Quesada, J. F., Kintsch, W., & Gomez, E. (2003). Automatic Landing Technique Assessment using Latent Problem Solving Analysis. In R. Alterman & D. Kirsh (Eds.), Proceedings of the 25th Annual Conference of the Cognitive Science Society. Boston, MA.

Automatic Landing Technique Assessment using Latent Problem Solving Analysis

José Quesada, Walter Kintsch ([quesadaj, wkintsch]@psych.colorado.edu)
Institute of Cognitive Science, University of Colorado, Boulder
Boulder, CO 80309-0344 USA

Emilio Gomez (egomez@ugr.es)
Department of Experimental Psychology, University of Granada
Campus Cartuja, S/N, Granada, Spain

Abstract

Latent Problem Solving Analysis is applied to model the decision processes of expert instructors judging professional pilots' landing technique in a B747 flying simulator, showing that that a memory-based model can do well in the absence of more conscious, logical processes.

Introduction

In previous work, Latent Problem Solving analysis has been applied to laboratory tasks (e.g., Quesada, Kintsch, & Gomez, 2002). This paper extends this idea to a different domain: landing technique evaluation. There is currently no methodology to automatically assess landing technique in a commercial aircraft or a flying simulator. Instructors are an important cost for training and evaluation of pilots, and their use also incorporates a subjective component that may vary from pilot to pilot. In this application of LPSA to landing technique evaluation, we assume that an expert uses his past knowledge to emit landing ratings by comparing the current situation to the past ones, and generates an expanded representation of the environment by composing the past situations that are most similar to the current one.

Selecting the set of variables that should be used to train the model is not a trivial task. Is the visual information to be considered? Which variables are relevant? Our approach was to develop a methodology based on two key ideas: (1) Expert triangulation: while an expert was able to monitor almost every single variable relevant (complete information expert) another one was limited to watching a real-time plot of a very limited set of variables chosen by him (reduced information expert). If the judgments of these two experts are highly correlated, the variables selected by the reduced information expert have sufficient explanatory power to perform the evaluation. (2) Modeling of the landing evaluation task using Latent Problem Solving Analysis (LPSA, e.g., Quesada et al., 2002) on the variables selected by the reduced information expert. The resulting system was able to evaluate landing performance automatically. We work under the assumption that an expert trusts her past knowledge to emit landing ratings in a significant way. As pointed by Landauer (2002), most people use conscious logic only to narrow realms where they also possess large volumes of hidden intuitive knowledge. Experts are supposed to be attuned to the constraints of their environments (e.g., Ericsson & Lehmann, 1996; Vicente & Wang, 1998) in a way that presumes automaticity. Our proposal does not deny that the expert is also employing some other, more analytical method. However, we would like to point out that a memory-based model can do well in the absence of more conscious, logical processes.

The landing task

The variation in the requirements of the landing task is immense with landing conditions such as wind, gust, and visibility. However, our data collection experiment was designed to be very simple with the idea of minimizing the variability due to uncontrolled factors. The manufacturer of the aircraft normally provides charts with the preferred value of a variable (e.g., Glideslope) given some possible values of other variables (e.g., the air speed). In other cases, it is the Government who provides the charts. The landing is usually divided into approach, flare, and touchdown. A graphical, simplified description of these three concepts can be viewed in Figure 1:

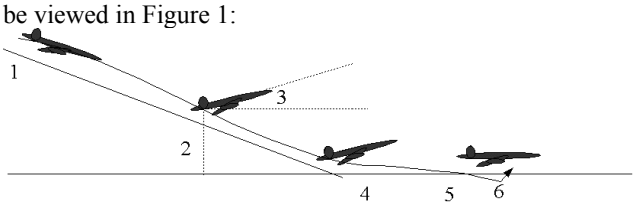

Figure 1: Basic scheme of a landing. (1) Glideslope. (2) Flare initiation height. (3) Pitch rate (4) Glideslope predicted intercept point. (5) Touchdown point. (6) Reversal in the direction of the vertical acceleration at touchdown point.

To evaluate the landing technique, we selected a set of five criteria consulting several landing technique instructors and simulator specialists. The list included: (1) Flare initiation height: The flare has to be initiated at a particular height; this height is not rigid as lower flares can be compensated by a higher pitch rate for example. Three levels (too high, correct, and too slow) were used. (2) Thrust Reduction: The reduction should be progressive, and it has to be started in a particular moment in time. It was judged using three levels: (too fast, correct, and too slow). (3) Pitch rate. The pitch was evaluated using five discrete levels, from too high to too low. (4) Overall landing score. This is a general rating that expressed how good the landing was, from one to five. In a sense, it is not a summary of the former measures, as it adds new information. Some landings can have, for example, an incorrect Flare initiation height, but end up

getting a five out of five, because it was compensated with other means. The possible ways in which these different grades can be observed and their interaction enables a complex set of data to model.

The problem of variable selection and complexity reduction

In some circumstances, the modeler has to be very knowledgeable about the task to be able to create a successful model (for example, chess modelers tend to be good chess players). Although this point may seem redundant and obvious, it is very important, since it is not always the case that the modeler can invest the long time required to master the task to model. The alternative approach to task modeling is to ask the experts what information they use, and what procedures they have developed to perform the task. In this line, expert knowledge elicitation techniques have been developed to try to 'extract' the knowledge from human experts and 'insert' it into the system. Thus, most expert systems are rule-based systems. The approach that we have taken here is different. A basic idea is that experts are able to confront very complex tasks because they have managed to reduce the dimensionality of the respective problem spaces of their jobs thanks to massive amounts of experience. There is a need to translate this dimensionality reduction to the system that is going to perform in their same environment.

Two reductions in the dimensionality of the task are performed to represent the variability in the environment in an efficient way: (1) The one suggested by the expert selection, using the triangulation methodology, and (2) the one performed by LPSA's SVD when the lower-dimensional corpus is created. They are explained in the following two sections.

The triangulation of expert judgments

In this section, we present a possible solution to the problem of variable selection. It uses a configuration of two experts, who perform the task in two very different conditions.

The question is: How do we know which variables a model should pay attention to? It is hard to imagine that our information processing system keep track of every dimension that could possibly be registered. For example, a high fidelity flying simulator can log up to 10000 variables, each with a precision (sampling ratio) of 1/100 seconds. Since it is recorded in the log files, we can assume that this amount of information is available to the pilot and copilot in the commercial aircraft simulated. Of course, in a particular temporal moment *t*, the human components of the system (pilots and ATCs), are aware of a very small proportion of these variables, and the focus of attention is changing from *t*1 to *t*2… to *t*n. It is computationally unfeasible for a cognitive system (either human or artificial) to work in such a high dimensional space.

As a step towards solving this problem, we present the expert triangulation method. It is very simple and susceptible to be applied in a variety of expertise domains. The basic idea is that if we cannot model an expertise field because of its complexity we can use two experts with different access to the information available to discriminate the importance of each variable in the task they perform. A first approach, quite used in modeling work, is the effort to model directly the expert behavior using as many as possible of the variables he can access in his normal, daily performance. Let us call this expert 'the complete information expert'. However, when the task is complex, trying to model the whole situation often proves itself to be an excessively difficult task. Some theories do propose ways of selecting the relevant parts to model. This selection is a priori, that is, the assertion 'The expert is using variable X but not Y' is part of the theory. What we propose is to use a second expert to do the variable selection in a non-theory-driven way. The second expert will have limited access to the variables in the system (for example, he can only plot a limited number). For that reason, this second expert is called 'the reduced information expert', and is forced to select a small set of variables. The model will be created to reproduce the behavior of this expert, and this is often a key step since the modeling task can change from being intractable to being tractable. Note that the theory does not have a priori assumptions about what are the task's most important variables: the expert does (see Figure 2).

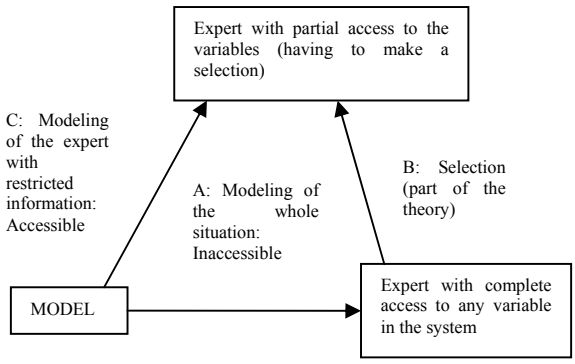

Figure 2: The triangulation technique

Construction of the reduced-dimensional space

The second way of reducing the complexity of the problem is performing dimension reduction. The dimension-reduction step and its properties to explain learning and generalization are important in several cognitive theories (e.g., Edelman & Intrator, 1997; Rumelhart, Smolensky, McClelland, & Hinton, 1986). The algorithm used in LPSA is the singular value decomposition (SVD) of the frequency matrix of states by landings, and the reduction in the number of singular values. As a result, we obtain a representation of both states and landings in the same space. Any new landing that is not in the space can be represented as a linear combination of the vectors of its states. We can predict the ratings of any new landing by averaging the ratings of the *k* known nearest neighbors of the vector representing the landing in this space (see e.g., Duda & Hart, 1973 p.103;

Hastie, Tibshirani, & Friedman, 2001, p. 415 for a description of nearest neighbors algorithms). To construct the space we used the variables that the reduced information expert was using, as suggested by the triangulation technique.

Method

10 pilots performed 40 landing each. We manipulated wind direction and intensity at 6 levels, using incomplete counterbalanced design to control for order effects. The levels selected were 30, 20, 10, 0, and -10 (tail wind) knots. Since wind conditions influence the landing procedure, the ideal experimental design would be to select a representative sample of all the possible wind directions during landing. However, with the limited number of participants and simulator time assigned we preferred to control the presentation order effect by means of a counterbalancing design. Another important factor is gusts. We only used front and tail wind and no gust, again to simplify the experimental conditions and maximize comparability.

The two experts used a set of criteria to rate the landings, described in 'the landing task' section. Both experts rated all landings. The reduced information expert was allowed to select and plot as many variables as he wanted, with the limitation that they should fit in his 20" computer screen. He plotted only the following five variables: Vertical acceleration, Radio altitude, Pitch rate, Rate of descent, Pitch angle. There was still some space left, which implies that the expert considered that he did not need to plot any more variables. It turned out that rate of descent and pitch angle were barely used, as they were never referred to when the expert explained his ratings to the experimenters. Thus, rate of descent and pitch angle were omitted from the analysis. The expert in the copilot seat (complete information expert) used a huge array of information, since he was exposed to the same environment as the pilot, being able to see the runway approach and feel the movements of the aircraft when the wheels touched the ground, for example. The basic idea was to calculate the agreement between the two human graders (reduced and complete information experts). If the agreement was very low, the judgments are too subjective, and a possible automated method of assessing landing technique is hard to validate. Then, the same agreement would be calculated for each human expert and LPSA.

To do the model selection part where we tried to find the right parameter set, the criterion used was the average correlation between the model and the human ratings of both human graders.

Corpus creation

The states in each landing were stripped off of all the variables except for the reduced information that the expert was actually using: flare initiation height, thrust reduction and pitch rate. Since the average duration of a landing when the starting point is 500 feet was about 15 seconds, and the sampling ratio was 10 samples a second, the average number of states per landing was 150.

The flare initiation height, expressed as feet, was transformed (rounded) to be multiples of ten (e.g., 112 feet would be 110, 89 would be 90, etc) and the vertical acceleration and thrust reduction were rounded to the nearest integer (e.g., a vertical acceleration of -9.8 would be -10, and a thrust value of 3.2 would be 3). This rounding is necessary because LPSA assumes that a landing is a sequence of states, and the continuous flow of these values has to be discretized. Since decimal values are not relevant, and humans would consider that, for example, an altitude of 45 feet is the same as 46 or 47 feet for most purposes, we applied the rounding in our model.

The original sampling ratio was 10 times a second. That made a total number of 569 unique states in 400 landings. Although LSA has been applied to text corpora with the same number of types, and even several orders of magnitude more, the limited number of landings imposes a severe restriction. Most known learning mechanisms, including LSA, need several repetitions of the units to learn them. That is, LPSA learns better when a good proportion of the states can be found in more than one context. The transformations and rounding that we performed were serving the purpose of decreasing the number of different states in the corpus. When the states are described using continuous variables, and these variables are sampled at a fast rate, an non-rounded corpus would have as many unique states as the total number of states (that is, each state would appear in the corpus only once, leaving little room for learning). The variables were joined with underscores to make them a single token, and use space as token separator. This way, a state in the system was represented as a token as follows: "flare initiation height_thrust Reduction_pitch rate". This token is the equivalent to a word in standard LSA. The matrix of states by landings was created, and an SVD was performed on it. After the decomposition, the biggest N (where N is a free parameter) dimensions were kept. The parameter manipulation is explained in the results section (model selection). A web interface to the 400 landings graphs (mimicking reduced information expert's display), experimental conditions and ratings used in this paper can be visited at http://lsa.colorado.edu/~quesadaj/adriVisor.cgi. The complete corpus is also available upon request.

Apparatus

The Netherlands' National Aerospace Laboratory (NLR) National Simulation Facility (NSF) simulator was used. A Boeing 747 cockpit was installed consisting on a side-by-side full glass airliner cockpit with a layout equipped with six programmable CRTs. The airport selected was San Francisco airport, because it is situated at sea level; this feature is desirable because the radio altitude and the barometric altitude tend to be the same.

Results

Significance tests. The polychoric correlation was selected because of its suitability for analyzing judgment data on ordinal scale. To test the hypothesis of the correlation being significantly different from zero, we used resampling methods, concretely a randomization test. That is, we used a Monte Carlo approach to estimate the probability of our results (correlations) being obtained due to a bias in the computation. For example, imagine that both the expert and the model say simply 'correct' all the time. The bias is 'say always correct'. The correlation human-model would be 1. As well, if we randomly rearrange the values of the model or the expert, so that they do not line up with each other (for example, the model rating for landing 1 would be matched to the expert judgment for landing 67, and so on), the correlation would still be one. In this extreme case of bias, having a high correlation between the model and the expert does not mean any merit for the model, since any random rearrangement of the data would obtain the same correlation. The randomization tests performed were conducted resampling 500 times.

Model selection. We created several corpora modifying the number of dimensions (100, 150, 200, 250, 300, 350, and maximum dimensionality, 400) and the number of nearest neighbors used to estimate the landing ratings (from 1 to 10). Another manipulation was the inclusion or exclusion of a time tag, and the type of weighting scheme used (log entropy vs. none). This way, the possible combinations of levels were (7 x 10 x 2 x 2) = 240. For each of these combinations of levels, we used leave-one-out to calculate the ratings for the landing excluded. The estimated ratings for each of the 400 landings were then correlated with the real ratings. The combination of levels that best correlated with both humans was selected, and that was: Corpus with 200 dimensions, 5 Nearest Neighbors, no weighting, no time tagging).

Model fitness

The first thing to observe is that the average agreement between human experts was not very high (polychoric correlation .48, see boxed bars in Figure 3). To our knowledge, there are no studies that report statistics on specifically landing technique experts, so we will use general expertise for comparison. Shanteau (2001, p. 237, table 13.2), presents data on consensus (agreement) between experts in different domains. The landing technique experts reported in this study had an inter-rater reliability in line with Clinical Psychologists (.40), Stockbrokers (<.32), polygraphers (.32) and Livestock Judges (.50). Their agreement is lower than the ones reported for Weather forecasters (.95), Pathologists (.55), Auditors (.76) and Grain Inspectors (.60).

The average correlation between the model, and the reduced information expert was about the same as the correlation between the two humans (.48 vs. .46, boxed bars in Figure 3). Note that the ceiling for the model is the correlation between two humans doing the task; a model that correlates with one human better than two humans correlate with each other is under suspicion. It seems that the judgment on thrust reduction is particularly difficult for the two experts to agree (human-human correlation of only .27).

One of the LPSA assumptions is that experts perform dimension reduction to represent their environments. The equivalent model (5 nearest neighbors, no weighting, no time stamping) without performing dimension reduction (that is, using 400 dimensions, which is the shortest dimension of the matrix) correlates with humans (on average for all criteria) only .26, which can be interpreted as evidence for dimension reduction in the representation.

The randomization test for the different criteria showed that all of the agreements between human judges were highly significant: Flare initiation height (.52, p = .002[1]), thrust reduction (0.27, p .002), pitch rate (.46, p= .002) and overall landing performance (.61, p = .002).

The equivalent model without dimensionality reduction (400 dimensions, 5 neighbors, no weighting, no timestamp) produced .37, .08, .57, .50 correlations for the above used criteria respectively.

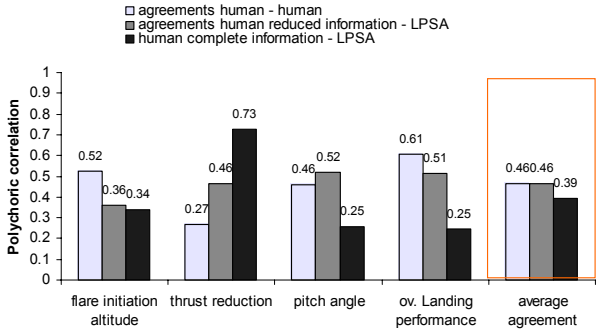

Figure 3: Agreement between the model and the reduced information expert for each of the rating criteria

In our design, we tried to mimic the reduced information expert (the one that had access to only a few selected variables) since the model used the variables this expert utilized. However, having a good correlation with the complete information expert (located in the copilot seat) is desirable too, so in the process of finding the right parameters, the models were selected for their correlation with both humans. Figure 3 presents the correlations obtained for the complete information expert. Note that the only criterion where the model correlates with any of the experts more than they correlate to each other is thrust reduction. Thrust reduction seems to be a very difficult feature to judge, since the agreement between human experts is the lowest (.27) and also it is the one in which the

[1] A *p* value of .002 indicates that none of the polychoric correlations for the randomizations was higher than the observed one, being the proportion 1/501 = .002

reduced information expert obtains the lowest test-retest reliability (0.538, see test-retest measures).

All the polychoric correlations between the reduced information expert and the model were significant (p = .002). So were the correlations between complete information expert and model.

Test – retest measures

One common method to assess how accurate human raters are is the test-retest correlation. It simply consists in having the same expert grade twice the same item in two different temporal moments, preferably distant in time. It is well known that humans have imperfect test-retest reliability. In our study, we asked the reduced information expert to reevaluate a random sample of 100 plots displayed in the same way he experienced during the experiment. The plot contained wind information, but all other information that could identify the landing (pilot name, landing number, ratings etc.) was removed from the graph. The reassessment took place about one 8 months after the end of the experiment. The reliabilities were .64, .53, .84, and .72 for flare, thrust, pitch and overall score respectively.

To our knowledge, there are no studies that report statistics on specifically landing technique experts, so we will use other domains of expertise to figure out how our reduced information expert stands. The average test-retest reliability (0.69) is better than some other studies of reliability of expert judgments reviewed in Shanteau (2001, p. 237, table 13.3), concretely better than for Clinical Psychologists (.41), Stockbrokers (<.40), Grain Inspectors (.62) and Pathologists (.50). His test-retest reliability is however lower than the one reported in the same work for Weather forecasters (.98), Livestock Judges (.96), Auditors (.90) and polygraphers (.91). It is worth noticing that a computational model such as LPSA has a test-retest reliability of 1, and that could be viewed by the trainees as a good feature.

Application of the model in a non-structured corpus

A cognitive system (human or machine) exposed to expert-level amounts of experience in a non-structured environment will show a very poor performance, similar to those of novices. Product theories of expertise (e.g., Vicente & Wang, 1998) propose that the amount of environment structure is the main explanatory factor for the expertise advantage, and LPSA should be able to reflect this fact. To test this hypothesis we run exactly the same simulations on an artificial corpus with 400 landings where the states for each landing were randomly sampled from the original corpus. This random corpus contained landings where all the variables changed randomly for the (average) 15 seconds that a landing lasts. In the hypothetical case of having a human exposed to a domain similar to such a non-structured environment, the amount of learning obtained by the human after the long-term experience would be very little. This poor learning would be reflected by a poor ability to predict future states, and the landing rating case, a poor rating skill, and the LPSA model reflects that in Figure 4.

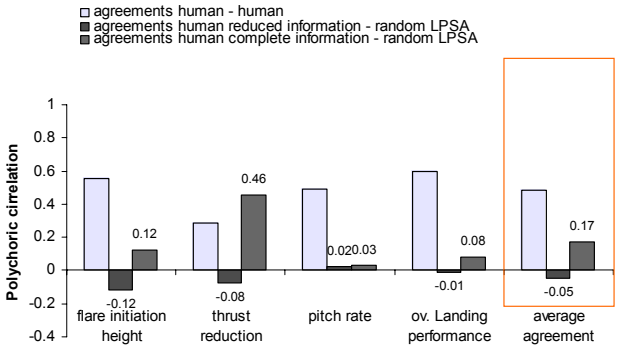

Figure 4: Agreement between the model and the reduced information expert for each of the rating criteria when the model has been trained on a corpus where the environment changes randomly.

Discussion and conclusions: theoretical and practical implications

Theoretical implications

The evaluation of landing technique is a complex task. It takes several years to learn the basics to be able to land a plane, and even longer to be able to evaluate the quality of a landing in a consistent way and give advice on how to improve it. Complex tasks are usually explained in cognitive science referring to constructs as problem solving, mental models, or reasoning. LPSA shows that simple ideas such as similarity-based processing and pattern matching could have a role even in cognitively complex tasks. LPSA is a very simple computational model based on the analysis of massive amounts of knowledge. It assumes that representation takes place in a representational space that has fewer dimensions than the external (distal stimuli) space represented. It also assumes that humans retrieve the most similar past experiences to the current one automatically. The response to the current situation (in this case, the grading of a landing) occurs partially because the ratings of past landings which are similar to this one 'come to mind', and the response is a composite of those ratings.

LPSA requires a lot of experience before it can do this retrieval-based rating, as do humans, and this experience is useless when the environment has no constraints. An important critic can be raised in that we are not giving the model the same amount of practice as the expert has, since we used only 400 landings in a very limited set of environmental conditions (6 different wind strengths, no changes in direction) in only one runway. Ideally, the system would be trained with the particular circumstances of relevance (several runways, wind conditions, aircrafts, etc.). The model has strictly 400 landings of practice in very limited wind conditions, and then, it cannot be really considered comparable to an expert instructor, who has experience in a much more varied environment. In this

sense, we do not want to argue that the current data and results presented are a complete model of landing technique evaluation, or that it can substitute instructors in their task. It must be demonstrated that the model as it is developed here can render similar performance in a wider set of conditions. However, there are reasons to believe that the system can scale up reasonably well. LPSA has been applied to corpora far bigger than the one used here. In the context of control of dynamic systems, the corpus used contained the equivalent of three years of daily practice. When the same ideas on knowledge representation are applied to semantics and text comprehension, the corpus used represents the exposure to printed text that an average human may have by the time she reaches college level, and this is several years of practice.

Practical implications

One important practical conclusion that we want to draw is that it is possible to construct systems that grade landing technique automatically as well as humans, if we consider that the limit of performance for such a model is the human-human agreement. The correlation human-human was low (0.46) but in the range of some other areas reported (Shanteau, 2001). Some authors are extremely critical with the efficiency of human experts doing their tasks: 'Expert did better [than the actuarial model] in only a handful of [the tasks reviewed], mostly medical tasks in which well-developed theory outpredicted limited statistical experience' (Camerer & Johnson, 1991, p. 197).

The advantages of automatic landing technique evaluation are many: (1) Reduced cost of the evaluation. (2) Increased objectivity in the evaluation, making comparisons between different pilots more reliable. Although all the instructors try to emit an accurate judgment, different experts have different subjective criteria for the evaluation. (3) Decrease the influence of the instructor. Lintern and collaborators (Lintern, 1990) pointed out that "Although probably a slight exaggeration, it is frequently asserted that the flight instructor is the greatest source of variance in the pilot training equation" (p. 326). Pilots might be more confident in their own recently acquired skills if they know that they evaluation has been done automatically and is equal for each one of them. (4) Perfect Test-retest reliability. (5) The model is not exposed to factor such as psychological or physical strain. (6) It is always available and can be triggered by the trainee at will. (7) The model can rate as many landings as time enables, etc. In a large-scale application of the model (for a training and evaluation department, for example), we can imagine that 500 pilots need to be evaluated. In that situation only a small proportion of randomly sampled landings (that can be kept from previous sessions) must be evaluated by humans; the rest is performed by the system. Since the model has different landing criteria, it could emit recommendations such as: 'In this landing, you initiated the flare too high, and reduced the thrust too late. Try to take it into account for the next one'.

Acknowledgments

The simulator and expert time was possible thanks to a grant supported by the European Community - Access to Research Infrastructure action of the Improving Human Potential Program under contract number HPRI-CT-1999-00105 with the National Aerospace Laboratory, NLR This research was also supported by Grant EIA – 0121201 from the National Science Foundation.
Our acknowledgements to Tom Landauer, Simon Dennis and Bill Oliver, for proposing interesting theoretical issues. We are grateful to Kim Vicente and John Hajdukiewicz for sharing experimental data and insightful discussions.

References

Camerer, C. F., & Johnson, E. J. (1991). The process-performance paradox in expert judgment. How can experts know so much and predict so badly? In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise*. Cambridge: Cambridge University Press.

Duda, R. O., & Hart, P. E. (1973). *pattern classification and scene analysis*. New York: John Willey and Sons.

Edelman, S., & Intrator, N. (1997). Learning as extraction of low-dimensional representations. In D. Medin & R. Goldstone & P. Schyns (Eds.), *Psychology of Learning and Motivation* (Vol. 36): Elsevier Science.

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology, 47*, 273-305.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *the elements of statistical learning*. New York: Springer.

Landauer, T. K. (2002). Some remarks on consciousness by a somewhat maverick cognitive scientist.

Lintern, G. (1990). Transfer of landing skills in beginning flight training. *Human Factors, 32*(3), 319-327.

Quesada, J. F., Kintsch, W., & Gomez, E. (2002). A theory of Complex Problem Solving using Latent Semantic Analysis. In W. D. Gray & C. D. Schunn (Eds.), *24th Annual Conference of the Cognitive Science Society* (pp. 750-755). Fairfax, VA.: Lawrence Erlbaum Associates, Mahwah, NJ.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, E. (1986). Schemata and sequential thought processes in PDP models. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge, MA: MIT Press.

Shanteau, J. (2001). What does it mean when experts disagree? In E. Salas & G. Klein (Eds.), *Linking expertise and naturalistic decision making*. Mahwah, NJ: Lawrence Erlbaum Associates.

Vicente, K. J., & Wang, J. H. (1998). An ecological theory of expertise effects in memory recall. *Psychological Review, 105*, 33-57.