlik, Alex Kosorukoff, and Sarah Miller  
University of Illinois at Urbana-Champaign, Urbana, IL

Expert judgment has been conceived in contrasting ways. The naturalistic decision making (NDM) paradigm has put forth a largely instance-based account, viewing experts as relying on a storehouse of cases, such as in the recognition-primed decision (RPD) model. Cognitive psychology has instead advanced largely heuristic- or rule-based accounts, such as in the lens model and fast-and-frugal heuristics. To clarify the relationship between these accounts, we performed two experiments in which novices and experts performed a task explicitly designed to reveal signatures in the data of the use of both rule- and instance-based strategies. Modeling revealed that expert judgment benefited from both the use of linear cue-weighting rules and instance memory. Instance memory use was reflected in experts' (but not novices') ability to handle task nonlinearity, and the finding that expert accuracy across instances was positively correlated with the number of times each instance was historically seen in past experiences.

### INTRODUCTION

Cognitive engineering and cognitive psychology have provided apparently conflicting accounts of the strategies underlying expert judgment. Klein's (1993) RPD model, for example, assumes that expertise draws upon a vast storehouse of past cases or instances. In contrast, cognitive psychology has painted a picture of expert judgment as best described by relatively simple, linear-additive cue integration rules (e.g., Hammond & Stewart, 2001), or even simpler "fast and frugal" heuristic rules (Gigerenzer, Todd, & the ABC Research Group, 1999). Here, we report the results of two experiments performed to clarify the relations between these contrasting views of expert judgment. Do they truly conflict, or could they in fact be complementary? While our research does not, of course, provide complete answers to these questions, it at least represents a step towards better understanding the relations between rule- and instance-based strategies and models of expert judgment.

### MULTI-CUE JUDGMENT & CATEGORIZATION

A multi-cue judgment task is comprised of a set of binary or continuously valued cues that are probabilistically related to a criterion to be inferred. In human factors contexts, for example, the cues may be the values of variables perceptually available from interface displays, and the criterion may be a remote system state or event that must be inferred, estimated, or predicted on the basis of these cues (Kirlik, 2006). In cognitive psychology, research on multi-cue judgment has focused on either linear-additive cue integration rules, as in the lens model (Brunswik, 1952; Hammond, 1955), or on simpler, "fast and frugal" heuristic rules for cue processing that make even fewer demands on cognition than do linear-additive strategies (Gigerenzer et al., 1999).

However, recent research in categorization, as opposed to multi-cue judgment, suggests an important role for alternative judgment strategies – in particular, those that are nonlinear, and consistent with Klein's (1993) instance-based RPD model (Juslin, Olsson, & Olsson, 2003). Juslin and his colleagues proposed a dual process model of judgment that draws upon

both Medin and Schaffer's (1978) instance-based model of categorization and a linear-additive cue integration model to incorporate both instance- and rule-based strategies within a unified framework. They performed two experiments designed to tease apart the conditions under which rule- versus instance-based strategies were used. Their results indicated that strategy use was tied to task characteristics: instance-based strategies were used when categorizing items described by features into a binary criterion (category) in deterministic conditions, while a rule-based strategy was used when estimating a continuously valued criterion on the basis of cues in uncertain conditions.

Both of Juslin et al.'s (2003) experimental tasks were linear, meaning that the relationship between the cue values and the criterion could be best approximated linear-additively (a regression function). As such, it is not clear whether these same results would hold in tasks with nonlinear, rather than linear, cue-criterion relations. It is no secret that on the whole, linear-additive rules outperform human experts (Goldberg, 1965; Dawes, 1971), once human expertise has been used to identify the relevant cues. This implies that in a linear task, even a human expert can do no better than the best linear model using the same cues. This result is almost trivial, given a computer's ability to flawlessly and consistently execute a linear-additive cue integration algorithm.

As such, even if expert judgment is, in some cases at least, instance-based, as the NDM perspective contends (e.g., Klein, 1993; Lipshitz, Klein, Orasanu, & Salas, 2001), direct evidence in support of this claim is not likely to be found by studying expertise in linear-additive ecologies (we consider indirect evidence, such as post-hoc verbal reports of strategies used during "critical incidents," to be suggestive, but less than fully convincing). How could such evidence be found? One answer to this question is to note that, from an instance-based perspective such as NDM, there is no reason to expect peak levels of expert performance to differ in linear and nonlinear tasks. This is so because judgments are assumed to arise largely from memory retrieval, rather than by implementation of a rule that, in even the best case, has no more complex form than a linear-additive function. If the NDM position is correct, experts should be able to transcend the linear-additive limit to judgment performance (Dawes, 1971) in nonlinear tasks by

pattern matching to prior experience. Evidence supporting an instance-based account of judgment could potentially be found by studying expert performance in a task that could not be perfectly performed by any linear-additive strategy, but could potentially be performed flawlessly by an instance-based strategy.

Olsson, Enkvist, and Juslin (2007) undertook this exact approach. Their experimental participants performed a radically nonlinear multi-cue judgment task, where the hypothesis was that participants would spontaneously shift from a futile attempt to find a judgment rule to an instance-based strategy – the only way their task could have been performed even remotely well. Instead, they found their participants trapped in persistent, ineffectual attempts to learn the very complex, nonlinear cue-criterion relation or rule. Furthermore, in order to elicit the hypothesized instance-based, rather than rule-based, behavior in a second experiment described in the same article, Olsson and her colleagues had to: 1) eliminate all uncertainty from their initial nonlinear judgment task; 2) double the number of training trials; and 3) explicitly instruct participants that the relations between the cues and the criterion were “too complex to be comprehended and that the only way to learn the task was by memorizing individual exemplars” (p. 1379). Only after all of these interventions did participants exhibit performance superior to the best linear-additive approximation to the nonlinear task.

We note, however, that the difficulty that Olsson and her co-workers had in generating evidence for instance-based judgment could be due to three features of their experimental paradigm. First, while their task (judging properties of fictional bugs based on cues such as leg length and nose length) is representative of the types of artificial tasks used in basic psychological research, their task likely held little meaning for their participants. Second, the radically nonlinear cue-criterion function underlying their task is unlikely to be found in nature or experience. Given the crucial role of the task environment in shaping cognitive processes (Brehmer, 1994; Cooksey, 1996; Hammond & Stewart, 2001), it remains possible that a less arbitrary, yet still nonlinear, task might be more likely to promote instance-based judgment. In addition, Olsson and co-workers hypothesized that 440 training trials in a single experimental session would be sufficient for allowing participants to develop an instance-based strategy. This amount of learning time pales in comparison to the many years of experience underlying the expertise of those studied by NDM researchers, and indeed by all cognitive engineering researchers aiming to support judgment in professional tasks.

The following two studies described here will attempt to address these possible limitations of the Olsson et al. (2007) experiments, by studying expert judgment in a familiar and meaningful nonlinear judgment task with which participants have at least 10 years of experience. Prior to describing the experiments, however, we first present the formal judgment modeling approach and techniques we use for data analysis. This is crucial because these techniques provide the basis for both our experimental logic and the logic we use to identify possible evidence of the use of both rule- and instance-based strategies in our empirical data.

### Judgment Analysis

The lens model equation (LME) (Hursch, Hammond, & Hursch, 1964; Tucker, 1964) provides a method for assessing performance in multi-cue judgment tasks. The LME is a formula for decomposing judgment performance into a number of meaningful parameters (Equation 1, Table 1).

$$r_a = GR_eR_s + C\sqrt{(1 - R_e^2)}\sqrt{(1 - R_s^2)} \quad (1)$$

| Component | Name                                | Description  |
|-----------|-------------------------------------|--|
| $r_a$     | Achievement                         | Correlation between the criterion variable and the judgment variable                                 |
| $R_e$     | Linear Environmental Predictability | Multiple correlation of the criterion variable with the proximal cues (Environmental predictability) |
| $R_s$     | Consistency of Strategy Usage       | Multiple correlation of the judgments with the proximal cues (Consistency of strategy execution)     |
| G         | Linear Knowledge                    | Correlation between the linear components of the criterion and judgment variables                    |
| C         | Nonlinear (Unmodeled) Knowledge     | Correlation between the nonlinear components of the criterion and judgment variables                 |

Table 1: Components of the lens model equation.

Achievement is a measure of overall judgment quality as represented by the correlation of human judgments with the objective task criterion (the correct response). Of particular interest to us are the LME parameters G and C. G, which can range between -1 and 1, is a measure of how well the judge has tailored his or her weighting of judgment cues to the optimal weights, or to the ecological validities (predictive values) of various judgment cues. Higher G values indicate better knowledge of the manner in which the judgment cues are *linearly* related to the criterion. In contrast, C, which can also range from -1 to 1, provides a measure of a judge’s ability to perform in a manner that is also adaptive to the *nonlinear* aspects of a task’s cue-criterion relationships. As such, a value of C significantly greater than zero indicates that a judge has an ability to use some type of strategy more sophisticated than mere linear-additive cue weighting. Given the task we have chosen, we hypothesize that if our expert participants can benefit from instance memory to out-perform the best linear-additive approximation to our nonlinear task, this ability will be revealed as a value of C significantly greater than zero.

### EXPERIMENT 1

American baseball was chosen as the backdrop for our nonlinear multiple-cue judgment task. This particular domain was selected due to the availability of highly experienced experts, as well as its meaningfulness to them. The task consisted of making judgments of the expected number of runs that would be scored in a given scenario comprised of multiple cues – the number of outs, and the presence of runners on each base. The possibility of 0, 1, or 2 outs, and the presence or absence of a runner on each of three bases create 24 possible cue combinations or scenarios. Actual major league baseball

(MLB) expected runs statistics were obtained for all scenarios. In general, fewer outs and more base runners yield a higher number of expected runs scored. However, the combinatorial nature of the task yields a nonlinear relationship between the cue values (0, 1 or 2 outs; 0 or 1 runner on each of 1<sup>st</sup>, 2<sup>nd</sup>, and 3<sup>rd</sup> base) and expected runs.

**Method**

*Participants.* Twenty-four paid participants completed this experiment. Participants were classified as either experts (n=12) or novices (n=12) based on their past experience with watching baseball games. To qualify as an expert, participants were required to have watched at least twenty-five games per season for the past ten years. Novices were required to have watched less than five games in their lifetime.

*Procedure.* We designed our experimental procedure to be identical to that used by Olsson et al. (2007) to provide the cleanest test of our hypothesis that a natural (rather than contrived), inherently meaningful task performed by participants with many years (as opposed to a single experimental session) of experience would more likely reveal the use of instance-based judgment strategies. Participants were run in a single, two-phase session. In the training phase, they were presented with visual representations of twelve outs-and-bases scenarios (pre-selected from the pool of twenty-four) in a randomized order, and asked to respond with a numerical estimate of the expected (average) number of runs that would be scored by the end of an inning given a scenario. Feedback in the form of a correct answer was supplied after each incorrect response, and participants were required to train to criterion – correctly answering the expected number of runs to 0.1 accuracy for all twelve training scenarios two times in a row – before being allowed to move on to the next phase.

In the testing phase of the study, the remaining twelve unused baseball scenarios were presented in random order for participants to again respond with a numerical estimate of the expected number of runs that would be scored by the end of an inning. Feedback was not provided at any point during this phase, and participants only received one chance to respond per scenario, regardless of the accuracy of their answers.

We should note that Olsson et al. (2007) were required to use an experimental design incorporating training because their participants had no experience with their task prior to experimentation. In contrast, given that our expert participants arrived to the laboratory with at least 10 years of experience relevant to our task, our experiment did not actually require a training phase. This is the reason we report two experiments in this paper. The first experiment allows us to compare our results most directly with those of Olsson et al. (2007), while our second experiment allowed a perhaps more direct measure of expert judgment performance uncorrupted by a somewhat artificial training intervention.

**Results and Discussion**

One expert participant and one novice participant were removed from analyses as outliers due to having given an

expected runs response in excess of six standard deviations away from the mean response for each group of participants.

Despite the nonlinear nature of this judgment task, it is possible to construct both 2-cue and 4-cue linear-additive approximation strategies, as in the following equations

$$Y = .81 - .52X + .43B \tag{2}$$

$$Y = .81 - .52X + .24B_1 + .41B_2 + .64B_3 \tag{3}$$

where:

- Y is the model’s prediction for number of expected runs
- X is the number of outs in the inning
- B is the total number of runners on all bases
- B<sub>1</sub> is a binary cue (1/0) for if there is a runner on 1<sup>st</sup> base
- B<sub>2</sub> is a binary cue for if there is a runner on 2<sup>nd</sup> base
- B<sub>3</sub> is a binary cue for if there is a runner on 3<sup>rd</sup> base

Both models are used in subsequent analyses to provide a linear baseline for assessing participant performance. The regression fit of the 2-cue linear-additive approximation (Eqn. 2) to the nonlinear task was 0.88 (R<sup>2</sup>), while the fit of the 4 cue linear additive approximation (Eqn. 3) was 0.93 (R<sup>2</sup>). Given the combinatorial nature of the expected runs scored task, the high fits of these relatively simple linear-additive approximations may come as a surprise. Yet it is in fact the robustness of simple linear-additive cue combination rules for many aspects of the human ecology that provides the theoretical basis for taking such rules seriously from a psychological perspective (see Hammond & Stewart, 2001).

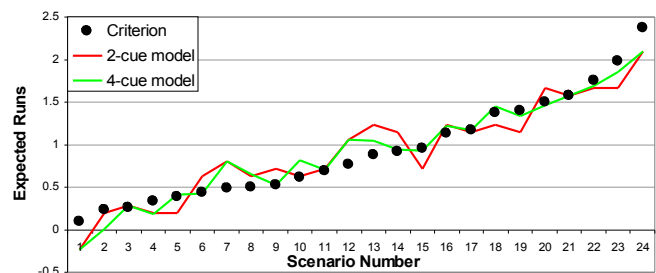


Figure 1. Nonlinear task environment, expected # runs for each scenario.

The nonlinear nature of the task can be seen in Figure 1, in which we plot scenario number on the x-axis, ordered from scenario 1 with the fewest expected runs (2 outs, no runners) to scenario 24 with the most expected runs (0 outs, bases loaded). The y-axis in Figure 1 indicates the number of expected runs, the red line indicates the predictions of the 2-cue linear-additive approximation model (Eqn. 2) and the green line indicates the predictions of the 4-cue approximation model (Eqn. 3). Note that these lines should not be interpreted as normal regression graphs typically are. One must recognize that each value (scenario) on the x-axis represents a vector of cue values, rather than a scalar variable. For the 2-cue model, this vector is (X, B) from the above. For the 4-cue model, this vector is (X, B<sub>1</sub>, B<sub>2</sub>, B<sub>3</sub>). The black circles indicate the value of the criterion for each scenario. The nonlinear nature of the task, given the cues defined above, can be seen in the less than perfect fits of the 2- and 4-cue linear-additive approximation models. If the task could be performed perfectly by a 2-cue

regression model, the red line would lie perfectly on the black dots, and if it could be performed perfectly by a 4-cue model, the green line would lie on these dots. The fact that the green line is able to lie closer to the criterion values than the red line reflects the added degrees of freedom in the 4-cue model.

**Lens Model Analysis.** A lens model analysis was performed for each participant. Values of the LME parameters averaged across all expert participants, all novice participants, and additionally those of the two linear approximation models are shown in Figure 2.

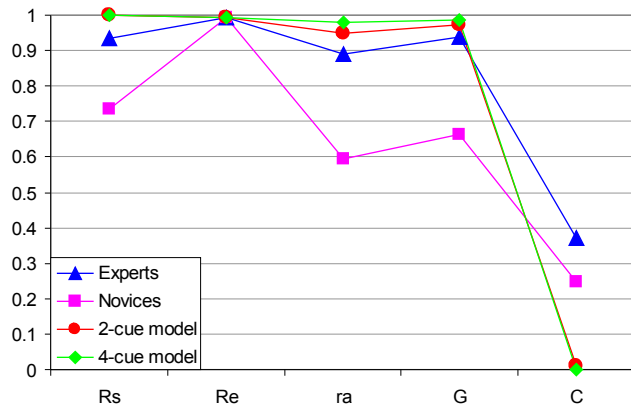


Figure 2. Lens model parameters for experts, novices, and 2 linear models.

On the whole, our baseball experts were able to attain a high level of performance comparable to that of the two linear models in terms of achievement ( $r_a$ ), consistency of strategy execution ( $R_s$ ), and linear knowledge ( $G$ ). Note that the high level of  $G$  for the experts (.9) indicates a high level of adaptivity to the strictly linear relations between the cue values and the criterion (e.g., less runners on base generally implies less runs will be scored; more outs generally implies less runs will be scored). In addition, 10 of 11 experts could be reliably modeled ( $p < 0.05$ ) with either the 2- or 4-cue linear additive models. In contrast, our baseball novices were not able to achieve this same level of performance, garnering lower values on the whole than experts for the same three parameters –  $r_a$ :  $t(20) = 2.30, p = 0.032$ ;  $R_s$ :  $t(20) = 2.24, p = 0.037$ ; and  $G$ :  $t(20) = 2.71, p = 0.014$ . In addition to better linear-additive cue use, domain expertise also contributed to significant use of nonlinear knowledge ( $C$ : 95% confidence interval = [0.164, 0.622]). As discussed previously, the implication of use of nonlinear knowledge on the part of domain experts is suggestive of the possible use of instance memory to overcome the barriers of linear-additive approximation rules for performing the task. Not surprisingly then, novices were not similarly able to make use of nonlinear strategies ( $C$ : 95% confidence interval = [-0.297, 0.553]).

**Individual Differences.** After training, one expert participant (#5) was able to surpass the performance of both the 2-cue and 4-cue linear models by way of near-perfect strategy execution and linear knowledge. This expert demonstrated a striking ability to adapt to task nonlinearity, providing evidence that at least this one human expert could exceed the performance of even the best 4-cue linear

approximation to a nonlinear judgment task (Figure 3). Two experts were able to outperform the 2-cue model, and several others came close. Expert performance was fairly stable across the board, while novices' showed more variation.

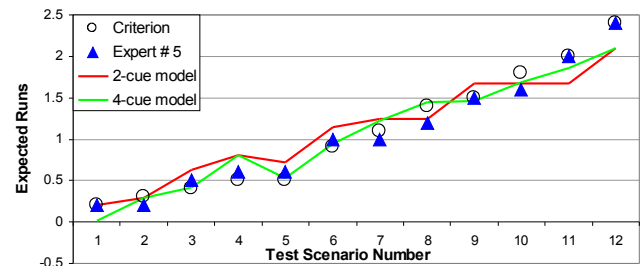


Figure 3: Expert #5 judgments compared to criterion and two linear models.

**Experience with Individual Scenarios.** Because instance-based strategies involve drawing on previous experience, one might expect to find a positive link between judgment performance on any particular scenario and the number of times that scenario has been experienced. Major league baseball statistics were obtained for how many times each of the twenty-four bases-and-outs scenarios is encountered on average. These were correlated with the accuracy of expert participants' expected runs estimates, but yielded weak to no correlation of marginal significance,  $r(143) = 0.144, p = 0.068$  (but see the results from Experiment 2 assessing expert "walk in the door" judgment performance).

## EXPERIMENT 2

Motivated by the observation that the inclusion of a training phase may have affected expert participants' judgment strategies in the first study, Experiment 2 was conducted as a follow-up to more closely examine the effects of (no) training on strategy use and judgment performance.

### Method

**Participants.** Another group of twelve paid baseball experts was recruited to participate in this experiment. The qualifications for being deemed a "baseball expert" were identical to those of the first study.

**Procedure.** In a single session lasting no more than one hour, participants were presented with all twenty-four outs and bases scenarios, and asked to provide a numerical estimate for the expected number of runs that would be scored by the end of an inning. Feedback was not provided, and participants were allowed a maximum of one answer per scenario, regardless of response accuracy. Presentation order of the twenty-four scenarios was randomized for each subject.

### Results and Discussion

Lens model parameters were comparable to those of trained experts found in Experiment 1. Correlating the number of times each scenario is experienced on average with the accuracy of untrained expert participants' judgments yielded a correlation of  $r(292) = 0.312, p < 0.01$ . Compared to the lack

of correlation in experiment 1, it seems that without an initial training phase, relative experience with a scenario is associated with judgment performance on that scenario, with more experience resulting in more accurate judgments. This finding thus provides an additional source of evidence pointing to the existence of instance-based processing in this task. The fact that we did not observe this effect in Experiment 1, in which even expert subjects received training, suggests that this training likely had the effect of prompting experts to acquire a somewhat new strategy for task performance that was less a sole reflection of their long-term knowledge and instead one that benefited from experience and task-specific laboratory training as well.

### GENERAL DISCUSSION

Expert judgment has been conceived in contrasting ways. The naturalistic decision making (NDM) paradigm has put forth a largely instance-based account, viewing experts as relying on a storehouse of cases or instances, such as in the recognition-primed decision (RPD) model. Cognitive psychology has instead advanced largely heuristic- or rule-based accounts, such as in the lens model and fast-and-frugal heuristics. To clarify the relationship between these accounts, we performed two experiments in which novices and experts performed a nonlinear judgment task explicitly designed to reveal signatures in the data of the use of both rule- and instance-based strategies.

Although NDM researchers tend to eschew formal modeling in favor of qualitative models, field observations, and interviews (Lipshitz, Klein, Orasanu, & Salas, 2001), it may be interesting to note that formal modeling enabled us to identify the empirical signature of instance-based strategies in an (arguably) more objective fashion than NDM has itself been able to achieve. Modeling revealed that expert judgment benefited by both the use of linear rules and instance memory. Reliance on instance memory use was reflected in experts' (but not novices') ability to handle task nonlinearity, and the finding that expert accuracy across instances was positively correlated with the number of times each instance was observed over the past 10 years. In addition, one "super" expert was observed to perform the nonlinear task better than even a 4-cue linear-additive approximation having an  $R^2$  value of 0.93 – truly a remarkable achievement.

However, despite uncovering some evidence of instance-based processing, a striking finding of this research was the sufficiency of linear-additive models in describing the judgments of experts in a task with participants specifically chosen to try to elicit evidence for instance-based strategies. This further validates the utility of models and methods based on quantitative methods to describe judgment and decision making of experts (Kirlirk, 2006; Kirlirk & Strauss, 2006).

Finally, we conclude by noting that despite its realism, our task is still artificial in that rarely do baseball fans generate absolute numerical estimates of expected runs scored while viewing a game scenario. Thus, in our current research we are using a paired-comparison design (which scenario will yield more runs?), as we believe this response mode is less contrived

and is likely to allow us to better tap into participants' expertise.

### ACKNOWLEDGMENTS

This research was supported by NSF grant DRMS-045216 to the University of Illinois.

### REFERENCES

- Brehmer, B. (1994). The psychology of linear judgment models. *Acta Psychologica*, 87, 137-154.
- Brunswik, E. (1952). *The conceptual framework of psychology*. Chicago: University of Chicago Press. (*International Encyclopedia of Unified Science*, vol. I, no. 10.)
- Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications*. New York: Academic Press.
- Dawes, R.M., 1971. A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist*, 26, 180-188.
- Einhorn, J. H., Kleinmuntz, D. N., & Kleinmuntz, B. (1979). Regression models and process tracing analysis. *Psychological Review*, 86, 465-485.
- Gigerenzer, G., Todd, P. & the ABC Research Group (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.
- Goldberg, L.R. (1965). Diagnosticians vs. diagnostic signs: The diagnosis of psychosis from the MMPI. *Psychological Monographs*, 79.
- Hammond, K. R. (1955). Probabilistic functionalism and the clinical method. *Psychological Review*, 62, 255-262.
- Hammond, K. R., & Stewart, T. R. (Eds.). (2001). *The essential Brunswik: Beginnings, explications, applications*. New York: Oxford University Press.
- Horrey, W. J., Wickens, C. D., Strauss, R., Kirlirk, A., & Stewart, T. R. (2006). Supporting situation assessment through attention guidance and diagnostic aiding: The benefits and costs of display enhancement and judgment skill. In A. Kirlirk (Ed.), *Adaptive perspectives on human-technology interaction* (pp. 55-70).
- Hursch, C. J., Hammond, K.R., & Hursch, J.L. (1964). Some methodological considerations in multiple-cue probability studies. *Psychological Review*, 71, 42-60.
- Justin, P., Olsson, H., & Olsson, A.-C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General*, 132, 133-156.
- Kirlirk, A. (2006). *Adaptive perspectives on human-technology interaction: Methods and models for cognitive engineering and human-computer interaction*. New York: Oxford University Press.
- Kirlirk, A., & Strauss, R. A. (2006). Situation awareness as judgment I: Statistical modeling and measurement. *International Journal of Industrial Ergonomics*, 36, 463-474.
- Klein, G. A. (1993). A recognition primed decision (RPD) model of rapid decision making. In G. A. Klein, J. Orasanu, R. Calderwood, & C. E. Zsombok (Eds.), *Decision making in action: Models and methods*. Norwood, NJ: Ablex.
- Lipshitz, R., Klein, G., Orasanu, J. & Salas, E. (2001). Focus article: Taking stock of naturalistic decision making. *Journal of Behavioral Decision Making*, 14(5), 331-352.
- Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85, 207-238.
- Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, 116, 2417-2424.
- Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review*, 7, 375-402.
- Olsson, A.-C., Enkvist, T., & Justin, P. (2007). Go with the flow: How to master a nonlinear multiple-cue judgment task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32, 1371-1384.
- Tucker, L. (1964). A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch, and by Hammond, Hursch, and Todd. *Psychological Review*, 71, 528-530.