Determinants of linear judgment: A meta-analysis of lens model studies[*]

Natalia Karelaia[1] & Robin M. Hogarth[2]

HEC Université de Lausanne[1], Lausanne

ICREA & Universitat Pompeu Fabra[2], Barcelona,

February 1, 2007

1

Abstract

The mathematical representation of Brunswik's lens model has been used extensively to study human judgment and provides a unique opportunity to conduct a meta-analysis of studies that covers roughly five decades. Specifically, we analyze statistics of the "lens model equation" (Tucker, 1964) associated with 259 different task environments obtained from 78 papers. In short, we find – on average – fairly high levels of judgmental achievement and note that people can achieve similar levels of cognitive performance in both noisy and predictable environments. Although overall performance varies little between laboratory and field studies, both differ in terms of components of performance and types of environments (numbers of cues and redundancy). An analysis of learning studies reveals that the most effective form of feedback is information about the task. We also analyze empirically when bootstrapping is more likely to occur. We conclude by indicating shortcomings of the kinds of studies conducted to date, limitations in the lens model methodology, and possibilities for future research.

Keywords: judgment, lens model, linear models, learning, bootstrapping

Since the 1960s, many psychologists have used the framework of Brunswik's (1952) lens model to study processes where humans make predictions of specific criteria (see, e.g., Brehmer & Joyce, 1988; Cooksey, 1996; Hastie & Kameda, 2005). For example, a person might make a judgment (i.e., prediction) about another person's intelligence, about the likelihood of rain, whether a job candidate will be successful, and so on. In all these cases, the simple beauty of Brunswik's model lies in recognizing that both the person's judgment and the actual criterion predicted can be thought of as two separate functions of cues that are available in the environment. Thus, the accuracy of human judgment depends on the extent to which the function that describes it matches its environmental counterpart.

But how good or accurate are people at making judgments and on what does this depend? These are important questions that have generated considerable controversy in the psychological literature (Cohen, 1981; Gigerenzer, 1996; Kahneman & Tversky, 1996). Whereas it is unlikely that these questions can be answered satisfactorily by any particular approach, an advantage of research conducted within the Brunswikian tradition is the use of a common methodology for formalizing the lens model. Thus, not only can researchers within this tradition communicate results within a common framework, it is possible to aggregate results quantitatively across many studies and make statements that reflect the accumulation of results. This is the purpose of the current paper in which we present a meta-analysis of studies conducted using the lens model over a period of five decades.

The paper is organized as follows. We first describe the mathematical formulation of the lens model. Second, we specify how we identified and included particular studies

in our analysis. Third, by summarizing the results of these studies we illuminate the issue of how accurate human judgment is and the factors that affect it. Since psychologists typically study judgment within laboratories but people use judgment outside laboratories, we pay particular attention to differences between laboratory and field studies. Fourth, since the topic of learning has been central to studies within the lens model tradition, we make a separate analysis of learning studies. Key topics center on how much learning occurs, what affects this and the impact of different types of feedback. Fifth, we contribute to the discussion of the relative advantages of clinical judgments and their paramorphic representations (Hoffman, 1960) or bootstrapping models (e.g., Goldberg, 1970; Dawes, 1971; Camerer, 1981) by analyzing the conditions under which people are more likely to be outperformed by models of their judgments. Finally we conclude by summarizing the main substantive conclusions of the analysis, indicating shortcomings of the kinds of studies conducted to date, and suggesting avenues for future research.

The mathematical formulation of Brunswik's lens model

The use of Brunswik's lens model received an important impetus in 1964 when a series of papers showed how statistical methods could be used to capture judgmental processes (Hammond, Hursch, & Todd, 1964; Hursch, Hammond, & Hursch, 1964; Tucker, 1964. See also Castellan, 1973). In this, human judgment, denoted Y_s, is modeled as a linear function of a set of k cues, X_j, $j = 1,...k$. Thus,

$$Y_s = \sum_{j=1}^{k} \beta_{s,j} X_j + \varepsilon_s \qquad (1)$$

where the $\beta_{s,j}$'s represent the weights that the person (or judge) gives to the different cues

and ε_s is the error term of the regression of Y_s on the X_j's.

Similarly, the environmental criterion, Y_e, can be modeled as a function of the

same cues, X_j, $j = 1,...k$. That is,

$$Y_e = \sum_{j=1}^{k} \beta_{e,j} X_j + \varepsilon_e \qquad (2)$$

where the $\beta_{e,j}$'s represent the weights that the environment gives to the different cues and

ε_e is the error term of the regression of Y_e on the X_j's – see Figure 1.

Insert Figure 1 about here

The logic of the lens model is that the person's decisions will match the

environmental criterion to the extent that the weights the judge gives to the cues match

those used by the model of the environment, i.e., the matches between $\beta_{s,j}$ and $\beta_{e,j}$ for all j

$= 1,...k$. Moreover, the correlation between criterion and judgment, $\rho_{Y_e Y_s}$ – the so-called

"achievement" index or r_a – can be expressed by the "lens model equation"

$$r_a = G R_e R_s + C \sqrt{\left(1 - R_e^2\right)\left(1 - R_s^2\right)} \qquad (3)$$

where $G = \rho_{\hat{Y}_e \hat{Y}_s}$ (the "matching" index) is the correlation between the predictions of both

models, i.e., between $\sum_{j=1}^{k} \beta_{e,j} X_j$ and $\sum_{j=1}^{k} \beta_{s,j} X_j$; R_e and R_s are, respectively, the multiple

correlations of the models of the environment and the judge, and capture, on the one

hand, environmental predictability (R_e), and on the other hand, the consistency with

which the judge executes the decision rule (R_s); and $C = \rho_{\varepsilon_e \varepsilon_s}$ is the correlation between

5

the error terms of the two models. If these are independent, i.e., $\rho_{\varepsilon_e \varepsilon_s} = 0$, then judgmental accuracy or achievement (r_a) is simply a multiplicative function of three terms: matching (G), environmental predictability (R_e), and response consistency (R_s), and neatly captures the effects of both cognitive and task variables on observed performance. In practice, C may actually differ from 0 if, say, a variable has been omitted from the analysis and/or cues are used in a nonlinear manner.

In addition to the basic lens model statistics indicated above, we are interested in the products of two of these statistics. First, it is illuminating to analyze the human component of achievement independently of the predictability of the environment. For situations where $C = 0$, this can be represented by the product of matching, G, and response consistency, R_s. This product, GR_s, named "performance" by Lindell (1976) and "linear cognitive ability" by Hogarth and Karelaia (2006), neatly captures the extent to which judges both match task requirements and are consistent in the execution of their strategies.

Second, the product of matching, G, and environmental predictability, R_e, is an estimate of the validity of the bootstrapping model of the judge (Goldberg, 1970; Dawes, 1971; Camerer, 1981). This product, GR_e, is interesting in that it captures the validity of the judge's strategy assuming that the strategy is applied in a perfectly consistent manner (i.e., when $R_s = 1$). In other words, it captures what would happen if a judge was replaced by his or her model.

Database for the meta-analysis

By searching several databases, key articles, and consulting leading contributors to the literature, we identified some 200 published and unpublished works that suggested that they might contain lens model data, specifically the components of Equation 3. [1] We excluded from consideration works that failed to model the environmental side of the lens (i.e., for which criterion data were missing), studies that used aggregate as opposed to individual judgments (see, e.g., Gifford, 1994) as well as research within the conflict resolution paradigm in which the criterion for one person is the judgments of others (see, e.g., Hammond, Wilkins, & Todd, 1966). From these 200 works we identified 78 that contained full (or almost full) lens model statistics. These works were published between 1954 and 2006, one half being published before 1984, and the other half afterwards.[2] Most of these studies examined judgments in more than one environment or experimental setting. Thus, we ended up with a total of 259 different environments in which judgments were made. The mean number of participants in the 259 environments was 19 (inter-quartile range, 10 to 24), each making, on average, 88 judgments (inter-quartile range, 25 to 91). The total number of individual judgments on which our results are based is thus large – almost 320,000.[3]

We characterized each of the 259 data points by the averages of the lens model statistics of the participants in each of these environments. These averages were either

---

[1] It is important to note that, on completing our analysis, we became aware of another recent meta-analysis of lens model studies conducted by Kaufmann and Athanasou (2007). The scope of their work, however, is more limited than ours and their criteria for including studies in the analysis are different. Their work should thus be considered complementary to what is presented here.

[2] We note, incidentally, that when splitting the 1954 - 2006 period into five-year periods, the 1972-1976 period contains the largest number of published papers, 21. Interest in the topic then declined, as judged by the number of papers we identified, but increased again at the end of the 1990s.

[3] In fact, the total is somewhat larger because this figure excludes all but the last block of learning trials.

taken directly from the papers, inferred (e.g., from graphs), or calculated by making use of the properties of the lens model equation (Equation 3). In many experimental papers, no value was given for C (the correlation between the residuals of the models of the judge and environment), and this was assumed to be 0 (a reasonable assumption given the way in which experimental data were generated). We emphasize that our unit of analysis is the average of statistics of individuals within each environment as opposed to the actual individual statistics. Unfortunately, only a few papers provided individual level data and so we are unable to comment on variation within the different environments.

In addition to the lens model statistics for all 259 observations, we encoded variables that characterized both the specific tasks and participants (these are described below and in Table 2). When studies explicitly considered learning over several blocks of trials, we limited our attention to statistics for the first and last blocks. The latter were used to capture general performance and aggregated with the non-learning data. The former were used as a baseline to capture the effects of learning relative to levels exhibited in the last blocks of trials (see below).

How accurate is human judgment overall?

The upper part of Table 1 reports mean values of the lens model statistics for the data we examined. Note that, with the exception of environmental predictability, R_e, our data points are themselves means, and that there is considerable variability in that almost all indices vary between their theoretically possible minima and maxima. Across all 259 observations, mean achievement, r_a, is 0.55, mean matching, G, is 0.81 (the median of G is notably higher, 0.91), and mean response consistency, R_s, reaches 0.80. On the

environmental side, predictability, R_e, is 0.79 on average, and mean non-linear component, C, equals 0.05.

It is interesting to note that mean R_e and mean R_s are close in value as this suggests a kind of overall (or mean) probability matching phenomenon. As just noted, mean R_e is a "true mean" whereas mean R_s is a "mean of means." Thus, whereas the reported standard deviation of R_s is smaller than that of R_e, this hides the fact that in any given study values of R_s at the individual level can vary quite a lot for fixed levels of R_e.

Insert Table 1 about here

As for the two composite statistics, mean linear cognitive ability, GR_s, is 0.66, with its median slightly higher at 0.73. The mean validity of bootstrapping models, GR_e, is 0.64 and surpasses mean achievement of clinical judgment, 0.55.

To explore relations between the various indices, the lower part of Table 1 presents pair-wise correlations. Several significant correlations ($p < 0.01$) come as no surprise.[4] In particular, consistent with Equation 3, there are high positive correlations between achievement, r_a, and (a) matching, G, 0.77; (b) response consistency, R_s, 0.52; and (c) environmental predictability, R_e, 0.40. Less obvious *a priori* is the significant correlation between the two statistics that characterize performance independent of environmental predictability, namely matching, G, and response consistency, R_s. This correlation is positive, 0.42, and suggests that decision makers who match the environment better are also more consistent in executing their judgment.

---

[4] For statistical purposes, we conduct our analysis by treating each of our 259 observations as though these are random drawings from an underlying population of environments. This is patently a false assumption. We therefore use statistical tests in the spirit of heuristic guides as opposed to "hypothesis tests."

Another striking result is that neither consistency, R_s, nor matching, G, correlates with environmental predictability, R_e (correlations of 0.08 and -0.01, respectively). That is, decision makers can reach similar levels of performance/linear cognitive ability in both noisy and predictable environments.

Finally, the fact that r_a and C, the non-linear component, are moderately correlated (0.28) suggests significant non-linear usage of cues and/or omitted variables (i.e., judges were using information of which investigators were unaware).

What factors affect the accuracy of human judgment?

For all 259 environments, we encoded (when available) seven variables that could potentially affect the level of human judgment. Six involve characteristics of environments or tasks and one of decision makers. These variables are also listed on the left hand side of Table 2.

Insert Table 2 about here

Environments vary, first, in the number of cues. This can be taken as a surrogate measure of task complexity (i.e., given limited information processing capacity). We code the data into three groups: two, three, or more than three cues.

Second, we distinguish between environments where cues are "given" as opposed to "achieved." For the former, decision makers are provided with the explicit values of the cues by the experimenter. For the latter, the values of the cues need to be inferred – and often even identified – by decision makers.

Third, we classify environments by the level of inter-cue redundancy as either "none" (no redundancy), "some" (if average cue inter-correlation is less than 0.4, or redundancy is described as being: "low," "moderate," "some"), or "high" (otherwise).

Fourth, we classify the distributions of the environmental weights $\beta_{e,j}$'s given to cues in three groups. In particular, we define weighting functions as non-compensatory if, when cue weights are ordered in magnitude, the weight of each cue exceeds the sum of those smaller than it (Martignon & Hoffrage, 1999; 2002). All other functions are classified as compensatory except for the special case of equal-weighting.

Fifth, we consider differences between laboratory experiments and field studies.

Sixth, we distinguish between environments in which participants were explicitly given the possibility to learn over several blocks of trials and environments without this possibility. We label the latter "stable" environments. For "learning" environments, we also record the number of learning trials and type of feedback given to participants. We classify feedback into five categories: none, outcome feedback, cognitive feedback, task information feedback, and other types of feedback. Outcome feedback is simply knowledge of the outcome of a judgment. Cognitive (or process) feedback refers to data involving the judge's decision policies (e.g., $\beta_{s,j}$'s, the weights given to the different cues). Task information feedback is information about true relations in the environment (e.g., $\beta_{e,j}$'s) rather than relations perceived by the judge (i.e., cognitive feedback).

Seventh, it is reasonable to assume that initial level of expertise may be important for achievement. We therefore classify participants in three groups: "novices," "experts" (two extreme categories), and "some training" (an intermediate category).

In what follows, we analyze the role that each of these variables plays in determining levels of human achievement and performance. The data described below are presented in Table 2. In particular, Table 2 classifies the data according to the variables enumerated above and specifies the numbers of environments falling into each category, average numbers of judges in the environments, average numbers of judgments made by each judge, and means of lens model indices. In what follows, we use the 0.05 significance level in statistical tests, unless indicated otherwise.

Number of cues. Given well-established limitations on human information processing, it is often argued that the linear model does not provide a good description of judgment when the number of cues is large (cf., Payne, Bettman, & Johnson, 1993). At the same time, when decision makers have many cues available, redundancies might increase achievement. What do the data say about these issues?

There are two significant effects of the number of cues in the task. First, in these data, the non-linear component, C, is smaller in the environments with three cues (0.00 vs. 0.07 and 0.08 for the two-cue and more-than-three-cues environments, respectively). Second, mean matching, G, is much smaller in environments with more cues. That is, with more than three cues, an average judge has a matching index of 0.71, while for the environments with fewer cues the analogous index is 0.88. Given that response consistency, R_s, does not correlate with the number of cues, the final impact of this task variable on the decision maker's performance, GR_s, is similar to the effect described above on G. In particular, in environments with more than three cues, participants had, on average, a lower level of GR_s, 0.58 (vs. an average of 0.72 for all other environments).

As for human achievement, r_a, it also reaches its lowest level in the environments with more than three cues, 0.51, which is significantly different from the 0.63 corresponding to the environments with two cues. Importantly, the two-cue and more-than-three-cue environments have similar levels of all lens model indices except for matching, G, and, therefore, the difference in human achievement, r_a, can be attributed to the difference in the levels of G. This picture changes, however, if the environments with three cues are used as a reference point in interpreting the figures in the environments with more than three cues. In this comparison, the effect of G (0.88 vs. 0.71) on achievement, r_a, is masked by the differences observed in the levels of the non-linear component, C (0.00 vs. 0.08). As a result, the difference between achievement levels in the environments with three cues (0.55) and more than three cues (0.51) is not significant.

Overall, we find that judges match the environment worse when there are more than three cues and that human achievement is negatively affected by a large number of cues (see also Einhorn, 1971). Below, we shall address further the issue of the relation between the number of cues and inter-cue redundancy.

Given/achieved cues. In addition to weighting and combining information, an important dimension of many judgmental tasks involves identifying and assessing levels of relevant information (Einhorn, 1972). In these data, we find that whether cues were directly given to participants or had to be achieved affects neither matching, G (0.82 vs. 0.79, respectively; the difference is not significant) nor response consistency, R_s (0.81 vs. 0.79, respectively; the difference is not significant). As a consequence, the difference between the levels of human achievement, r_a, in the studies where cues are achieved by judges (0.60) and in the studies where the cues are provided directly by experimenters

(0.55) is not statistically significant, even though the non-linear component, C, is much larger in the environments with achieved as opposed to given cues (0.13 vs. 0.03, respectively). The difference in C suggests that when cues had to be "achieved," judges were relying on information that investigators did not include in the corresponding models of environments.

For further insight, we analyzed the effect of this variable separately in environments with different numbers of cues (see previous section). While we do not have a sufficient number of environments with achieved cues within the groups of two- and three-cue environments, we observed a significant effect of this task variable within the environments with more than three cues. In particular, achievement, r_a, is greater when cues are achieved (0.62, n = 27) than when these are given (0.46, n = 67). Interestingly, within the environments with more than three cues, we find additionally significant effects of this variable on both matching, G, and consistency, R_s. In particular, judges match the environment better when cues are achieved (0.80, n = 27 vs. 0.67 in 67 environments with given cues), but show greater judgmental consistency when cues are given (0.83, n = 67 vs. 0.75 in 27 environments with achieved cues).

Overall, we find no evidence in the data that either matching or response consistency is affected by the way judges obtain the cue values. However, in a subset of data – environments with more than three cues – we find that judges match environmental weights better but are less consistent in applying their models when they need to infer cue values as opposed to when these are directly provided by experimenters.

Inter-cue redundancy. Inter-cue redundancy is a functional element of decision environments in that it facilitates what Brunswik (1943; 1952) referred to as "vicarious

functioning" or the interchangeability of cues. It thereby contributes to improving the reliability of overall judgments and can help to limit information search without significant reductions in judgmental accuracy (Connolly & Miklausich, 1978; Einhorn, Kleinmuntz, & Kleinmuntz, 1979).

In these data, however, we do not observe any beneficial effect of inter-cue redundancy on human achievement, r_a. Indeed, we find an inverse relation: mean achievement, r_a, is significantly higher when there is no redundancy (0.61) than when there is "some" (0.53) or it is "high" (0.54). A similar tendency is observed with respect to the average level of matching, G (0.89 vs. 0.78 and 0.76, respectively). Interestingly, the non-linear component, C, is higher in environments with high redundancy (0.10 vs. 0.03 in all other environments).

Although surprising at first, the effect of redundancy may be due (at least partially) to a positive correlation of this task variable with the number of cues, 0.33 (n = 208, p < 0.001). In particular, only 15% of environments with more than three cues contain no inter-cue redundancy which is well below the 56% of environments with two cues, and 80% of environments with three cues. (In addition, among the environments with three cues, none is classified as containing high redundancy.)

To distinguish the effects of inter-cue redundancy and the number of cues on human judgment, we regressed (separately) achievement, r_a, matching, G, and response consistency, R_s, on the number of cues and the level of cue redundancy. To control for the linear predictability of environments, we included two more predictors: the non-linear component, C, and environmental predictability, R_e. We used a robust weighted least squares (WLS) regression procedure. The results of these regressions showed no

significant effect of cue redundancy on ether achievement, r_a, or response consistency, R_s. However, when explaining the variance of matching G, the regression coefficient for redundancy is negative and significant (robust t = -2.65, p = 0.009).[5] That is, in these data, subjects matched environmental weights better when cue redundancy was smaller, controlling for the number of cues and linear environmental predictability.

We conclude therefore that the data do not show any positive effect of cue redundancy on human judgment and even suggests that matching, $G,$ suffers from higher redundancy.

Weighting function form. An important topic of lens model research has focused on how well experimental participants handle different types of functional relations between cues and the criterion (see, e.g., Brehmer, 1980). The data reveal a significant effect of how cues are weighted by the environment. In particular, of the three types of functions we consider, the lowest level of mean human achievement, r_a, (0.51) corresponds to environments where this is non-compensatory. Mean achievement is highest (0.66) in equal-weighting environments whereas the mean for compensatory environments lies between these extremes (0.57).

We note that two variables seem to explain this effect. First, in non-compensatory environments, participants did the worst job in matching the environmental weights (mean G is 0.80 vs. 0.84 in compensatory environments and 0.91 in environments with equal weights). Second, mean response consistency, R_s, is also lowest in the non-compensatory environments: 0.74 vs. 0.82 in all other environments.

---

[5] A robust OLS procedure with While-adjusted standard errors gives similar results, except that the value of t-statistics of the coefficient for redundancy decreases to -1.81 (p=0.072).

The joint effect of these two variables magnifies the difference in performance/linear cognitive ability, GR_s, between the different types of environments. According to the data, the most "difficult" environments involve non-compensatory weighting schemes (mean $GR_s = 0.62$), followed by compensatory environments (mean $GR_s = 0.70$). The equal-weighting environments are best suited to applying a linear model (mean $GR_s = 0.75$).

Laboratory and field studies. An important dimension of the Brunswikian research philosophy centers on the concept of *representative design* (Brunswik, 1956). While field studies are naturally *representative*, laboratory studies may or may not reflect formal properties of naturalistic environments. Thus, it is of interest to compare laboratory and field studies on two dimensions. First, do the conditions of laboratory studies mirror those of field studies? Second, do participants have differential achievement and performance in the two kinds of environments?

In field studies (n = 48), mean matching, G, was lower than in laboratory studies (n = 208), 0.74 vs. 0.83 (the difference is significant). There are no other differences between field and laboratory studies in the data. The overall effect of this task variable on performance/linear cognitive ability, GR_s, and achievement, r_a, is negligible.

For further insight, we analyze the effect of this task variable controlling for the number of cues in the environments. In fact, all field studies contained more than three cues, while only 50 of the 208 laboratory studies (24%) did. We therefore limit further analysis to the environments with more than three cues. In doing so, we find that the difference in consistency, R_s, becomes significant ("field" being more consistent, on

average, than "laboratory", 0.84 vs. 0.78), but the difference in matching, G, becomes insignificant. The overall effect on achievement, r_a, remains negligible.

Since the number of cues affects inter-cue redundancy, we split the data along two dimensions: laboratory/field and the level of redundancy (Table 3). We find that the field studies better reflect naturalistic decision environments in that they contain, on average, more redundancy than laboratory studies. In particular, none of the field studies lacked redundancy, about half contained some, and the other half a lot. The majority of laboratory studies (62%) had no redundancy. In the environments with high redundancy, the difference in mean human achievement, r_a, between laboratory and field studies is striking: 0.36 vs. 0.61 (the difference is significant, though the validity of this finding is weakened by the low number of observations, 7, in one group).

Insert Table 3 about here

Does environmental predictability, R_e, have the same importance for human achievement in laboratory and field studies? We analyze the correlation between R_e and other components of human achievement in laboratory and field studies separately and find that, in laboratory studies, environmental predictability, R_e, correlates with neither matching, G, nor response consistency, R_s. In the more naturalistic field studies, however, we find that when environments are more predictable, participants are more consistent ($\text{corr}(R_e; R_s) = 0.52$, n = 45, p < 0.001). Matching, G, however, is unaffected by environmental predictability, R_e (within field studies, $\text{corr}(R_e; G) = 0.22$, n = 45, ns.). As a result, in field studies, greater environmental predictability implies greater linear

cognitive ability or performance, that is, the correlation between R_e and GR_s is 0.35 (n = 45, p = 0.020).

Parenthetically, we note that some thirty years ago, Brehmer (1976) showed – in a small sample – that R_e and R_s were positively correlated in field studies and claimed that this relation was also observed in laboratory studies. Our larger sample of evidence does not support the latter claim.

Expertise. It is possible to point to individual studies of judgmental achievement involving acknowledged experts that indicate both abysmal (Einhorn, 1972) and incredibly accurate performance (Stewart, Roebber, & Bosart, 1997). However, what are the general trends?

In these data, the initial level of expertise (as opposed to expertise acquired through learning in the experimental studies) does not affect the level of achievement, r_a (all relevant differences in Table 2 are not significant). However, for matching, G, and consistency, R_s, we find some unexpected trends. In particular, participants classified as "experts" have the lowest level of matching, G (0.68, significantly different from 0.83, the mean G among "novices"), and are not more consistent than their less experienced colleagues (the mean R_s's for experts and novices, 0.83 and 0.79, are not significantly different).

Note, however, that environments that involved experts had, on average, a greater non-linear component, C (0.17 vs. 0.03 in environments that involved novices), and a lower level of linear environmental predictability, R_e (0.74 vs. 0.81 in environments that involved novices). To separate the effects of the task characteristics from the effect of expertise, we regressed (separately) achievement, r_a, matching, G, and response

consistency, R_s, on the non-linear component, C, environmental predictability, R_e, and the variable characterizing the level of expertise. We used a robust WLS regression procedure. The results of these regressions showed no significant effect of expertise on ether achievement, r_a, or matching, G. However, when explaining the variance of response consistency, R_s, the regression coefficient for expertise is positive and significant (robust t = 2.59, p = 0.010). That is, controlling for linear environmental predictability and the non-linear component of environments, we find that the experts were more consistent than novices in applying their decision policies (cf., quasi-rationality hypothesis in Brehmer, 1994).

We conclude that the initial level of expertise does not affect how well individuals match environmental structures but is important for consistency in applying individual policies. The effect of the initial level of expertise on judgmental consistency is, however, insufficient in these data to impact the overall level of human achievement. We next examine the effect of expertise acquired through learning during the experimental trials.

How effective is learning? The role of feedback

We first compare the results of environments that involved learning trials and those in which participants did not have the possibility to learn the task (i.e., "stable" environments). We next focus only on environments involving learning to uncover additional factors (such as different types of feedback) that contribute to acquiring expertise.

In these data, learning as opposed to stable environments are more predictable (mean R_e of 0.82 vs. 0.72, Table 2) and involve, on average, less of the non-linear

component, C (0.03 vs. 0.09). As for human performance, both mean matching, G, and mean response consistency, $R_{s,}$, are larger in learning environments (0.85 vs. 0.70, and 0.82 vs. 0.77, respectively), in line with the values of the environmental parameters R_e and C. As a result, in learning as compared to stable environments, average cognitive linear ability, GR_s, is also larger (0.71 vs. 0.54), as well as average achievement, r_a (0.60 vs. 0.44).

To understand whether it is environmental predictability, R_e, or the possibility to learn the task that accounts for better matching and consistency in learning environments, we regressed (separately) matching, G, consistency, R_s, cognitive linear ability, GR_s, and achievement, r_a, on environmental predictability, R_e, non-linear component, C, and the two-level dummy variable, *learning/stable*. We used a robust WLS regression procedure. All four regressions show a significant positive effect of the possibility to acquire expertise. That is, when controlling for C and R_e, learning improves matching, G (t(238) = 3.55, p < 0.001), consistency, R_s (t(239) = 2.11, p = 0.036), cognitive linear ability, GR_s (t(233) = 5.12, p < 0.001), and achievement, r_a (t(241) = 5.45, p < 0.001). Interestingly, the three predictor variables (R_e, C, and the dummy *learning/stable*) jointly explain about 48% (adjusted R^2) of the variance of human achievement.

Insert Table 4 about here

Since the presence of redundancy is important for human performance (Brunswik, 1952), we next examine the effect of learning in environments with different levels of redundancy (Table 4). In the group of environments with high inter-cue redundancy, learning has a large effect on linear cognitive ability, GR_s. In particular, mean GR_s is only

0.46 in stable environments while it reaches 0.83 in learning environments (the difference is significant). The effect of learning is less pronounced, but still notable, when cue redundancy is lower. Here mean GR_s is 0.58 in stable environments and 0.69 in learning environments (the difference is also significant). When all cues are independent, however, there is no effect of learning. This finding emphasizes the importance of redundancy not only for human performance but also for better learning and the acquisition of expertise.

In these data, field studies contained, on average, more redundancy than laboratory studies. Therefore, we next analyze whether the effect of learning is more pronounced in field than laboratory studies. First, within stable environments there is no difference in mean GR_s between laboratory and field studies (0.55 vs. 0.52, the difference is not significant). Within learning environments, however, mean GR_s is larger for field than laboratory studies (0.81 vs. 0.70, t = - 2.18, p < 0.05). That is, positive effects of learning are present in both laboratory and field studies, but more so in field studies. This finding suggests that field studies provide better conditions for learning than laboratory studies, possibly due to larger levels of inter-cue redundancy in the field environments.

We next limit our analysis to environments that involved learning to understand better what factors increase the magnitude of learning effects. We analyze the changes in linear cognitive ability, GR_s, and achievement, r_a, that occurred due to learning (i.e., changes between the first and last blocks of trials). From 186 learning environments, 163 contain both pre-learning (i.e., the first block) and post-learning (i.e., the last block) lens model statistics. On average, there were 103 learning trials in the environments. The mean improvement of linear cognitive ability, GR_s, through these trials was 0.20 (39%

more than the pre-learning level), and the mean improvement of achievement, r_a, was 0.18 (44% more than the pre-learning level) – see Table 5.

Insert Tables 5 and 6 about here

To identify task and individual characteristics that affect the magnitude of learning we regressed the changes in linear cognitive ability, GR_s, and achievement, r_a, (separately) on the following variables: number of cues, type of cues (achieved or given), cue redundancy, weighting function form, type of study (laboratory or field), expertise, number of learning trials, and four dummies representing the presence of outcome feedback, cognitive feedback, task information, and other types of feedback. In addition, we controlled for the pre-learning levels of human performance since these can limit the space for learning. That is, when explaining the changes in GR_s, an additional explanatory variable was the initial level of GR_s; in the regression of the changes in r_a, this was the initial level of r_a.

We are especially interested in the effect of different kinds of feedback on learning. Previous literature has shown that outcome feedback is helpful in simple (e.g., two cue) tasks (Doherty, Tweney, O'Connor, & Walker, 1988), but not in complex, uncertain tasks (Brehmer, 1980; Hoffman, Earle & Slovic, 1981). Outcome feedback may even deter learning under uncertainty (Hammond, Summers, & Deane, 1973). Regarding other types of feedback, people learn more from task information feedback than from cognitive feedback (Balzer, Doherty, & O'Connor, 1989). Moreover, when combined with cognitive or outcome feedback, task information feedback is more effective (Reilly & Doherty, 1992; Balzer, Sulsky, Hammer, & Sumner, 1992). However, sometimes

providing task information only may be sufficient (Reilly & Doherty, 1992; Remus, O'Connor, & Griggs, 1996).

The results of the regressions are presented in Table 6. We used both an OLS procedure with White-adjusted standard errors (with and without intercept, Models 1A and 1B) and a robust WLS procedure (Model 2). All regressions show that GR_s increases due to learning more when (1) the initial level of GR_s is lower, and (2) task information is available to judges. The effects of initial expertise (negative), number of cues (positive), and type of cues (more learning when cues are achieved) are less robust but significant.

Indeed, when task information was given, GR_s increased by, on average, 0.35, while mean improvement was 0.28 when no feedback was available (Table 5). When cues were achieved, mean GR_s improvement was 0.47, much larger than the analogous learning effect in the environments with given cues (0.17). The magnitude of the difference between conditions should, however, be interpreted with caution given the differences in the initial values (see the last two columns of Table 5). When task information was given, GR_s increased by, on average, 0.35, while mean improvement was 0.28 when no feedback was available.

As for the improvement in achievement, r_a, all three regressions indicate that it is larger when the initial level of r_a is lower. Positive effects of the number of learning trials, number of cues, achieved (vs. given) cues, and greater cue redundancy are also significant, although not robust, similar to a negative effect of initial expertise.

We do not find any evidence that the availability of outcome or cognitive feedback improves human performance. Given the regression results described above, the differences in the levels of performance improvement that occurred in the presence of

outcome or cognitive feedback and without any feedback (Table 5) can be totally attributed to the differences in pre-learning levels of performance and other task characteristics.

Interestingly, within learning environments, we found a high positive correlation between pre-learning and post-learning levels of consistency (R_s) of 0.74 (n = 130, p < 0.001). That is, participants with better initial performance kept their advantage after the learning trials.

To summarize, we find that, first, when the possibility to acquire expertise is available, individuals reach better levels of matching, G, response consistency, R_s, linear cognitive ability, GR_s, and achievement, r_a. Second, the effect of learning is questionable in the environments with low cue redundancy. Third, positive effects of learning are especially notable in naturalistic environments, such as field studies, as opposed to laboratory experiments. Fourth, the availability of task information magnifies the effect of learning. Neither outcome nor cognitive feedback helps to learn. Fifth, individuals learn more when they have to infer cue values from the context rather than when cue values are explicitly provided. And finally, in these data, there is some evidence that more learning trials imply greater improvements in performance.

What factors affect the accuracy of bootstrapping models?

The validity of bootstrapping models can only really be tested on out-of-sample cross-validation. However, it can be instructive to analyze the potential sizes of effects due to different variables based on past samples of data. We therefore next analyze the

data to isolate the conditions under which the application of bootstrapping would seem differentially advantageous.

We define the advantage of bootstrapping models as the difference between linear achievement with perfect consistency (i.e., GR_e) and human achievement (i.e., r_a). We report this measure in the last column of Table 2. As might be expected, the effect of eliminating judgmental inconsistency more than outweighs any advantages of the non-linear component, C, and bootstrapping is always more effective (cf., Goldberg, 1970; Camerer, 1981). (All entries in the last column of Table 2 are positive.) What is more interesting, however, is to identify the task and judge characteristics that potentially favor bootstrapping models.

To do this, we regress the bootstrapping advantage, $(GR_e - r_a)$, on various task and judge characteristics, using a robust WLS procedure to account for outliers. We report significant regression coefficients in Table 7. Two regression models were examined: the first contained R_e and R_s among the predictor variables; the second included instead the difference between these terms, i.e., $(R_e - R_s)$. We find that the advantage of bootstrapping is larger when: (1) cues are given (vs. achieved); (2) redundancy is lower (vs. higher); (3) judges initially have less expertise; and (4) judges do not have the possibility of acquiring additional expertise through learning. In addition, there is some evidence that the advantage of bootstrapping over clinical judgment is larger in laboratory as opposed to field studies.

A possible explanation of the effect for the type of cues is that inferring (i.e., achieving) cue values may itself be the key to high performance levels regardless of the accuracy of the subsequent processes of matching and executing (consistently or not) a

particular decision strategy. The role that cue redundancy "vicariously" (Brunswik, 1952) plays in human judgment (e.g., interchangeability of the cues) could explain why bootstrapping models lose their advantage when redundancy increases. In what concerns experience, bootstrapping may be less advantageous than clinical judgment when judges possess more expertise (either initial or acquired) precisely because expertise may allow judges to integrate in their judgments non-linear elements that cannot be captured otherwise by a model of the judge. Finally, the advantage of bootstrapping over clinical judgment in laboratory as opposed to field studies is probably a consequence of the fact that, in the former, the non-linear component, C, was often constrained to 0.

Insert Table 7 about here

Regarding the parameters of the lens model, the advantage of bootstrapping models is larger when environments are more predictable (positive coefficient of R_e), judges are less consistent (negative coefficient of R_s), and when environments are more predictable than judges (negative coefficient of $(R_e - R_s)$). These findings correspond to what was found in earlier studies of the effectiveness of bootstrapping (e.g., Camerer, 1981). Finally, we find that the advantage of bootstrapping models is larger when the non-linear component, C, is smaller (negative coefficient of C). This, however, comes as no surprise since bootstrapping models are *linear* models of judges.

Discussion

Summary. We consider that the environments we examined have essentially demonstrated three important findings but that these are subject to a number of

limitations (of generalization). In short, the findings are these: (1) People are capable of achieving high levels of judgmental performance. Moreover, we identified several task and judge characteristics that determine how good this performance can be. (2) People learn best from feedback that instructs them about the characteristics of the tasks they face. (3) People are inconsistent in the application of their judgmental rules such that models of their judgments are typically more accurate than they are themselves (i.e., getting rid of inconsistency is generally better than getting rid of idiosyncratic knowledge that is not captured by linear models). Once again, several task and judge parameters delimit the conditions under which bootstrapping is most effective.

We next summarize the factors that, according to our results, affect the accuracy of human judgment. First, when the number of cues is large, judges match environmental models worse and, consequently, the levels of judgmental performance are lower (cf., Payne, Bettman, & Johnson, 1993).

Second, inter-cue redundancy adds more difficulty to matching environmental models. (But, see below our arguments of why the lens model methodology may not be the best way to capture the effects of redundancy on final judgments). We find, however, that the presence of redundancy is crucial for successful learning. The effect of learning is questionable in the environments with low cue redundancy.

Third, controlling for task complexity (as measured by the number of cues), the levels of human achievement are higher when judges infer cue values as opposed to using the values directly provided by experimenters. This occurs because judges match environmental weights better while inferring cue values, even though in these conditions individuals are less consistent in applying their models.

Fourth, human achievement is higher in environments that weigh different cues in a more equal manner. In environments with differentially weighted cues, both the matching and consistency that judges demonstrate are smaller.

Fifth, in field studies, judges are more consistent in applying their decision strategies when environments are more predictable. This is not true for laboratory experiments (cf., Brehmer, 1976).

And finally, controlling for linear predictability and non-linear components of environments, we find that experts are more consistent (and therefore more predictable) in applying their decision strategies than novices. We do not find, however, any difference in how well experts and novices match environmental models.

Among other things, these results suggest that individuals may have preconceived, simplified expectations of decision environments and try to apply decision strategies coherent with these expectations (see also, Brehmer, 1980; 1994). In our data, redundancy-free and equal-weighting environments are most favorable to the strategies that judges use and, in fact, equal-weighting strategies generally provide a good default (Dawes & Corrigan, 1974). The presence of redundancy and differential cue weights creates a misbalance between individual expectations and environmental structure, and therefore hurts individual performance. In the presence of such imbalance, a good amount of learning is needed to improve human performance. This implies that correct application of decision strategies that rely heavily on a single cue or few cues (e.g., representativeness heuristic, Kahneman & Tversky, 1972; availability heuristic, Tversky & Kahneman, 1973; "take-the-best" heuristic, Gigerenzer & Goldstein, 1996) requires a

certain level of expertise.[6] Performance levels that can be achieved by novices when applying such strategies are rather limited, especially in the environments when cue redundancy, and therefore cue interchangeability, is low.

Regarding the effects of learning, we find that, in addition to redundancy, it is information feedback that improves learning, while neither outcome nor cognitive feedback helps to learn. The effectiveness of task feedback has indeed been emphasized in the literature (e.g., Balzer, Doherty, & O'Connor, 1989), while the effectiveness of outcome feedback has been questioned (e.g., Brehmer, 1980; Hammond, Summers, & Deane, 1973; Hoffman, Earle & Slovic, 1981). In terms of "surprising" results, we were puzzled to find that there were larger effects for learning in field as opposed to laboratory environments. Our surprise was because we thought it would be easier for judges to appreciate relations in a laboratory than in the field. However, the reason may be that greater redundancy in the latter facilitated performance.

We identified several relevant task and judge characteristics concerning the conditions under which the application of bootstrapping models is differentially advantageous. In particular, bootstrapping outperforms clinical judgment more when: (1) environments are more predictable, and in particular, when these are more predictable than judges (cf., Camerer, 1981); (2) the non-linear component of environments is smaller; (3) cue values are given rather than inferred by judges; (4) cue redundancy is low; (5) judges are less experienced; and (6) there are no opportunities for learning.

In terms of more detailed findings, it is useful to consider what the data tell us about the effects of the different lens model statistics represented in Equation 3. Although

---

[6] Interestingly, various studies indeed have reported that experts use surprisingly little information (e.g., Goldberg, 1970).

(as shown in Table 1), there was variation in ranges across all the statistics, it is of interest to recall that environmental predictability, R_e, hardly varied between the factors identified in Table 2 and was always close to mean R_s, response consistency (the correlation between the means of the two measures being close to zero). In addition, the pattern of correlations between the various indices revealed that achievement, r_a, was more strongly correlated with R_s than with R_e (0.52 vs. 0.40) and that whereas R_s was correlated with G (matching), R_e was not. Thus, although environmental predictability, R_e, undoubtedly limits achievement, in these data it explains little variance in differential achievement. Instead, this variance is more adequately captured by G and R_s, i.e., by the particular strategy the judge uses and how consistently this is executed.

Limitations and further research. The limitations in our conclusions result from the fact, that taken as a whole, the 259 environments that we studied could hardly be described as being generated by principles of *representative design* (Brunswik, 1956, see also Dhami, Hertwig, & Hoffrage, 2004). For example, most of the laboratory studies had little or no inter-cue redundancy, an important component of realistic task environments, whereas this feature was present in field studies. In particular, very few laboratory learning studies (8%) contained much redundancy, and yet, we found that greater levels of redundancy lead to better learning.

Interestingly, the presence of redundancy is an important ingredient of Brunswik's psychological framework and suggests that people use different combinations of cues across different trials (so-called vicarious functioning). Unfortunately, by estimating unique sets of weights for individuals across trials, the linear lens model methodology does not capture this aspect of how people may be processing information. This,

therefore, points to the need to develop methodology that can capture this dimension of behavior within a lens model framework and thereby also allow a better understanding of the effects of redundancy. In other words, the methodology assumes that participants are processing the information in a linear manner which, for many tasks, may not be the case.

Recently, there have been some promising and illuminating examples of how lens model research can be conducted in more representative and naturally occurring environments. Specifically, Gosling and his colleagues have investigated overall achievement and matching (of "cue validities" with "utilization coefficients") in judgments of personality made on the basis of the target person's office or bedroom (Gosling, Ko, Mannarelli, & Morris, 2002), websites (Vazire & Gosling, 2004), musical preferences (Rentfrow & Gosling, 2006), and sounds experienced over two days (Mehl, Gosling, & Pennebaker, 2006). We see this work as being very much in the right direction as it neatly captures what people actually do in their natural ecologies.

One advantage of the mathematical formulation of the lens model (i.e., Equation 3) is the neat expression of results in terms of correlational statistics. However, underlying this feature is the implicit assumption that errors in judgment should, in effect, be penalized by a (symmetric) squared error loss function. It may be that in some situations – and particularly in field studies – that this assumption is not appropriate. Work in extending the mathematical framework would thus be most important. It is possible that some of the results we have obtained should be modified.

An important limitation of our investigation was that few studies reported individual level data and thus we were forced to make our analyses on the basis of averages. This limited our ability to comment on individual variability but reflects

reporting practices in science as opposed to specific limitations of lens model studies *per se*. Improvements in information processing and storage in recent years could be harnessed to alleviate this problem in the future. It would be useful, for example, to make use of multi-level hierarchical techniques (see, e.g., Raudenbush & Bryk, 2002) to understand simultaneous group and individual level behavior in lens model studies.

Castellan (1973; 1992) has provided an illuminating critique of the meaning of the matching index, *G*, in lens model studies pointing out some limitations in its interpretation due to mathematical constraints. In our summary data, however, we find little evidence for Castellan's critiques. One reason could be the artificial nature of many of our studies (with orthogonal cues) that allow less ambiguous inferences. Second, most studies involved only two or three cues although it is true that *G* was lower with more than three cues and, particularly, in field studies.

Our results regarding the factors that affect the accuracy of human judgment and the effectiveness of learning also suggest promising directions for further research. First, interactions between task variables can be studied to identify the most optimal conditions for using clinical judgment. For example, in the data we examined, learning is only possible when environments contain cue redundancy. Another important mediator of learning is task information feedback. It would be interesting to investigate the joint effect of the two factors. The data in our sample are insufficient to perform a complete analysis, but our preliminary investigation shows that the positive effect of task information feedback is more likely to occur when cue redundancy is present.

Second, it would be illuminating to study the interactions of task variables and judge characteristics. For example, we found that judges match environmental models

better in redundancy-free environments. Does this finding apply to both novices and experts? Do these two groups of judges react similarly when facing redundancy? One of our other findings suggests that the presence of redundancy may only be beneficial if judges possess sufficient expertise to face it. The interaction of cue redundancy and expertise should be addressed explicitly in further research.

Parenthetically, we note that advances in technology can greatly help collecting data within the paradigm of Brunswik's (1956) paradigm of representative design and linking this with lens model analysis. Hogarth (2006) and Hogarth, Portell, and Cuxart (2007), for example, have exploited the SMS capacity of cell telephones to conduct Experiential Sampling Method (ESM) studies of decision making and the perception of risk. And Mehl et al. (2006) have pioneered the use of the Electronically Activated Recorder (EAR) to sample snippets of ambient sounds in people's environments which can subsequently be used as cues for judgments made by others (see also Mehl, 2006). Moving forward, it is hard not to be optimistic about harnessing these and related technological developments provided researchers are willing to avail themselves of these opportunities.

Concluding remarks. Experimental sciences – like psychology – advance in incremental fashion. New studies appear each year often as a response to immediately preceding papers and what might be called "local" issues (i.e., those that mark certain points in time). One can understand, therefore, why – at the level of individual studies – researchers often adopted simple research designs involving only a few orthogonal cues. It is interesting to ask, therefore, how studies might have been planned some fifty years ago had a future meta-analysis been considered a goal of the research program. How

would the studies have differed? What else would be known today had we been able to plan studies in 259 environments in advance?

This question cannot, of course, be answered unless one first decides on the appropriate research questions. In broad terms, therefore, and using hindsight, the main questions that dominated the research program center not so much on "how good" people are at making judgments *per se*, but on defining the individual and tasks conditions that lead to differential levels of judgment and this includes, of course, learning. This being the case, it can be regretted now that more attempts were not made to widen the kinds of environmental tasks that participants faced. At the same time, the pioneers of lens model studies probably did not envisage the possibilities of meta-analysis which is a fairly recent methodological innovation. However, current researchers are aware of this methodology and, since the lens model paradigm lends itself so well to the methodology, we hope that future research can take our analysis as a starting point. Indeed, the spread-sheet on which our analysis is based is available at www.xxxxxxxx.[7]

Going forward, we note several challenges to research within the lens model paradigm. One – just noted – is to develop methodology that is more flexible in modeling how judges use information. The second – also noted above – is the systematic use of representative design. From many lens model studies, it is not at all obvious to which populations results should or could be generalized. For example, some of our learning environments involved situations where cues had to be "achieved" by participants, and, in these environments, learning effects on human achievement were greater than in those where cues were "given." More generally, this point also speaks to

---

[7] It is our intention to make this spreadsheet both available and user-friendly in the near future.

the issue of studying substantive experts and finding the means to replicate expertise within laboratory settings.

Finally, whereas we have been critical of the limitations of the current linear technology of lens model analysis, we are impressed by the richness of the findings we have uncovered. With more flexible technology, and clearer ideas of how knowledge can be accumulated, we believe that Brunswik's lens model still holds many insights that await discovery.

References

*Adelman, L. (1981). The influence of formal, substantive, and contextual task properties on the relative effectiveness of different forms of feedback in multiple-cue probability learning tasks. *Organizational Behavior and Human Performance*, 27, 423-442.

*Armelius, B-A, & Armelius, K. (1974). The use of redundancy in multiple-cue judgments: Data from a suppressor-variable task. *American Journal of Psychology*, 87 (3), 385-392.

*Ashton, A. H. (1982). An empirical study of budget-related predictions of corporate executives. *Journal of Accounting Research*, 20 (No. 2 Pt I), 440-449.

*Ashton, R. H. (1981). A descriptive study of information evaluation. *Journal of Accounting Research*, 19 (1), 42-61.

*Athanasou, J. A., & Cooksey, R.W. (2001). Judgment of Factors Influencing Interest: An Australian Study. *Journal of Vocational Education Research*, 26 (1). Web version.

Balzer, W. K., Doherty, M. E., & O'Connor, R. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin*, 3, 410-433.

*Balzer, W. K., Sulsky, L. M., Hammer, L. B., & Sumner, K. E. (1992). Task information, cognitive information, of functional validity information: Which components of cognitive feedback affect performance. *Organizational Behavior and Human Decision Processes*, 53, 35-54.

*Bisantz, A. M., & Pritchett, A. R. (2003). Measuring the fit between human judgments

and automated alerting algorithms: A study of collision detection. *Human Factors*, 45 (2), 266-280.

*Bisantz, A. M., Kirlik, A., Gay, P., Phipps, D. A., Walker. N., & Fisk. A. D. (2000). Modeling and analysis of a dynamic judgment task using a lens model approach. *IEEE Transactions on Systems, Man, and Cybernetics – Part A, Systems and Humans*, 30 (6), 605-616.

*Brehmer, B. (1969). Cognitive dependence on additive and configural cue-criterion relations. *American Journal of Psychology*, 82 (4), 490-503.

*Brehmer, B. (1973a). Effects of cue validity on interpersonal learning of inference tasks with linear and nonlinear cues. *American Journal of Psychology*, 86 (1), 29-48.

*Brehmer, B. (1973b). Effects of task predictability and cue validity on interpersonal learning of inference tasks involving both linear and nonlinear relations. *Organizational Behavior and Human Performance*, 10, 24-46.

*Brehmer, B. (1974). The effect of cue intercorrelation on interpersonal learning of probabilistic inference tasks. *Organizational Behavior and Human Performance*, 12, 397-412.

Brehmer, B. (1976). Note on clinical judgment and the formal characteristics of clinical tasks. *Psychological Bulletin*, 83 (5), 778-782.

Brehmer, B. (1980). In one word: Not from experience. *Acta Psychologica*, 45, 223-241.

Brehmer, B. (1994). The psychology of linear judgement models. *Acta Psychologica*, 87, 137-154.

*Brehmer, B., & Hagafors, R. (1986). Use of experts in complex decision making: A paradigm for the study of staff work. *Organizational Behavior and Human*

Decision Processes, <u>38</u>, 181-195.

Brehmer, B., & Joyce, C. R. B. (Eds.) (1988). *Human judgment: The SJT view*. Amsterdam: North-Holland.

*Brehmer, B., & Kuylenstierna, J. (1980). Content and consistency in probabilistic inference tasks. *Organizational Behavior and Human Performance*, <u>26</u>, 54-64.

Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychological Review*, <u>50</u>, 255-272.

Brunswik, E. (1952). *The conceptual framework of psychology*. Chicago: University of Chicago Press.

*Camerer, C. (1979). A general theory of judgment improvement. Unpublished manuscript, University of Chicago, cited in Camerer, C. F. (1981).

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

Castellan, N. J., Jr. (1973). Comments on the ¨lens model" equation and the analysis of multiple-cue judgment tasks. *Psychometrika*, 38 (1), 87-100.

Castellan, N. J., Jr. (1992). Relations between linear models: Implications for the lens model. *Organizational Behavior and Human Decision Processes*, <u>51</u>, 364-381.

*Chasseigne, G., Grau, S., Mullet, E., & Cama, V. (1999). How well do elderly people cope with uncertainty in a learning task? *Acta Psychologica*, <u>103</u>, 229-238.

*Chasseigne, G., Mullet, E., & Stewart, T. R. (1997). Aging and multiple cue probability learning: The case of inverse relationships. *Acta Psychologica*, <u>97</u>, 235-252.

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *The Behavioral and Brain Sciences*, <u>4</u>, 317-370.

Connolly, T., & Miklausich, V. M. (1978). Some effects of feedback error in diagnostic decision tasks. *Academy of Management Journal*, 21, 301-307.

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications*. New York, NY: Academic Press.

*Cooksey, R. W., Freebody, P., & Davidson, G. R. (1986). Teachers' predictions of children's early reading achievement: An application of social judgment theory. *American Educational Research Journal*, 23 (1), 41-64.

*Cooksey, R. W., Freebody, P., & Bennett, A. J. (1990). The ecology of spelling: A lens model analysis of spelling errors and student judgments of spelling difficulty. *Reading Psychology: An International Quarterly*, 11, 293-322.

*Cooksey, R. W., Wyatt-Smith, C., & Freebody, P. (2006). *Assessment as judgment-in-context: Analysing how teachers evaluate students' writing*. Submitted.

*Cooper, R. P., & Werner, P. D. (1990). Predicting violence in newly admitted inmates: A lens model analysis of staff decision making. *Criminal Justice and Behavior*, 17 (4), 431-447.

*Dalgleish, L. I. (1988). Decision making in child abuse cases: applications of SJT and signal detection theory. In Brehmer B. & C. R. B. Joyce (1988). (Eds.), *Human judgment: The SJT view*. Amsterdam, Netherlands: North Holland, 317-360.

Dawes, R. M. (1971). A case study of graduate admissions: Applications of three principles of human decision making. *American Psychologist*, 26, 180-188.

Dawes, R., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, 81, 95–106.

*Deane, D. H., Hammond, K. R., & Summers, D. A. (1972). Acquisition and application of knowledge in complex inference tasks. *Journal of Experimental Psychology*, 92 (1), 20-26.

Dhami, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin*, 130(6), 959-988

*Djamasbi, S., Remus, W., & O'Connor, M. (2004). Does mood influence judgment accuracy? From proceedings: *Decision Support in an Uncertain and Complex World: The IFIP TC8/W8.3 International Conference*, pp. 213-222.

*Doherty, M. E., Tweney, R. D., O'Connor, R. M., Jr., & Walker, B. (1988). *The role of data and feedback error in inference and prediction*. Final report for ARI contract MDA903-85-K-0193, Bowling Green State University.

*Doughtery, T. W., Ebert, R. J., & Callender, J. C. (1986). Policy capturing in the employment interview. *Journal of Applied Psychology*, 71 (1), 9-15.

*Dudycha, L. W., & Naylor, J. C. (1966). Characteristics of the human inference process in complex choice behavior situations. *Organizational Behavior and Human Performance*, 1, 110-128.

*Dunwoody, P. T., Haarbauer, E., Mahan, R. P., Marino, C. J., & Tang, C. C. (2000). Cognitive adaptation and its consequences: A test of cognitive continuum theory. *Journal of Behavioral Decision Making*, 13, 35-54.

*Ebert, R. J., & Kruse, T. E. (1978). Bootstrapping the security analyst. *Journal of Applied Psychology*, 63 (1), 110-119.

Einhorn, H. J. (1971). Use of nonlinear, noncompensatory models as a function of task and amount of information. *Organizational Behavior and Human Performance*, 6,

l-27.

*Einhorn, H. J. (1972). Expert measurement and mechanical combination. *Organizational Behavior and Human Performance*, 7, 86-106.

Einhorn, H.J., Kleinmuntz, D.N., & Kleinmuntz, B. (1979). Linear regression and process-tracing models of judgment. *Psychological Review*, 86, 465-485.

Gifford, R. (1994). A lens-mapping framework for understanding the encoding and decoding of interpersonal dispositions in nonverbal behavior. *Journal of Personality and Social Psychology*, 66 (2), 398-412.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, 103 (3), 592-596.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103, 650-669.

*Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin*, 73 (6), 422-432.

*Gorman, C. D., Clover, W. H., & Doherty, M. E. (1978). Can we learn anything about interviewing real people from "interviews" of paper people? Two studies of the external validity of a paradigm. *Organizational Behavior and Human Performance*, 22, 165-192.

Gosling, S. D., Ko., S. J., Mannarelli, T., & Morris, M. E. (2002). A room with a cue: Personality judgments based on offices and bedrooms. *Journal of Personality and Social Psychology*, 82 (3), 379-398.

*Grebstein, L. C. (1963). Relative accuracy of actuarial prediction, experienced clinicians and graduate students in a clinical judgment task. *Journal of Consulting Psychology*, 27 (2), 127-132.

Hammond, K. R., Hursch, C. J., & Todd, F. J. (1964). Analyzing the components of clinical inference. *Psychological Review, 71,* 438–456.

*Hammond, K. R., & Summers, D. A. (1965). Cognitive dependence on linear and nonlinear cues. *Psychological Review, 72,* 215–224.

*Hammond, K. R., Summers, D. A., & Deane, D. H. (1973). Negative effects of outcome-feedback in multiple-cue probability learning. *Organizational Behavior and Human Performance*, 9, 30-34.

Hammond, K. R., Wilkins, M. M., & Todd, F. J. (1966). A research paradigm for the study of interpersonal learning. *Psychological Bulletin, 65,* 221–232.

Hastie, R., & Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychological Review*, 112 (2), 494-508.

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin*, 57, 116-131.

*Hoffman, P. J., Earle, T. C., & Slovic, P. (1981). Multidimensional functional learning (MFL) and some new conceptions of feedback. *Organizational Behavior and Human Performance*, 27, 75-102.

Hogarth, R. M. (2006). Is confidence in decisions related to feedback? Evidence from random samples of real-world behavior. In K. Fiedler & P. Juslin (eds.), *Information sampling and adaptive cognition* (pp. 456-484). Cambridge, UK: Cambridge University Press.

Hogarth, R. M., & Karelaia, N. (2006). *On heuristic and linear models of judgment: Mapping the demand for knowledge*. Working paper, Universitat Pompeu Fabra, Barcelona.

Hogarth, R. M., Portell, M., & Cuxart, A. (2007). What risks do people perceive in everyday life? A perspective gained from the experience sampling method (ESM). Working paper, Universitat Pompeu Fabra, Barcelona.

*Holzworth, R. J., & Doherty, M. E. (1976). Feedback effects in a metric multiple-cue probability learning task. *Bulletin of the Psychonomic Society*, 8 (1), 1-3.

Hursch, C. J., Hammond, K. R., & Hursch, J. L. (1964). Some methodological considerations in multiple-probability studies. *Psychological Review*, 71 (1), 42-60.

*Jarnecke, R. W., & Rudestam, K. E. (1976). Effects of amounts and units of information on the judgmental process. *Perceptual and Motor Skills*, 13, 823-829.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3, 430-454.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review*, 103 (3), 582-591.

Kaufmann, E., & Athanasou, J. A. (2007). *A meta-analysis of judgment achievement in the Lens Model Equation*. Working paper, University of Mannheim, Germany & University of Technology, Sydney.

*Kessler, L., & Ashton, R. H. (1981). Feedback and prediction achievement in financial analysis. *Journal of Accounting Research*, 19 (1), 146-162.

*Koele, P. (1980). The influence of labelled stimuli on nonlinear multiple-cue probability learning. *Organizational Behavior and Human Performance*, 26, 22-31.

*Lafon, P., Chasseigne, G., & Mullet, E. (2004). Functional learning among children, adolescents, and young adults. *Journal of Experimental Child Psychology*, 88, 334-347.

*Lee, J-W., & Yates, J. F. (1992). How quantity judgment changes as the number of cues increases: An analytical framework and review. *Psychological Bulletin*, 112(2), 363-377.

*Libby, R. (1976). Man versus model of man: Some conflicting evidence. *Organizational Behavior and Human Performance*, 16, 1-12.

*Lindell, M. K. (1976). Cognitive and outcome feedback in multiple-cue probability learning tasks. *Journal of Experimental Psychology: Human Learning and Memory*, 2 (6), 739-745.

*Luft, J. L., & Shields, M. D. (2001). Why does fixation persist? Experimental evidence on the judgment performance effects of expensing intangibles. *The Accounting Review*, 76 (4), 561-587.

*MacGregor, D., & Slovic, P. (1986). Graphical representation of judgmental information. *Human-Computer Interaction*, 2, 179-200.

Martignon, L., & Hoffrage, U. (1999). Why does one-reason decision making work? A case study in ecological rationality. In G Gigerenzer, P. M. Todd and the ABC Research Group. *Simple heuristics that make us smart* (pp. 119-140). New York, NY: Oxford University Press.

Martignon, L., & Hoffrage, U. (2002). Fast, frugal, and fit: Simple heuristics for paired comparison. *Theory and Decision*, 52, 29-71.

Mehl, M. R. (2006). The lay assessment of subclinical depression in daily life.

Psychological Assessment, $\underline{18}$ (3), 340-345.

Mehl, M. R., & Gosling, S. D., & Pennebaker, J. W. (2006). Personality on its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, $\underline{90}$ (5), 862-877.

*Muchinsky, P. M., & Dudycha, A. L. (1975). Human inference behavior in abstract and meaningful environments. *Organizational Behavior and Human Performance*, $\underline{13}$, 377-391.

*Naylor, J. C., & Schenk, E. A. (1968). The influence of cue redundancy upon the human inference process for task of varying degrees of predictability. *Organizational Behavior and Human Performance*, $\underline{3}$, 47-61.

*Newton, J. R. (1965). Judgment and feedback in a quasi-clinical situation. *Journal of Personality and Social Psychology*, $\underline{1}$ (4), 336-342.

*Nystedt, L., & Magnusson, D. (1973). Cue relevance and feedback in a clinical prediction task. *Organizational Behavior and Human Performance*, $\underline{9}$, 100-109.

*O'Connor, M., Remus, W., & Lim, K. (2005). Improving judgmental forecasts with judgmental bootstrapping and task feedback support. *Journal of Behavioral Decision Making*, 18, 246-260.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. New York: Cambridge University Press.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Reilly, B. A., & Doherty, M.E. (1992). The assessment of self-insight in judgment policies. *Organizational Behavior and Human Decision Processes*, $\underline{53}$, 285– 309.

Remus, W., O'Connor, M., & Griggs, K. (1996). Does feedback improve the accuracy of recurrent judgmental forecasts? *Organizational Behavior and Human Decision Processes*, 66 (1), 22-30.

Rentfrow, P. J., & Gosling, S. D. (2006). Message in a ballad: The role of music preferences in interpersonal perception. *Psychological Science*, 17 (3), 236-242.

*Roose, J. E., & Doherty, M. E. (1976). Judgment theory applied to the selection of life insurance salesmen. *Organizational Behavior and Human Performance*, 16, 231-249.

*Rothrock, L., & Kirlik, A. (2003). Inferring rule-based strategies in dynamic judgment tasks: Toward a noncompensatory formulation of the lens model. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, 33 (1), 58-72.

*Rothstein, H. G. (1986). The effects of time pressure on judgment in multiple cue probability learning. *Organizational Behavior and Human Decision Processes*, 37, 83-92.

*Rudestam, K. E., Sherman, R. C., & Jarnecke, R. (1974). Effect of lens-model and outcome feedback in a social judgment analogue. *Psychological Reports*, 35, 1223-1233.

*Schmitt, N., Coyle, B. W., & King, L. (1976). Feedback and task predictability as determinant of performance in multiple cue probability learning tasks. *Organizational Behavior and Human Performance*, 16, 388-402.

*Schmitt, N., Coyle, B. W., & Saari, B. B. (1977). Types of task information feedback in multiple-cue probability learning. *Organizational Behavior and Human*

Performance, 18, 316-328.

*Smith, L., Gilhooly, K., & Walker, A. (2003). Factors influencing prescribing decisions in the treatment of depression: A social judgment theory approach. *Applied Cognitive Psychology*, 17, 51-63.

*Steinmann, D. O. (1974). Transfer of lens model learning. *Organizational Behavior and Human Performance*, 12, 1-16.

*Steinmann, D. O. (1976). The effects of cognitive feedback and task complexity in multiple-cue probability learning. *Organizational Behavior and Human Performance*, 15, 168-179.

*Steinmann, D. O., & Doherty, M. E. (1972). A lens model analysis of a bookbag and poker chip experiment: A methodological note. *Organizational Behavior and Human Performance*, 8, 450-455.

*Stewart, T. R., Roebber, P. J. and Bosart, L. F. (1997). The importance of the task in analyzing expert judgment. *Organizational Behavior and Human Decision Processes*, 69, 205-219.

*Stewart, T. R., Middleton, P., Downton, M., & Ely, D. (1984). Judgments of photographs vs. field observations in studies of perception and judgment of the visual environment. *Journal of Environmental Psychology*, 4, 283-302.

*Stewart, T. R., Moninger, W. R., Grassia, J., Brady, R. H., & Merrem, F. H. (1989). Analysis of expert judgment and skill in a hail forecasting experiment, *Weather and Forecasting*, 4 (1), 24-34.

*Strauss R., & Kirlik, A. (2003). A Systems Perspective on Situation Awareness II: Experimental Evaluation of a Modeling & Measurement Technique. Technical Report AHFD-03-13/NTSC-03-3.

*Summers, S. A., Summers, R. C., & Karkau, V. T. (1969). Judgments based on different functional relationships between interacting cues and a criterion. *American Journal of Psychology*, 82 (2), 203-211.

*Tape, T. G, Kripal, J., & Wigton, R. S. (1992). Comparing methods of learning clinical prediction from case simulations. *Medical Decision Making*, 12, 213-221.

*Tape, T. G., Heckerling, P. S., Ornato, J. P., & Wigton, R. S. (1991). Use of clinical judgment analysis to explain regional variations in physicians' accuracies in diagnosing pneumonia. *Medical Decision Making*, 11(3), 189-197.

*Todd, F. J. (1954). A methodological analysis of clinical judgment. Unpublished doctoral dissertation, University of Colorado, cited in Hursch, C. J., Hammond, K. R., & Hursch, J. L. (1964).

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 4, 207-232.

Tucker, L. R. (1964). A suggested alternative formulation in the developments by Hursch, Hammond and Hursch and by Hammond, Hursch and Todd. *Psychological Review*, 71, 528-530.

*Uhl, C. N. (1966). Effects of multiple stimulus validity and criterion dispersion on learning of interval concepts. *Journal of Experimental Psychology*, 77 (4), 519-527.

Vazire, S., & Gosling, S. D. (2004). e-perceptions. Personality impressions based on personal websites. *Journal of Personality and Social Psychology*, 87 (1), 123-132.

*Wiggins, N., & Kohen, E. S. (1971). Man versus model of man revisited: The forecasting of graduate school success. *Journal of Personality and Social Psychology*, 19 (1), 100-106.

*Wiggins, N., Gregory, S., & Diller, R. (1974). Unpublished study cited in Dawes, R. M., & Corrigan, B. (1974).

*Wigton, R. S., Poses, R. M., Collins, M., & Cebul, R. D. (1990). Teaching old dogs new tricks: Using cognitive feedback to improve physicians' diagnostic judgments on simulated cases. *Academic Medicine*, 65 (9), S5-S6.

*Wright, W. F. (1977). Financial information processing models: An empirical study. *The Accounting Review* (July): 676-689.

*Wright, W. F. (1979). Properties of judgment models in a financial setting. *Organizational Behavior and Human Performance*, 23, 73-85.

*Yntema, D. B., & Torgerson, W. S. (1961). Man-computer cooperation in decisions requiring common sense. *IRE Transactions of the Professional Group on Human Factors in Electronics, HFE-2,* 20-26.

*York, K. M., Doherty, M. E., & Kamouri, J. (1987). The influence of cue unreliability in a multiple cue probability learning task. *Organizational Behavior and Human Decision Processes*, 39, 303-317.

*Youmans, R. J., & Stone, E. R. (2005). To they own self be true: Finding the utility of cognitive information feedback. *Journal of Behavioral Decision Making*, 18, 319-341.

* indicates studies used in meta-analysis.

Table 1: Descriptive statistics of lens model indices

	Mean	Median	Min	Max	Stan. dev	n
r_a	0.55	0.57	-0.06	1.00	0.24	259
G	0.81	0.91	-0.01	1.00	0.23	249
R_e	0.79	0.81	0.00	1.00	0.19	256
R_s	0.80	0.83	0.29	1.00	0.14	250
C	0.05	0.00	-0.11	1.00	0.16	246
GR_e	0.64	0.68	-0.01	1.00	0.24	242
GR_s	0.66	0.73	-0.01	1.00	0.24	243

Correlations

	r_a	G	R_e	R_s	C	GR_e
r_a	x					
G	**0.77**	x				
R_e	**0.40**	0.02	x			
R_s	**0.52**	**0.42**	0.08	x		
C	**0.28**	0.10	-0.11	-0.01	x	
GR_e	**0.90**	**0.79**	**0.61**	**0.38**	0.06	x
GR_s	**0.81**	**0.91**	0.06	**0.72**	0.05	**0.76**

Bold: $p < 0.01$

Table 2: Mean lens model indices by different variables.

	Number of tasks	Average number of: Judges	Judgments	r_a	G	R_e	R_s	C	GR_s	GR_e	GR_e-r_a
Numbers of cues											
2	69	26	48	**0.63**	**0.88**	0.79	0.79	**0.07**	**0.71**	**0.71**	**0.08**
3	90	19	93	**0.55**	**0.88**	0.80	0.81	**0.00**	**0.72**	**0.69**	**0.13**
more than 3	97	16	111	**0.51**	**0.71**	0.79	0.81	**0.08**	**0.58**	**0.55**	**0.07**
Type of cues											
Given	211	18	55	0.55	0.82	0.79	0.81	**0.03**	0.67	0.64	**0.10**
Achieved	40	23	263	0.60	0.79	0.81	0.79	**0.13**	0.63	0.63	**0.05**
Redundancy											
None	106	20	50	**0.61**	**0.89**	0.82	0.81	**0.03**	**0.73**	**0.72**	**0.11**
Some	78	19	97	**0.53**	**0.78**	0.79	0.83	**0.03**	**0.66**	**0.62**	**0.09**
High	25	26	101	**0.54**	**0.76**	0.76	0.80	**0.10**	**0.64**	**0.58**	**0.04**
Function form											
Equal weighting	42	29	65	**0.66**	**0.91**	0.82	**0.81**	0.02	**0.75**	**0.75**	0.10
Compensatory	88	16	99	**0.57**	**0.84**	0.79	**0.83**	0.04	**0.70**	**0.66**	0.10
Non-compensatory	54	22	40	**0.51**	**0.80**	0.84	**0.74**	0.04	**0.62**	**0.65**	0.14
Type of study											
Lab	208	20	87	0.56	**0.83**	0.80	0.80	0.04	0.67	**0.66**	**0.10**
Field	48	16	89	0.52	**0.74**	0.76	0.84	0.09	0.63	**0.57**	**0.06**
Expertise											
Novice	200	20	67	0.56	**0.83**	**0.81**	**0.79**	**0.03**	0.67	0.67	**0.11**
Some training	29	16	208	0.59	**0.84**	**0.72**	**0.85**	**0.09**	0.73	0.61	0.05
Expert	27	13	110	0.49	**0.68**	**0.74**	0.83	**0.17**	0.57	0.51	**0.01**
Learning											
Stable	67	22	155	**0.44**	**0.70**	**0.72**	**0.77**	**0.09**	**0.54**	**0.49**	0.07
Learning	186	19	64	**0.60**	**0.85**	**0.82**	**0.82**	**0.03**	**0.71**	**0.70**	0.10

Note: **Bold** letters denote significant differences for a measure (p < .05)

Table 3: Mean achievement, r_a, in lab and field studies,

by different levels of redundancy.

	Type of study				
	Lab		Field		Difference?
Redundancy					
None	0.61	(106)	--	(0)	--
Some	0.53	(58)	0.52	(20)	ns.
High	0.36	(7)	0.61	(18)	t(23)=-2.53, p=0.019.

Note: Number of observations is given in parentheses.

Table 4: Mean linear cognitive ability, GR_s, in learning and stable studies,

by different levels of redundancy.

	Type of study				
	Stable		Learning		Difference?
Redundancy					
None	0.72	(12)	0.73	(86)	ns.
Some	0.58	(20)	0.69	(57)	t(75)=-2.03, p=0.046.
High	0.46	(13)	0.83	(12)	t(23)=-4.01, p=0.001.

Note: Number of observations is given in parentheses.

Table 5: Improvement of performance due to learning, by type of cues and feedback (means).

| | Improvement of: | | | | Number of | Pre-learning level of: | |
	GR_s		r_a		learning trials	GR_s	r_a
Overall	0.20	(131)	0.18	(138)	103	0.51	0.41
By type of cues:							
Given	0.17	(118)	0.16	(125)	95	0.56	0.43
Achieved	0.47	(13)	0.38	(13)	188	0.08	0.22
By type of feedback:							
No feedback	0.28	(10)	0.26	(10)	77	0.49	0.33
Outcome feedback	0.17	(92)	0.15	(99)	115	0.53	0.44
Cognitive feedback	0.25	(25)	0.21	(25)	74	0.48	0.34
Task information	0.35	(19)	0.29	(33)	105	0.35	0.23
Other types	0.09	(4)	0.07	(4)	86	0.54	0.37

Note: Number of observations is given in parentheses.

Determinants of linear judgment

Table 6: Regression models of differences due to learning.

Dependent variable:	Change in GR_s						Change in r_a					
	Model 1A		Model 1B		Model 2		Model 1A		Model 1B		Model 2	
Predictor:	Coefficient	t-statistic	Coefficient	t-statistic	Coefficient	t-statistic	Coefficient	t-statistic	Coefficient	t-statistic	Coefficient	t-statistic
Intercept			--	--	0.48	4.31			--	--	0.37*	2.82
Initial GR_s (r_a)	-0.42*	-4.12	-0.29*	-3.20	-0.57*	-10.70	-0.37*	-3.26	-0.30*	-2.89	-0.25*	-3.75
Number of learning trials							0.0006	1.98				
Number of cues			0.07*	2.40					0.05	2.11		
Achieved/given					0.24	3.58					0.41	5.30
Cue redundancy									0.09*	2.22		
Weighting function												
Lab/field study												
Feedback												
Outcome												
Cognitive												
Task Information	0.10	2.31	0.12*	2.70	0.08	2.07						
Other												
Expertise			-0.16	-2.19	-0.13	-2.43			-0.17*	-2.38		
R^2 (adjusted)	0.54		0.68		0.82		0.46		0.63		0.53	
n	91		91		91		98		98		98	

Notes:
(1) We only report coefficients that are statistically significant (p < .05)
(2) * p < .01
(3) Models 1A and 1B are obtained using an OLS procedure with White-adjusted standard errors; Model 2 is a robust WLS procedure.

Table 7: Task and judge characteristics explaining the advantage of bootstrapping: regression results.

Predictor:	Model 1 Coefficient	Model 1 t-statistic	Model 2 Coefficient	Model 2 t-statistic
Intercept	0.51	11.46	0.18	4.93
R_e	0.12	5.05	-	-
R_s	-0.56	-16.10	-	-
R_s-R_e	-	-	-0.23	-8.80
C	-0.48	-11.36	-0.35	-6.36
Number of cues				
Achieved/given	-0.04*	-2.35	-0.04*	-2.15
Cue redundancy	-0.02	-3.13		
Weighting function				
Lab/field study	0.05	2.78		
Learning/stable study			-0.06	-3.86
Expertise	-0.02*	-2.49	-0.03	-2.71
R^2 (adjusted)	0.77		0.53	
n	150		150	

Notes:

1. A robust WLS procedure has been used.

2. * $p < 0.05$; all other $p < 0.01$. Only significant coefficients are reported.

Figure 1: Diagram of lens model