

Supporting Joint Human-Computer Judgment Under Uncertainty

Sarah Miller, Alex Kirlik, Alex Kosorukoff, & Jennifer Tsai
The Beckman Institute and The Human Factors Division
University of Illinois at Urbana-Champaign
Urbana, Illinois

In this paper we present a concept and interface design aimed at combining expert human judgment with computational support. The goal of this design is to leverage the strengths and simultaneously compensate for the weaknesses of both the expert and a computational model. In order to test the design, we created a task modeled after fantasy baseball, which requires competitors to predict the performance of actual Major League Baseball (MLB) players over the course of the season. The most substantial and challenging aspects of the design involved how to both welcome expert input on a case-by-case basis, yet also provide visual guidance for how these inputs should reflect an appropriate degree of regression to the mean, or reliance on base-rate information. Results showed that the joint human-model system resulted in better performance than a model, which was based, in part, on past performance. The joint system also outperformed unaided or partially-aided experts in some cases but only equally as well in other cases. Design implications and future directions are discussed.

INTRODUCTION

Decision support systems and other types of automation are becoming increasingly integrated into a variety of operational contexts, including air traffic control (Wickens, Mavor, Parasuraman, & McGee, 1998), medical diagnosis (Morrow, Wickens, & North, 2006), and airport luggage screening (Weigmann, McCarley, Kramer, & Wickens, 2006). In these domains, expert judges are constantly required to act on the basis of conflicting, incomplete, uncertain, and constantly changing information (Thomas & Cook, 2005).

Designing systems to support expert judgment under uncertainty requires a delicate balance between allowing the judge to provide input where it is appropriate, while providing support via computational techniques where they are appropriate. This requires identifying those aspects of a task that are best performed by formal models or methods that are not subject to known cognitive biases and limitations, and those best performed via computational modeling. In addition, supporting judgment under uncertainty also requires identifying which aspects of a task benefit from human intervention, either because of some intrinsic or technical limitation on the ability to formally model human expertise, or because of the human's access to information unavailable to the computational system. Once these issues have been addressed, we then turn our attention toward creating an interface that promotes a symbiotic interaction between the human and computational components.

In order to illustrate and test this concept, we take a problem that is often encountered in operational contexts, wherein the judge is asked to provide a judgment or prediction about a specific instance or case. In this type of task, judges must reason about, and appropriately balance, case-specific information pertaining to an individual judgment with base-rate information characterizing the entire population or environment about which judgments are being made. Our design utilizes information visualization and computational support to aid judges in finding an appropriate balance between base-rate and case-specific information, which is very difficult for judges to do when the environment is highly

uncertain (Strauss & Kirlik, 2006).

In the following sections, we first identify those task components that are likely to be best performed by expert human judges, those components that are likely to be best performed by computational, linear-additive models, and the information visualization techniques that can be used to integrate input from the human judge with support from the computational model. We then discuss the experiment conducted to test our design.

INTERFACE AND MODEL DEVELOPMENT

Almost without exception, computational models using linear-additive rules have been found to consistently outperform expert judges in accurately and consistently weighting and integrating sources of information into a judgment (Dawes, 1971; Goldberg, 1968). Experts, on the other hand, have been found to perform better than computational models at identifying relevant information, or cues, that are not already included in the model (Meehl, 1954). Combining human input for cue values with computational support for weighting and adding the cues together is known as bootstrapping. This technique has been found to be an effective tool for supporting combined human-model systems (Camerer, 1981).

In operational contexts, different degrees of uncertainty may exist in a given task, so judges must appropriately balance case-specific and population-specific information depending on this degree of task uncertainty. The appropriate balance is based on the degree of uncertainty inherent in the task environment. When the environment is less predictable, the prediction should be closer to the population mean. As case-specific information become more reliable, the optimal prediction weights case-specific cue values more heavily. To illustrate this concept, let's say a weather forecaster is asked to predict what the temperature will be one hour from now, and what it will be one year from now. Since the temperature one hour from now is likely to not be much different than the current temperature, it can be well predicted using the current temperature. In other words, the forecaster should weight the case-specific information more heavily than the base-rate, or

average climate temperature. However, the temperature one year from today is not likely to be well predicted from the current temperature, so she should weight the average seasonal, or mean, forecast more heavily than the current temperature. In highly uncertain environments, judges often exhibit regression bias, where they fail to appropriately regress their judgments toward the mean and instead place too much weight on the case-specific information (Horrey, Wickens, Strauss, Kirlik, & Stewart, 2006; Kirlik & Strauss, 2006; Strauss & Kirlik, 2006). Therefore, if human judges are to provide input regarding cue values, support is needed to overcome regression bias. Our approach is to provide this support through a combination of information visualization techniques, which can be used to help judges reason about uncertainty, and computational techniques, which can be used to overcome biases and limitations. The next section discusses a method for overcoming regression bias, and provides a technique for measuring overall judgment quality.

JUDGMENT ANALYSIS AND THE SKILL SCORE

One method to analyze judgment quality is through an extension to Brunswik’s (1956) lens model (also see Kirlik, 2006) called the skill score decomposition (Murphy, 1988). The skill score decomposition can also be used to understand the difficulties judges have in dealing with uncertain information. Skill score, a scalar measure of the quality of judgment performance, is a function of the correlation between human judgments and the criterion being judged (called “achievement” in lens modeling), regression bias, and base rate bias. Equation 1 and Table 1 summarize the relationships among these quantities.

$$SS = r_a^2 - \left(r_a - \frac{s_s}{s_e} \right)^2 - \left(\frac{\bar{Y}_s - \bar{Y}_e}{s_e} \right)^2 \tag{1}$$

Table 1. Components of Murphy’s skill score decomposition. In addition to the previously defined factors, s_s and s_e represent the standard deviation of the judgment and criterion \bar{Y}_s and \bar{Y}_e represent the mean of the judgment and the criterion.

Component	Name	Description
SS	Skill Score	A relative measure of “actual” judgment quality.
r_a	Achievement	Degree of linear association between judgment and situation. “potential skill”
$\left(r_{yo} - \frac{s_y}{s_o} \right)^2$	Regression Bias	Degree that s.d. of judgments is not scaled to s.d. of criterion distribution and imperfect achievement
$\left(\frac{\bar{Y} - \bar{O}}{s_o} \right)^2$	Base Rate Bias	Degree that average judgment does not match the base rate of occurrence in the situation

According to Equation 1, in order to make high-quality judgments, it is necessary to not only have high lens model achievement, or correlation, but to also exhibit low regression and base rate biases. A base rate bias exists when the mean of the judgment distribution is not equal to the mean of the criterion distribution. In other words, a base rate bias means

that the judge’s predictions are consistently under- or over-estimated with respect to the criterion. As described in the previous section, a regression bias exists when the judge does not adaptively balance base-rate and case-specific information. Regression bias is zero only when the judgment distribution has a standard deviation that is equal to the product of achievement and the standard deviation of the criterion distribution. A judge who can achieve this goal, therefore, can be seen as adaptively regressing judgments to the mean when his or her own predictive ability (achievement) is low, and in contrast, making increased use of case-specific instead of base-rate information when his or her ability to make accurate predictions from case-specific information is high.

METHOD

We first selected and performed a preliminary analysis of a task involving predicting player performance in fantasy baseball. This task was chosen for its richness and because we could obtain highly experienced experts. Specifically, we focused on developing a system to aid fantasy baseball experts in making predictions of the end-of-the-year “dollar” values for baseball players.

Space limitations prevent in-depth discussion about the rules and goals of fantasy baseball, except for the most relevant components of the task. In typical fantasy baseball leagues, five categories, or cues, are used to determine player worth, or “dollar value.” For hitters, the five relevant cues are runs scored, home runs, runs batted in, stolen bases, and batting average. Five different, but analogous, cues are used to describe pitcher performance. These five cues (separately for hitters and pitchers) can be combined using a linear-additive model to determine the contribution of each cue to the overall dollar value. Perhaps not surprisingly, we found that the ultimate dollar value of a player, representing his contribution to a fantasy baseball team over the course of a season, could be well predicted ($R > 0.95$) for both hitters and pitchers alike with knowledge of these cue values. The reason this is not surprising is that these are the same cues that determine a fantasy team’s ranking in league standings. This means that the dollar value prediction task could be performed extremely well with simple linear regression, but with one important caveat. The cue values themselves are quite unpredictable from year to year, as will be seen in graphical form in a following figure. As such, the concept motivating the design of our hybrid human-computer system was to elicit input from experts for these cue values, but then to use linear regression to translate these values into dollar value predictions more accurately and reliably than could unaided human experts can do. Both hitters and pitchers were used because in general, pitchers are less consistent from year-to-year than hitters. In this way, we were able to create both low- and high-uncertainty experimental conditions, represented by hitters and pitchers, respectively.

To evaluate the resulting system and interface design, we compared a linear-additive and regression bias visualization aid (LA+RB) to a no-aid (NA) condition as well as a simple linear-additive aid (LA) condition consisting solely of the linear-additive aid with no graphical support. The no-aid condition provided no statistical or graphical support. The LA

condition provided participants with an interface that broke down the judgment into five cue judgments. Participants were able to use their expertise to modify or “tweak” the cue values associated with these judgments, and then the linear regression model combined and weighted these cues into a summary prediction of player dollar value. The full-aid condition combined the task decomposition and linear regression aid present in the partial condition with a set of graphical aids and visualizations designed to help participants adapt their modifications or tweaks of the cue values to the validities and standard deviations of their respective criterion distributions. This is required to balance weight given to case-specific and base-rate information in uncertain judgments.

Experimental Design. The experiment was a 3x2 between-within-subjects design. The three between-subjects display conditions were no aid (NA), linear-additive aid (LA), and linear-additive and regression bias aid (LA+RB). These will be described in a following section. Thirty-six subjects were randomly assigned to one of the three display conditions. The within-subjects variable was the low versus high-uncertainty manipulation associated with hitter versus pitcher stimulus items, or cases.

Participants. Thirty-six paid fantasy baseball experts from the University of Illinois at Urbana-Champaign participated in the study. Twelve participants were randomly assigned to each of the three display conditions. To qualify as an expert, participants were required to be in a fantasy baseball league this year as well as at least one additional year of the past two. The experiment took approximately one hour to complete. Each participant was paid \$10 for his or her time.

Stimulus Items. A representative sample of 84 baseball players was chosen based on several criteria. First, players were only from the National League (NL), since the teams closest to our area are NL teams (e.g. Chicago, St. Louis). All players had positive dollar values for both the 2005 and 2006 seasons. Since more hitters than pitchers fit these criteria, we selected a representative group of hitters to ensure equal numbers of both player types based on 2006 dollar values.

No-Aid (NA): For the no-aid condition, participants were presented with the final dollar values for each player for the past two years (2005 and 2006). Participants were asked to simply type in the predicted dollar value for the end of the current season (2007). Participants were instructed to make predictions on the basis of not only the prior dollar values displayed, but all relevant information and experience the participant brought to the task after recently selecting his or her own actual fantasy team or teams for the 2007 season.

Linear-Additive Aid (LA): For the LA condition, participants were shown each player’s performance in each of five predictive categories (with associated cues) for the past two years (2005 and 2006). Participants were asked to predict performance for 2007 in each of the five categories by typing in a number for each category. The final dollar value for each player was then automatically calculated based on a linear-regression model where the optimal beta weights for each cue were multiplied by the predictions in each category. The beta weights were obtained through a linear regression based on the cues and final dollar values from 2006. The design of this aid reflects a bootstrapping approach, where the human may have

more accurate or recent information about the cue values than the aid does, but the aid integrates the resulting cue values into a prediction more consistently and reliably than the unaided human can.

Linear-Additive and Regression Bias Aid (LA+RB): For the LA+RB condition (see Figure 1), participants were provided with the linear regression aid as well as a regression bias aid in the form of a restricted-range slider for adjusting cue values, and visual guidance for appropriately tailoring the distribution of their judgments over the course of the experiment. The same linear regression model that was used in the LA condition was used in this condition to aid judges in predicting final dollar values by weighting and combining the cues.

As in the LA condition, participants were given performance data in each of the five categories for the past two years. In this condition *all* players in the league are represented in a graphical format for *every* judged case. Each player was represented by a blue plus sign (+), while the current player is a bold red dot. The y-axis provides performance for 2005, while the x-axis shows the performance for 2006. The axes were set up this way so our reflective regression bias aid could be visualized in a graphical display as described in later paragraphs.

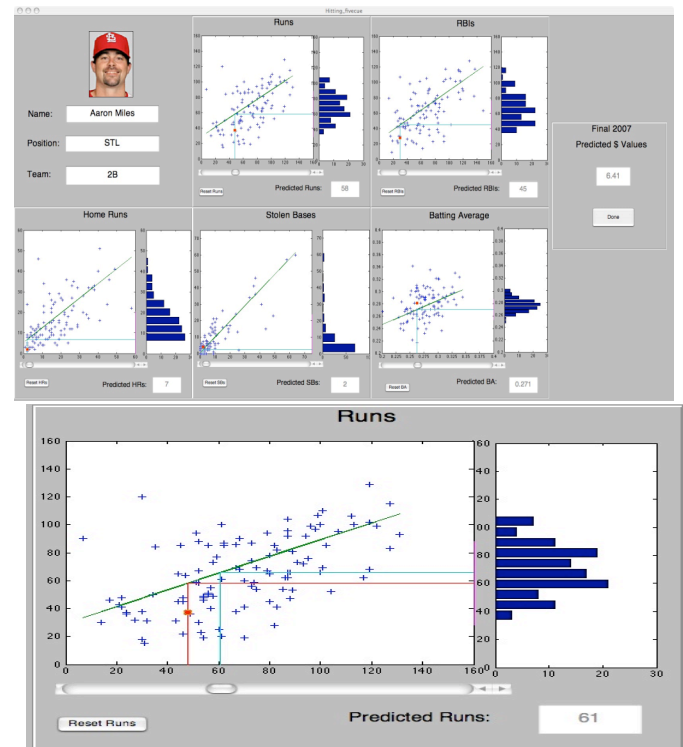


Figure 1. Full Aid Condition (task decomposition, linear additive, graphical visualization). The top figure is the full display. The bottom figure shows a single cue.

The different degrees of scatter (year to year predictability) of the data points across the 5 categories and the green lines shown for each of the categories (Figure 1) are two central aspects of the regression bias aid. The degree of scatter (lack of perfect correlation) was intended to help communicate the difficulty of predicting year-to-year performance, and how this

difficulty varies as a function of category. When this display first appeared for a given player, an initial prediction for each of the 5 categories was presented based solely on performance in the previous year (the light blue lines in Figure 1). The green regression lines were intended to act as visual “reflectors,” designed to aid participants in optimally adjusting, or re-scaling, their adjustments or tweaks of the cue value based on the degree of uncertainty or predictability of each cue as measured historically. Notice that the slope of the regression line differs among cues. The slope of the line determines the degree of re-scaling required to reduce regression bias to zero, and differs across the various cues as a function of their predictive validities.

The aid worked in the following way. First, the participant was presented with a prediction based simply on last year’s performance. They adjusted this prediction by moving the slider along the x-axis. This value was then projected up, reflected off of the green line, and over to the right, y-axis. The value where the reflected line crosses the right y-axis thus reflects an expert prediction of the future value of this cue, but one that is rescaled to appropriately balance expert knowledge and regression to the mean. For cues that are less predictable from one year to the next (e.g. Batting Average: lower left in Figure 1), the slope of the line approaches zero, while the intercept approaches the mean for all players. As such, when the slope is closer to zero, as in the batting average cue, the base rate correction provided by the “reflector” aids performers in making judgments closer to the mean (i.e. the range is restricted). When cues are more predictable from year to year (e.g. Stolen Bases), nearly the full range of values from the x-axis is projected to the y-axis (i.e. range is less restricted). In the current design, our experts’ judgments were strictly limited to the restricted range allowed by the regression aid. This means that the minimum value that can be entered for each cue is the intercept of the y-axis.

One potential limitation of our approach was that our calculations of optimal regression to the mean were based solely on the historical variability of each cue value year to year. As such, the recommended adjustments may be too restrictive if the expert’s adjustments increase the *apparent* predictability of a cue from year to year, and too liberal if the expert’s adjustments decrease this predictability. We use the term “apparent” to highlight that the year-to-year predictability of a cue, such as home runs, is a property of the historical data, and as such is not itself something human expertise can influence. However, it may be the case that experts are able to take into account additional cues, such as injuries, when making a prediction of say, future home runs from past home runs. In such a case, the optimal degree of range restriction may have to reflect the predictive validity of both the displayed cues and these additional cues. Of course, we cannot know prior to the collection of at least a first round of experimental data whether experts’ judgments will be better or worse than historical predictions optimally adjusted for regression to the mean. Results on whether experts’ inputs yield final predictions that are either better or worse than these predictions may require us to readdress this issue in future research.

One final limitation of the current experiment was that there were some unanticipated differences in the final dollar value predictions among display types due, in part, to the restricted range of the cue values in the LA+RB condition. For example, if a player did not play during the season, each of the conditions would produce a different dollar value result for this player. In the no aid condition, our experts would value the player at \$0, since he wasn’t playing. In our LA condition, our experts would enter a cue value of 0 for each of the five individual cues. Based on the computational aid designed to weight and add the cue values, this would result in a dollar value of \$-46. Finally, in the LA+RB condition, our judges would enter the minimum value for each cue based on the restricted range of the regression aid. This would result in a slightly negative dollar value (\$-3). Note the large difference in the predicted dollar values based on the same intended cue or dollar value entry. This design problem will be addressed in the results section.

RESULTS

Performance was analyzed using Murphy’s (1988) skill score. Recall that skill score is a measure of overall judgment quality, which combines correlation, base rate bias, and regression bias. Skill score measures were computed for each of the three aid conditions (NA, LA, and LA+RB). Participants’ performance was compared against “expert” experts’, essentially pre-season dollar value predictions made by ESPN analysts (ESPN). We also compared performance against the computational model without any participant modification. The model was based on the regression bias aid using player performance from 2006 as the input (Aid_no expt).

Before analyzing our results, we dropped the lowest 10% of both hitters and pitchers based on either plate appearances or innings pitched (i.e. 4/42 hitters and 4/42 pitchers were dropped). This was done for two reasons. The first reason was that potential injuries are essentially unpredictable and our participants likely did not take the potential for future injury into account. The second reason, as explained in the previous section, was that there were large, unanticipated, differences in dollar value outputs among the three conditions at the extreme low end of the scale. Since skill score is a measure of Mean Squared Error, a few players with very large errors can represent most of the error. In the current iteration, our design was intended to evaluate a different aspect of the combined model-judge system without taking this type of major source of error into account. Therefore we dropped the players who were injured the majority of the season.

The results for the skill score components for the top 90% of both hitters and pitchers are shown in Figure 2. The skill score was analyzed using a 2 (within: hitters, pitchers) x 3 (between: NA, LA, LA+RB) repeated measures ANOVA. There was a main effect of position (hitters vs. pitchers) ($F_{1,33} = 201.251, p < 0.001$) and an interaction effect of position x aid condition ($F_{2,33} = 17.190, p < 0.001$). A single-factor ANOVA revealed a marginally significant difference among display conditions for hitters ($F_{2,33} = 2.6, p < 0.10$), and a significant difference for pitchers ($F_{2,33} = 21.4, p < 0.01$).

The LA+RB condition was compared with the model (Aid_no expt) and ESPN analysts' performance. Recall that the Aid_no expt condition took the player's 2006 performance as the input and modified the prediction based on the uncertainty. Essentially this was the LA+RB condition without any expert input. To compare performance for the LA+RB condition against the ESPN and Aid_no expt model, we first found the 95% confidence interval for our LA+RB condition and determined if the Aid_no expt or ESPN scores fell within the interval. We did this because ESPN and the model were single scores. Eleven of the twelve participants in the hitting condition outperformed the model and ten of the twelve outperformed the model in the pitching condition. Participants in the LA+RB condition did not outperform the Aid_no expt model for either hitters or pitchers at the 95% confidence interval (Hitters: [0.35, 0.54], Aid_no expt=0.38; Pitchers: [0.08, 0.17], Aid_no expt=0.09). However, the Aid_no expt performed worse than the LA+RB condition for both hitters and pitchers at the 90% confidence interval (hitters [0.39, 0.50]; pitchers [0.10, 0.16]). The skill score for the ESPN experts was not significantly different from our LA+RB condition for hitters (ESPN=0.42), but was significantly worse for pitchers (ESPN= -0.68) at the 95% confidence interval.

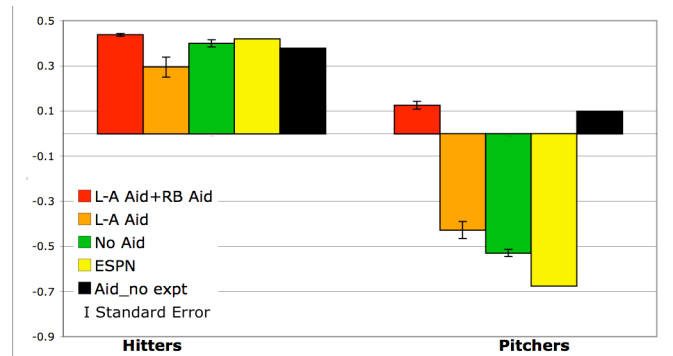


Figure 2. Skill scores for the top 90% of hitters and pitchers based on plate appearances (hitters) and innings pitched (pitchers).

DISCUSSION

The goal of the current research was to develop a joint human-computer system that was designed to integrate human judgment with computational support. This design combined information visualization and formal computational models that were intended to leverage the strengths and simultaneously compensate for the weaknesses of both the human and computational models. Our results showed some success and revealed some areas where our design was limited.

Perhaps most significantly, our LA+RB condition outperformed the Aid_no expt condition with 90% reliability. This means that our combined judge-model system likely outperformed the model alone. Almost no studies have found that a combined judge-model system is able to outperform a linear regression model (but see Yaniv & Hogarth, 1993).

Also, comparing our LA+RB condition against the LA and NA conditions results for the hitters and pitchers, which represented lower- and higher-uncertainty conditions

respectively, showed different patterns. For the lower-uncertainty condition (i.e. hitters), there were no significant differences (at p<0.05) in skill scores among the display conditions, but there were significant differences for pitchers. These results suggest that in conditions of higher uncertainty, our design appears to provide more support than when environments are more certain.

Future iterations in the design will explore how to improve the balance between case-specific input from the judge and base-rate information from the model. The current experiment asked participants to make case-specific judgments for each player. However, the results showed that the model provides good support in most instances. The next experiment is designed to provide meta-cognitive support to judges in highlighting instances where the model is likely to provide good support, versus instances where the model is not equipped to handle the situation at hand.

ACKNOWLEDGEMENTS

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

REFERENCES

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. (2nd ed.). Berkeley: University of California Press.

Camerer, C. F. (1981). General conditions for the success of bootstrapping models. *Organizational Behavior and Human Performances*, 27, 411-422.

Dawes, R. M. (1971). A Case Study of Graduate Admissions: Application of Three Principles of Human Decision Making. *American Psychologist*, 26, 180-188.

Goldberg, L. R. (1968). Simple models of simple processes? Some research on clinical judgments. *American Psychologist*, 23, 483-496.

Horrey, W. J., Wickens, C. D., Strauss, R. A., Kirlik, A., & Stewart, T. R. (2006). Supporting situation assessment through attention guidance and diagnostic aiding: The benefits and costs of display enhancement on judgment skill. In A. Kirlik (Ed.), *Adaptive Perspectives on Human-Technology Interaction* (pp. 55-70). New York: Oxford University Press.

Kirlik, A., & Strauss, R. A. (2006). Situation awareness as judgment I: Statistical modeling and measurement. *International Journal of Industrial Ergonomics*, 36, 463-474.

Meehl, P. E. (1954). *Clinical versus statistical prediction*. University of Minnesota.

Morrow, D. G., Wickens, C. D., & North, R. (2006). Reducing and mitigating human error in medicine. In R. S. Nickerson (Ed.), *Annual Review of Human Factors and Ergonomics* (Vol. 1). Santa Monica, CA: Human Factors and Ergonomics Society.

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.

Strauss, R. A., & Kirlik, A. (2006). Situation awareness as judgment II: Experimental demonstration. *International Journal of Industrial Ergonomics*, 34, 475-484.

Thomas, J. J., & Cook, K. A. (Eds.). (2005). *Illuminating the Path: The Research and Development Agenda for Visual Analytics*.

Weigmann, D., McCarley, J., Kramer, A., & Wickens, C. D. (2006). Age and automation interact to influence performance of a simulated luggage screening task. *Aviation, Space, and Environmental Medicine*, 77, 825-831.

Wickens, C. D., Mavor, A., Parasuraman, R., & McGee, J. (1998). *The future of air traffic control: Human operators and automation*. Washington DC: National Academy Press.

Yaniv, I., & Hogarth, R. M. (1993). Judgmental versus statistical prediction: Information asymmetry and combination rules. *Psychological Science*, 4, 58-62.