Psychological Bulletin
2008, Vol. 134, No. 3, 404–426

0033-2909/08/$12.00 DOI: 10.1037/0033-2909.134.3.404

Determinants of Linear Judgment: A Meta-Analysis of Lens Model Studies

Natalia Karelaia
Université de Lausanne

Robin M. Hogarth
ICREA (Institució Catalana de Recerca i Estudis Avançats)
and Universitat Pompeu Fabra

The mathematical representation of E. Brunswik's (1952) 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 5 decades. Specifically, the authors analyzed statistics of the "lens model equation" (L. R. Tucker, 1964) associated with 249 different task environments obtained from 86 articles. On average, fairly high levels of judgmental achievement were found, and people were seen to be capable of achieving similar levels of cognitive performance in noisy and predictable environments. Further, the effects of task characteristics that influence judgment (numbers and types of cues, inter-cue redundancy, function forms and cue weights in the ecology, laboratory versus field studies, and experience with the task) were identified and estimated. A detailed analysis of learning studies revealed that the most effective form of feedback was information about the task. The authors also analyzed empirically under what conditions the application of bootstrapping—or replacing judges by their linear models—is advantageous. Finally, the authors note shortcomings of the kinds of studies conducted to date, limitations in the lens model methodology, and possibilities for future research.

Keywords: judgmental accuracy, lens model, linear models, learning, bootstrapping

Supplemental material: http://dx.doi.org/10.1037/0033-2909.134.3.404.supp

In a seminal contribution, Hammond (1955) suggested using the conceptual framework of Brunswik's (1952) lens model to study processes of clinical judgment. The actual application involved was the assessment of IQ on the basis of a Rorschach test, but since that time, many psychologists have used the same conceptual framework to study the more general process by which humans make predictions of criteria that are probabilistically related to cues (see, e.g., Brehmer & Joyce, 1988; Cooksey, 1996; Hastie & Kameda, 2005). Consider, for example, an analyst examining financial indicators to predict corporate bankruptcy, a manager using behavior in interviews to assess job candidates, or a physician looking at symptoms that indicate the severity of a disease.

In all of these cases, the simple beauty of Brunswik's lens model lies in recognizing that the person's judgment and the criterion being predicted can be thought of as two separate functions of cues available in the environment of the decision. The accuracy of judgment therefore depends, first, on how predictable the criterion is on the basis of the cues and, second, the extent to which the function describing the person's judgment matches its environmental counterpart (or "the ecology" in Brunswik's, 1955, p. 198, terms).

But how good are people at making judgments and what factors affect this process? These are important questions that have generated considerable controversy in the psychological literature, whether the issues have been studied in terms of logical coherence (e.g., Cohen, 1981; Kahneman & Tversky, 1996) or judgmental accuracy (sometimes called "correspondence"; Hammond, 1996, pp. 103–110). Moreover, given the complexity of human judgment, it is unlikely that these questions can be answered satisfactorily by any single approach.

An advantage of research conducted within the Brunswikian tradition, however, is that, following the development of statistical methods in the 1960s, many researchers have used the same measures for capturing the contribution of different factors that determine the accuracy of judgment within the lens model paradigm. Thus, it is possible to aggregate these measures across many studies and make statements that reflect the accumulation of results. This is the purpose of the current article, in which we present a meta-analysis of studies conducted using the lens model over a period of 5 decades. Consistent with the Brunswikian tradition, a

Natalia Karelaia, Faculty of Business and Economics, Université de Lausanne, Lausanne-Dorigny, Switzerland; Robin M. Hogarth, ICREA (Institució Catalana de Recerca i Estudis Avançats) and Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, Spain.

This research was funded within the EUROCORES European Collaborative Research Projects (ECRP) Scheme, jointly provided by ECRP funding agencies and the European Science Foundation. The research was specifically supported by Swiss National Science Foundation Grant 105511-111621 (to Natalia Karelaia) and Spanish Ministerio de Educación y Ciencia Grants SEC2005-25690-E/SOCI and SEC2006-27587-E/SOCI (to Robin M. Hogarth).

We are indebted to Thomas Stewart, Michael Doherty, and the library at Universitat Pompeu Fabra for helping us locate many lens model studies. We also particularly thank the many researchers who answered our numerous inquiries and requests for data. Finally, we are grateful for the constructive comments on earlier versions of the manuscript received from Mandeep Dhami, Michael Doherty, and Chris White.

Correspondence concerning this article should be addressed to Natalia Karelaia, Faculty of Business and Economics, Université de Lausanne, Internef, 1015 Lausanne-Dorigny, Switzerland, or to Robin M. Hogarth, Department of Economics and Business, Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, 08005, Barcelona, Spain. E-mail: natalia.karelaia@unil.ch or robin.hogarth@upf.edu

major objective is to assess how task characteristics influence judgment.

We are aware, of course, that different formal methods and research traditions exist for capturing outcomes and processes of judgment (see, e.g., Anderson, 1981; Kenny & Albright, 1987). Moreover, these approaches can illuminate aspects of judgment that are not considered within the lens model paradigm. However, the study of judgment beyond this paradigm lies outside of the scope of the present review. As we demonstrate here, the significance of contributions within this tradition justifies limiting the present focus.

The article is organized as follows. We first describe the mathematical formulation of the lens model and, in particular, how judgmental performance, or "achievement", can be decomposed into different measures that form the basis of our subsequent meta-analysis. Second, we discuss various task and individual characteristics that, as shown in the extensive lens model literature, moderate judgmental performance. Third, we specify how we identified, included, and analyzed studies in our meta-sample. Fourth, we present the results of our meta-analysis organized according to four central issues: (a) the overall accuracy of human judgment, (b) task and individual characteristics that affect accuracy, (c) effects of learning and the role of different types of feedback in this process, and (d) whether and when it is advantageous to replace human judgments by models of those judgments, so-called paramorphic representations (Hoffman, 1960). This is known as bootstrapping and has important practical implications (see, e.g., Russo & Schoemaker, 2002). Finally, we summarize the main conclusions of the analysis, indicate shortcomings of the kinds of studies conducted to date, and suggest 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 articles 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 paradigm, human judgment, denoted Y_s, is modeled as a linear function of a set of k cues, X_j, $j = 1, \ldots, k$. Thus,

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

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_js.

Similarly, the environmental criterion, Y_e, can be modeled as a function of the same cues, X_j, $j = 1, \ldots, k$. That is,

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

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_js (see Figure 1). Note that both human judgment and the

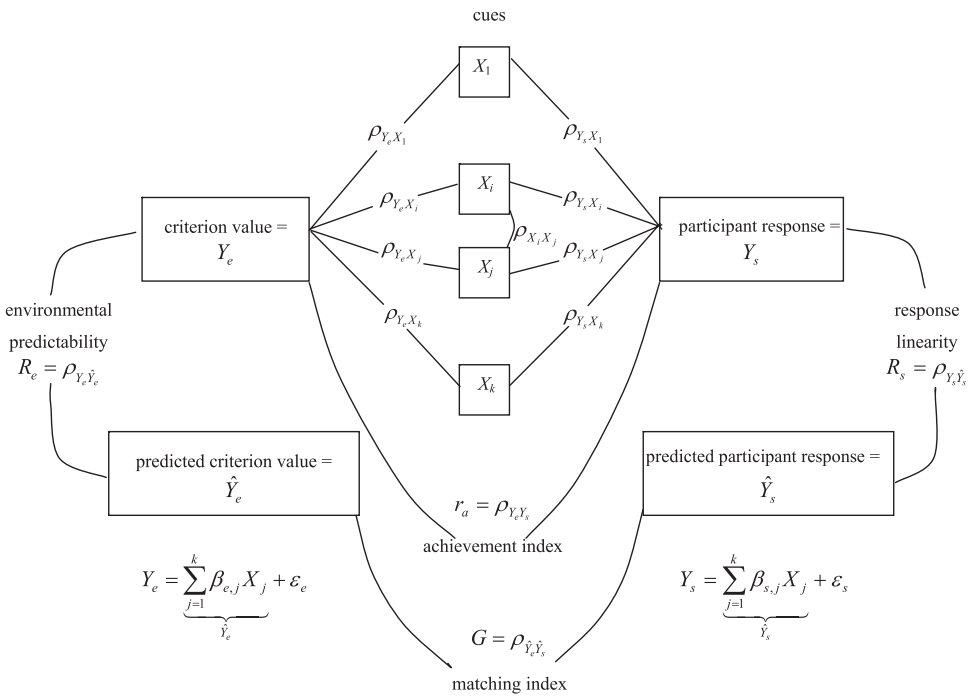

Figure 1. Diagram of the lens model. From "Heuristic and Linear Models of Judgment: Matching Rules and Environments" by R. M. Hogarth and N. Karelaia, 2007, *Psychological Review*, *114,* p. 734. Copyright 2007 by the American Psychological Association. See "The Mathematical Formulation of Brunswik's Lens Model" section of the introductory text for a description of terms.

environment are *probabilistic* models of the cues, in that the cues are never perfectly reliable or valid indicators of the criterion (Brunswik, 1952).

The logic of the lens model is that the person's decisions will best match the environmental criterion to the extent that the judge's reliance on specific cues matches the model of the environment. 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 (Tucker, 1964, p. 528):

$$r_a = GR_eR_s + C\sqrt{(1 - R_e^2)(1 - R_s^2)}, \quad (3)$$

where $G = \rho_{\hat{Y}_e \hat{Y}_s}$ (the matching index) is the correlation between the predictions of both models, that is, 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 the multiple correlations of the models of the environment and the judge, respectively, 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 the error terms of the two models. If these are independent, that is, if $\rho_{\varepsilon_e \varepsilon_s} = 0$, 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 or nonadditive manner.

The Linear Assumption

High linear predictability of the environment (R_e) and judgmental consistency (R_s) observed in various field contexts has led to the conclusion that both environments and judges are often well modeled by linear functions. Linear models can indeed often provide good higher level representations of underlying processes (Einhorn, Kleinmuntz, & Kleinmuntz, 1979). For example, Werner, Rose, and Yesavage (1983) analyzed predictions of imminent dangerousness of psychiatric patients (i.e., engaging in an assault during the first few days after being admitted to an acute care psychiatric unit) made by 30 experienced psychologists and psychiatrists. In this study, both environmental predictability, R_e, and judgmental consistency, R_s, were high (.82 and .84, respectively), even though overall judgmental achievement was low (.18). In another study, Ashton (1982) found that Time Inc. employees were consistent in predicting the actual number of annual advertising pages that appeared in *Time* magazine over several years (R_s of .89). The environment was also highly predictable (R_e of .94) and the task was familiar for these employees who were involved in budgeting tasks. Mear and Firth (1987) came to the same conclusion of high linearity in judgments (R_s of .87) when analyzing the prediction of security returns by financial analysts. However, environmental predictability was lower (R_e of .52).

Are environmental predictability and judgmental consistency related? On the basis of a sample of 15 studies, Camerer (1981) reported that judgmental consistency and environmental predictability were weakly (positively) correlated and that judgmental consistency tended to be generally larger than environmental predictability. Chasseigne, Grau, Mullet, and Cama (1999) systematically varied environmental predictability in a task (from .96 to .32) and found that for each decrement in task predictability, there was a corresponding equal decrement in judgmental consistency.

These findings raise more specific questions: Does the apparent linearity apply to both field and laboratory studies? Is there a relation between linear predictability of environments and judgmental consistency? and, Is the latter generally greater than the former?

Matching

The matching index (G) reflects how well the weights and function forms applicable to the cues in the ecology are represented in the linear model of the person's judgment. This is sometimes stated to be a measure of a judge's "knowledge" of the linear relation in the environment. However, care should be exercised in this interpretation because G can be large even if the weights used by the judge differ considerably from the weights in the ecology. This can occur, for example, in the presence of high inter-cue correlation (Castellan, 1992; Dawes & Corrigan, 1974; Einhorn & Hogarth, 1975).

Residual Correlation

Residual correlation (C), or the correlation between the residuals of the models of the environment and the person, captures the part of judgmental achievement related to cues that have been omitted from the models, nonlinearities in the cue–criterion relations, and possible configurality. Thus, high values of C may reflect: (a) accurate nonlinear or configural use of cues presented by the investigator, (b) accurate linear, nonlinear, or configural use of cues that the investigator did not include in the analysis (i.e., nonmodeled knowledge); or (c) some combination of both (Gorman, Clover, & Doherty, 1978). Configurality concerns interactions between cues (i.e., the impact of one cue on the criterion depends on the level of another cue), whereas nonlinearity refers to nonlinear transformations of individual cues before they are combined. Expertise is associated with greater reliance on configural rules. However, there is little evidence that this improves judgmental performance (Camerer & Johnson, 1997).

Composite Indices

In addition to the above lens model statistics, we are interested in the products of two indices. 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), termed "performance" by Lindell (1976, p. 741) and "linear cognitive ability" by Hogarth and Karelaia (2007, p. 738), quantifies the human, as opposed to the environmental, contribution to achievement and captures the extent to which judges both match task requirements and are consistent in the execution of their strategies.

Second, GR_e—or the product of matching (G) and environmental predictability (R_e) —is an estimate of the validity of the model created when a person is replaced by his or her strategy, that is, by bootstrapping (Camerer, 1981; Dawes, 1971; Goldberg, 1970). This product (GR_e) is of interest because it captures the validity of

the person's strategy assuming that this is applied in a perfectly consistent manner (i.e., when $R_s = 1$).

Many studies have shown that bootstrapping does better than individual judges in clinical decision making (Goldberg, 1970), graduate admissions (Dawes, 1971), employment interviews (Dougherty, Ebert, & Callender, 1986), predicting violent behavior of newly admitted inmates (Cooper & Werner, 1990), and other contexts (sometimes even better than a composite judge, i.e., the average of individual predictions; e.g., Ashton, 1982). The implication is that decision making procedures in many organizations (individual judgment or consensus) should be replaced by models derived from human decision makers. Some studies, however, report similar performance by judges and their models (e.g., Mear & Firth, 1987) or even superior performance by judges (e.g., Libby, 1976, but see Goldberg, 1976).

Generally speaking, bootstrapping does better when the judge has more expertise in the form of valid linear knowledge (high G) and the environment is predictable (i.e., high R_e) (Dawes, Faust, & Meehl, 1989; Kleinmuntz, 1990). In addition, because the bootstrapping technique relies on a linear combination of cues identified by the investigator, linear policy models should be expected to do better than unaided judgment when the residual correlation (C) is low. Finally, Camerer (1981) concluded that bootstrapping improves individual judgment when the criterion for judgment is missing or vague. We examine these issues below in greater depth.

What Factors Affect the Accuracy of Human Judgment?

One of our major objectives was to illuminate the roles of task and individual characteristics in explaining variability in judgmental achievement and its components (see above). Therefore, we now briefly discuss issues studied within the lens model paradigm that helped us define the characteristics we considered in the meta-analysis.[1]

1. Tasks vary in the 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).

2. In addition to combining information, an important dimension of many tasks involves identifying and assessing levels of relevant information (Einhorn, 1972). Therefore, we distinguish between studies where cues are "given" as opposed to "achieved." For the former, decision makers are provided with the explicit values of the cues by the investigator. For the latter, the values of the cues need to be inferred—and often even identified—by decision makers. For example, in a study by Dougherty et al. (1986), judges predicted future performance of job applicants after watching audiotape recordings of employment interviews. Job attitudes, applicants' compatibility with others, and other cues had to be inferred by the judges from the recordings. What are the effects of achieving cue values prior to making judgments?

3. Inter-cue redundancy is an important functional element of decision environments. In particular, it facilitates the interchangeability of cues that Brunswik (1943, 1952) referred to as "vicarious functioning" on the judgment side and "vicarious mediation" on the environment side. Redundancy thus contributes to improving the reliability of overall judgments and can help limit information search without significant reductions in accuracy (Connolly & Miklausich, 1978; Einhorn et al., 1979). Naylor and Schenck (1968) showed that learning increases with positive inter-cue correlations, but Lindell and Stewart (1974) provided evidence that judgment does not improve as a direct function of the magnitude of cue redundancy. We therefore distinguish in the meta-analysis between different levels of inter-cue redundancy.

4. Several studies have focused on how well participants handle different types of functional relations between cues and the criterion (see, e.g., Brehmer, 1980). We distinguish between linear and nonlinear forms of functional relations between the criterion and cues in the ecology. Studies with nonlinear cue–criterion relations on the environmental side of the lens model include curvilinear or configural components in the design of environmental structures (e.g., Rothstein, 1986). Learning nonlinear relations is a difficult task and even when people acquire such knowledge, they experience difficulty in applying this knowledge consistently (Deane, Hammond, & Summers, 1972).

5. In task environments with linear and curvilinear cue–criterion relations (free of any configurality), an additional important characteristic is the dispersion of weights in the ecology ($\beta_{e,j}$). We distinguish three cue-weighting schemes. First, a weighting function is additive noncompensatory 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; see also Hogarth & Karelaia, 2005). Second, all other weighting functions are additive compensatory. However, third, among the latter, we distinguish the special case of equal weighting.

6. An important dimension of the Brunswikian research philosophy is the concept of *representative design* (Brunswik, 1955). The idea behind this concept is that greater generalizability of experimental results can be achieved by conducting experiments under conditions that are representative of people's natural ecologies. We therefore consider differences between laboratory experiments and field studies. In field studies, cue and criterion values are, by definition, representative of the natural ecology of the tasks studied, that is, sampled from naturally occurring stimuli. For example, Kessler and Ashton (1981) used the data on 34 industrial companies listed on Compustat and asked participants to predict corporate bond ratings assigned by Moody's to these companies. In Stewart, Moninger, Grassia, Brady, and Merrem's (1989) study of the accuracy of hail forecasts, the experimenters used a stratified random sampling procedure to select stimuli to guarantee that the base rate (i.e., proportion of volume scans for which hail was verified) in the sample matched that in the larger population of volume scans. In contrast, laboratory studies that use simulated (i.e., hypothetical) values of cues and criterion are typically not representative of natural ecologies (although, in principle, stimuli could be appropriately constructed and respect, for example, the ecological cue–criterion and inter-cue correlations).

7. Whereas field studies are contextually situated, laboratory experiments have involved both contextual and abstract tasks. It is possible that judgmental achievement and learning may be more effective in meaningful, contextual tasks by enhancing judges'

[1] To simplify our presentation, we provide a description of the coding scheme used in the meta-analysis (see Method section) following the same numbers and variables discussed here (e.g., 1 represents number of cues; 2 represents whether cues are given or achieved, and so on).

interest in getting things right (see, e.g., Miller, 1971).[2] For example, in Mear and Firth (1987), security analysts made judgments about security risk on the basis of firms' financial profiles; in Tape, Kripal, and Wigton (1992), medical students predicted the risk of cardiovascular death on the basis of the presence or absence of various concrete risk factors. In contrast, abstract tasks employ nonmeaningful environments with unlabeled cues and/or criteria (e.g., Brehmer, 1974; Jarnecke & Rudestam, 1976).

8. Initial level of expertise in the task domain (i.e., familiarity with the task and having made similar judgments before) is important for achievement. We therefore distinguish between inexperienced judges and experts. However, it is possible to point to individual studies of judgmental achievement involving acknowledged experts that indicate both abysmal (Einhorn, 1972; Roose & Doherty 1976) and incredibly accurate (Stewart, Roebber, & Bosart, 1997) performance. What are the general trends?

9. Learning has been an important topic within the lens model paradigm, where numerous studies have focused on how people learn to utilize cues that are only probabilistically related to a criterion. In the so-called multiple-cue probability learning studies, judgmental accuracy is measured over several blocks of trials (e.g., Brehmer, 1969; Chasseigne, Mullet, & Stewart, 1997; Hammond, Summers, & Deane, 1973), and feedback is often provided over the course of these trials. The participants in studies are frequently nonexperts (Cooksey, 1996, p. 63). Many studies have compared the effectiveness of outcome feedback, cognitive feedback, and task information feedback. *Outcome feedback* is simply knowledge of the criterion value associated with a specific judgment case or profile. *Cognitive* (or process) *feedback* refers to data involving the judge's decision policy (e.g., the weights associated with the different cues in the model of the judge, the $\beta_{s,j}$s). *Task information feedback* is information about true relations in the environment (e.g., $\beta_{e,j}$s and/or function forms) rather than about relations implied by the judge's decisions (i.e., cognitive feedback). Task information feedback is often described as feed-forward because relevant information is provided to judges before they make their judgments.

Previous literature has shown that outcome feedback is helpful in simple, straightforward tasks (e.g., two cues or high predictability of the criterion; Adelman, 1981; Doherty et al., 1988; Hirst & Luckett, 1992; Muchinsky & Dudycha, 1975; cf., Tape et al., 1992) but not in complex, uncertain tasks (Brehmer, 1980; Hoffman, Earle, & Slovic, 1981). Outcome feedback may even impair learning under uncertainty (i.e., when cue–criterion relations are probabilistic) by decreasing consistency in the use of appropriate task knowledge (Schmitt et al., 1976; Schmitt, Coyle, & Saari, 1977) and/or by impeding the development of task knowledge (Hammond et al., 1973; Holzworth & Doherty, 1976).

Task information feedback has been shown to be effective (e.g., Kessler & Ashton, 1981) and to work better than cognitive feedback (Balzer, Doherty, & O'Connor, 1989). Moreover, when combined with cognitive or outcome feedback, task information feedback is even more effective (Balzer, Sulsky, Hammer, & Sumner, 1992; Reilly & Doherty, 1992). However, sometimes providing only task information may be sufficient (Reilly & Doherty, 1992; Remus, O'Connor, & Griggs, 1996). In addition, experience is an important determinant of judges' ability to benefit from feedback. In particular, Steinmann (1976) found that experienced judges used cognitive feedback and task information feedback to improve their judgments to a greater degree than less experienced judges.

Method

Database for the Meta-Analysis

By extensively searching several common databases (e.g., PsycINFO), identifying references to key articles, and consulting leading contributors to the literature, we identified more than 200 published and unpublished works that suggested that they might contain lens model data, specifically the components of Equation 3.[3] Of particular assistance was the "Annotated Bibliography of Cue Probability Learning Studies" prepared by R. J. Holzworth, which includes 315 references to journal articles, book chapters, doctoral dissertations, and technical reports that appeared between 1955 and 1999.[4] Furthermore, we placed requests for any related data on electronic mailing lists for members of the Brunswik Society and the Social Psychology network and thereby accessed several recent published and unpublished manuscripts.[5] Finally, we searched the reference sections of all retrieved reports. In cases in which some data were missing from reports or were unclear, we sent requests to authors to obtain the data. We did not explicitly limit the linguistic coverage of items searched; however, all identified reports in our database were in English.

Exclusion Criteria

We excluded from consideration experimental reports that did not model the environmental side of the lens, that is, for which criterion data were missing (e.g., Kuo & Liang, 2004), research within the conflict resolution paradigm in which the criterion for one person is the judgment of others (e.g., Hammond, Wilkins, & Todd, 1966), and studies in which the unit of analysis was aggregate (typically mean), as opposed to individual, judgments (e.g., Gifford, 1994). We note, parenthetically, that whereas there are numerous studies of the last category (especially in social psychology), they are quite contrary to the Brunswikian tradition in that they confuse idiographic (within-individual) and nomothetic (within-group) levels of inference. Finally, we excluded works where the data on judgmental achievement (r_a) were not provided and it was not possible to estimate them by substituting the available statistics into the lens model equation (e.g., Miklich &

---

[2] We thank one of the reviewers for this idea. In addition, Cooksey (1996, p. 88) discussed different combinations of expertise and context within the multiple-cue probability learning paradigm.

[3] 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 is more limited than ours, and their criteria for including studies in the analysis are different (e.g., they excluded studies that addressed learning and/or the effect of feedback on human judgment). However, Kaufmann and Athanasou's analyses covered issues that are not considered in our article. For example, they compared the levels of human achievement across different areas of decision making, such as medical, business, educational, and others. Their work thus should be considered complementary to what is presented here.

[4] This resource is available at: http://www.brunswik.org/resources.

[5] These resources are available at http://www.brunswik.org/index.html and http://www.socialpsychology.org.

Gillis, 1975). Among all retrieved reports, 86 were retained after applying the exclusion criteria.

Defining Studies

Many experimental reports contained more than one study, that is, they examined judgments in more than one environment or experimental setting using a between-subjects design. Task features, characteristics of participants, or experimental procedures differed across environments within the same report. We defined a study as the smallest experimental unit described. A study is therefore a task environment with unique (within the report) features of the experimental setting. A few reports contained results on more than one task administered using a within-subject design. In such cases, we averaged the results across experimental tasks and included only these averaged figures in the database to guarantee the independence of data points. Thus, all data points in the database corresponded to different participants and did not overlap. The split of the 86 experimental reports into studies led to a total of 249 different experimental studies, or environments in which judgments were made.

Coding Studies

We characterized the 249 studies by the averages of the lens model statistics of the participants in each of these environments. These averages were taken directly from the articles, inferred (e.g., from graphs), or calculated by utilizing the properties of the lens model equation (Equation 3). In 35 studies, we had to assume the correlation between the residuals of the models of the judge and the environment (C) to be 0 in order to calculate missing components of the lens equation (judgmental achievement, matching, or consistency). We later excluded these 35 null values from further analyses of the residual correlation (C). In the final data sample of 249 environments, 13 values of G, 3 values of R_e, 12 values of R_s, and 45 values of C—but no values of r_a—were missing.

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 nonlearning 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).

We emphasize that our unit of analysis is the average of statistics of individuals within each environment (i.e., "average judge") as opposed to the actual individual statistics. Unfortunately, the vast majority of articles did not provide individual-level data, so we are unable to comment on variation within the different environments.

In addition to the lens model statistics, we encoded, when available, the following variables characterizing the specific tasks and the participants.

Number of cues. We recorded the exact number of cues in each study and additionally coded the data into three groups: two, three, or more than three cues.

Type of cues. Studies where participants were provided with explicit cue values by the experimenter were coded as "given" (0). When the values of the cues needed to be inferred and/or identified by participants, we coded the study as "achieved" (1).

Inter-cue redundancy in the ecology. We classified environments by the level of inter-cue redundancy as either "none" (no redundancy), "some" (if the average of absolute intercorrelations was less than or equal to .40 or redundancy was described as being *low*, *moderate*, or *some*), or "high" (otherwise).

Function form in the ecology. We classified studies as "nonlinear" (1) if they explicitly included either curvilinear or configural relations between criterion and cues in the ecology (e.g., Rothstein, 1986). All other studies were coded as "linear" (0). Among nonlinear studies, we also identified studies where the criterion was a monotonic function of the cues and where this was not the case (i.e., U-shaped or inverted U-shaped functions).

Cue weights in the ecology. In studies without any configural element (i.e., cue interaction), we classified the distributions of the ecological cue weights into noncompensatory (the weight of each cue exceeds the sum of those smaller than it in the ordered sequence of cues) or compensatory (otherwise), except that we also coded the special case of equal weighting.

Type of study. We classified a study as "field" (1) if it involved cue and criterion values that were representative of the natural ecology of the task, that is, sampled from real stimuli. Studies that used hypothetical values of cues and criterion were coded as "laboratory" (0).

Context. We coded studies as "abstract" (0) if they used nonmeaningful environments with unlabeled cues and/or criteria and as "concrete" otherwise (1).

Expertise. We classified participants in three groups by the level of initial expertise (i.e., before learning trials, if any): "novices," "experts" (two extreme categories), and "some training" (an intermediate category).

Learning and feedback. We distinguished between studies where examining learning over multiple trials was an explicit objective and/or in which participants were explicitly given the possibility to learn through feedback over several blocks of trials (and therefore the analysis was done within each block) and studies without this possibility. We labeled the former as *multiple-block* studies and the latter as *one-shot* studies. For multiple-block studies, we also recorded the number of learning trials and type of feedback given to participants. We classified feedback into four categories: "outcome feedback" (1, if provided on learning trials between the first and last blocks; 0, otherwise), "cognitive feedback," "task information feedback," and "other types of feedback." For studies with outcome feedback, we additionally coded two variables identifying whether the feedback was provided on the first block of trials and on the last block of trials. Under other types of feedback, we classified studies where, for example, the information on the relation between the ecology and the judge's decision strategy (i.e., functional validity feedback; see Balzer et al., 1992) was provided to the participants (e.g., O'Connor, Remus, & Lim, 2005).

The coding form contained 18 variables, a product of author discussion (N. Karelaia and R. Hogarth), and included moderator variables suggested by previous literature (see above). All 86 experimental reports were coded independently by each of us. Disagreements were rare and were resolved through discussion of the issues involved. We did not quantify intercoder agreement because the items to be coded were relatively simple and did not

involve subjective judgments, such as, for example, overall quality of study methodology (Hunter & Schmidt, 2004, p. 471).[6]

Study Characteristics

The 86 experimental reports included in our analysis appeared between 1954 and 2007, one half dating before 1981 and the other half afterwards. When splitting the 1954–2007 interval into 5-year periods, we found that the 1974–1978 period contained the largest number of articles: 17. Interest in the topic then declined, as judged by the number of articles we identified, but increased again at the end of the 1990s. Of the 86 experimental reports, 4 were unpublished manuscripts, 2 were technical reports, 1 was a book chapter, 1 was a conference proceeding, and the rest were journal articles. The 86 experimental reports involved 143 different authors.

The mean number of participants in the 249 environments was 20 (interquartile range = 10 to 24), each participant making, on average, 58 judgments (interquartile range = 25 to 120). Thus, the total number of individual judgments on which our results were based was large: about 303,000.[7]

Analyses

The main dependent variables of interest were the components of Equation 3. Thus, the data we analyzed consisted of the average values of these components—that is, average correlations—across the individual participants within each of the 249 studies (see also above). We did not apply Fisher's Z transformation to these average correlations prior to combining them, as such transformed estimates have been shown to produce substantial upward biases (larger than downward biases associated with untransformed correlations coefficients) in meta-analytic models, especially if there is variation in population correlations across studies[8] (Field, 2001; Hunter & Schmidt, 2004, pp. 82–83; Strube, 1988).[9] Notably, in 130 of 249 studies, Fisher's Z transformation had been applied to the individual correlations of each participant before they were averaged. In addition, we applied Fisher's Z transformation to individual correlations when such data were available (39 of the remaining 119 studies) and then calculated averages of the transformed correlations.

To combine the average correlations from different studies and to correct for study-level sampling error in the estimation of the population correlations, we used frequency weights (Hunter & Schmidt, 2004, p. 81). Frequency weights were determined by multiplying the number of participants in a given study by the number of judgments each participant made in the study.[10] The results described below are obtained on the weighted data.

Several analyses of the weighted data were performed. First, we examined average (weighted) lens model indices and confidence intervals at the aggregate level. To assess variability in the data, we calculated the within-group heterogeneity statistic, Q (e.g., Ader & Mellenbergh, 1999, p. 304), as well as the recently proposed I^2 index (Higgins & Thompson, 2002). A significant estimate of Q suggests that variability in the meta-data is greater than would be expected from sampling error alone, that is, that the population correlations vary from study to study. The I^2 index also quantifies this variability and is easier to interpret; it corresponds to the total variability across the meta-data due to true heterogeneity, that is, to between-study variability. Higgins and Thompson (2002) proposed to interpret I^2 percentages of around 25%, 50%, and 75% as low, medium, and high heterogeneity, respectively.

When analyzing the meta-data, we used random-effects, instead of fixed-effects, models that allow population correlations to vary from study to study (DerSimonian & Laird, 1986). This choice was deliberate because the data contained between-study random differences that go beyond within-study sampling variability; that is, the assumption of between-study homogeneity was suspect (see analysis below). Moreover, the random-effects approach is better suited to generalizing results to a wider population of studies beyond those included in the meta-analysis (e.g., Field, 2001; Hedges & Vevea, 1998; Hunter & Schmidt, 2004, p. 202; Rosenthal & DiMatteo, 2001). Random-effects models use adjusted weights that incorporate the random-effects variance component (e.g., Lipsey & Wilson, 2001, p. 119) and are generally more conservative than fixed-effects models in that confidence intervals are larger.

When considerable between-study heterogeneity was detected, we stratified the studies to identify moderator variables that might account for variability in the averaged lens model indices. It is important to note that subgroup meta-analyses are observational by nature, and reliance on statistical significance tests can be misleading (e.g., Hunter & Schmidt, 2004, p. 70). We therefore considered, as a rule of thumb, the overlap of 95% confidence intervals of summary estimates to quantify the magnitude of the differences between subgroups. In addition, we fit meta-regression models with several study-level explanatory variables to explore possible simultaneous effects. Meta-regressions differ from simple regressions in that they are weighted regressions and random-effects models that allow for residual heterogeneity among studies not modeled by explanatory variables.

We assessed the presence of possible publication bias (or availability bias; Hunter & Schmidt, 2004, p. 493). This refers to a tendency for published and available (larger) studies to report, on average, larger mean effect sizes than unpublished and unavailable (smaller) ones (e.g., Lipsey & Wilson, 2001, p. 165). Parenthetically, in the lens model literature, the objective is not to show that correlational indices are statistically significant but rather to quan-

---

[6] The spreadsheets that detail the coded data on which our analysis was based are available online as supplemental material at http://dx.doi.org/10.1037/0033-2909.134.3.404.supp.

[7] In fact, the total was somewhat larger because this figure included only judgments corresponding to the last block of trials of the studies that report judgments across more than one block. In what follows, we present analysis of judgments corresponding to the last block of trials, except when exploring the effect of learning.

[8] There was substantial variability in our data that obliged us to use random-effects meta-analytic models (see Results section, *How Accurate is Human Judgment Overall?*).

[9] At very high levels of average correlations, there is no consistent advantage to using transformed instead of untransformed correlations and vice versa. In particular, the magnitude of underestimation of the average population parameters with untransformed correlations becomes similar to the overestimation with transformed correlations (Field, 2001; see also Law, 1995).

[10] The choice of frequency weights captures the notion that little weight should be given to studies with either few judgments per individual participant and/or few participants.

Table 1
Descriptive Statistics of Lens Model Indices

Lens model index	*M* (weighted)	95% confidence interval	*n*	*Q*	I^2 (%)	τ^2	Correlations					
							r_a	G	R_e	R_s	C	GR_e
r_a	.56	.53–.59	249	17,319*	.057	99	—					
G	.80	.76–.83	236	19,829*	.067	99	.78**	—				
R_e	.81	.79–.84	246	10,706*	.035	98	.43**	.10	—			
R_s	.80	.79–.82	237	5,644*	.019	96	.56**	.43**	.14*	—		
C	.04	.02–.06	204	6,249*	.023	97	.23**	.03	−.23**	−.05	—	
GR_e	.65	.61–.68	236	20,668*	.070	99	.91**	.82**	.63**	.41**	−.08	—
GR_s	.66	.63–.69	236	17,469*	.060	99	.83**	.92**	.12	.72**	−.03	.78**
$GR_e - r_a$	.10	.09–.11	236	2,461*	.008	90						

Note. See "The Mathematical Formulation of Brunswik's Lens Model" section of the text for a description of the lens model indices. Q represents within-group heterogeneity; I^2 is the percentage of variation attributable to between-study heterogeneity; τ^2 is the DerSimonian and Laird (1986) estimate of between-study variance.
* $p < .05$. ** $p < .01$.

tify them. Indeed, many published studies have reported low values of the lens model components, such as judgmental consistency, matching, and achievement, and these findings have been used to stress the need to use statistical models instead of unaided human judgment (e.g., Cooper & Werner, 1990; Einhorn, 1972). Therefore, publication bias, as understood in meta-analysis literature (i.e., based on statistical significance), was not an issue here. Nonetheless, we performed formal statistical tests to detect any publication bias and quantify its magnitude. We used two recently proposed statistical tests: Begg's test (Begg & Mazumdar, 1994) and Egger, Davey Smith, Schneider, and Minder's (1997) regression method.[11] When both procedures identified publication bias, we inspected funnel plots[12] (e.g., Lipsey & Wilson, 2001, pp. 142–143) to determine whether small studies reporting small correlations were indeed underrepresented and applied Duval and Tweedie's (2000) nonparametric trim-and-fill imputation technique. This method adjusts meta-data to incorporate the theoretical missing studies with an appropriate fixed- or random-effects model.

All analyses were performed with specific Stata macros written for meta-analytic data (Sterne, Bradburn, & Egger, 2001).

Results

We structured our results around four questions: (a) How accurate is human judgment overall? (b) What task and individual factors affect the accuracy of human judgment? (c) How effective is learning and what is the role of feedback in this process? (d) What factors affect the accuracy of bootstrapping models? Our answer to the first question is essentially limited to providing descriptive statistics of the different components of the lens model equation identified above. The answers to the remaining questions involve, first, providing breakdowns of the lens model components by different task and individual characteristics (variable-by-variable analysis) and, second, extensive use of meta-regression techniques to capture the separate and simultaneous effects of different moderator variables.

How Accurate Is Human Judgment Overall?

The left-hand part of Table 1 reports mean (weighted) values and 95% confidence intervals of the lens model indices for the aggregated data, as based on random-effects models. Across all observations, mean achievement (r_a) was .56, mean matching (G) and mean response consistency (R_s) were both .80, and mean residual correlation (C) was .04. On the environmental side, predictability (R_e) was, on average, .81. High average values for environmental predictability and linear response consistency suggest that both environments and individuals can be modeled well by linear functions of cues. Contrary to Camerer's (1981) analysis, R_s was not generally larger than R_e.

As for the two composite statistics, mean linear cognitive ability (GR_s) was .66 and the mean validity of bootstrapping models (GR_e) was .65. The latter value exceeded mean achievement of individual judgment by .10.

To explore relations between the various indices, we considered the pairwise correlations, as presented in the right-hand part of Table 1 (DerSimonian & Laird's, 1986, random-effects weights were applied). Several significant correlations ($p < .01$) were not unexpected. In particular, consistent with Equation 3, high positive correlations were observed between achievement (r_a) and (a) matching (G): .78; (b) response consistency (R_s): .56; and (c) environmental predictability (R_e): .43. In addition, the fact that r_a and C, the correlation between residuals, were moderately correlated (.23) suggests significant nonlinear and/or nonadditive usage of cues and/or omitted variables. Less obvious a priori was the significant correlation between the two statistics that characterize performance independent of environmental predictability, namely, matching (G) and response consistency (R_s). This correlation was positive (.43) and suggests that judges who match the environment better are also more consistent in executing their judgment.

Studies with higher environmental predictability tended to report slightly higher judgmental consistency. Consistency (R_s) had a weak positive correlation with environmental predictability (R_e):

[11] Begg's test examines whether the Kendall's rank correlation between effect sizes and their standard errors is zero. It is fairly powerful for large meta-analyses (Ader & Mellenbergh, 1999). The regression method tests for a linear association between effect sizes and their standard errors.

[12] Funnel plots graph effect sizes against study sample size (or sampling error).

Table 2
Mean (Weighted) Lens Model Indices by Different Study Characteristics

Study characteristic	No. studies	Average no. Judges	Average no. Judgments	*M* (weighted) *Adj.* (*n*) r_a	G	R_e	R_s	C	GR_s	$GR_e - r_a$
No. of cues										
Two	72	24	48	.63	.89 *.84 (20)*	.80	.78	.10	.71	.10
Three	77	20	49	.55	.86 *.82 (17)*	.81	.80	−.02*	.71	.13
More than three	99	18	73	.52	.70	.82	.82	.04	.59	.08
Type of cues										
Given	202	19	48	.56	.81 *.77 (36)*	.80 *.77 (25)*	.81 *.78 (41)*	.04	.67	.10
Achieved	43	28	104	.58	.77	.86	.76	.06*	.62	.10
Cue redundancy										
None	102	21	47	.66	.89	.85	.81	.04*	.75	.11
Some	85	21	54	.48	.71	.80	.83	.03	.55	.09
High	23	22	109	.54	.75	.79	.80	.06*	.61	.08
Function form in the ecology										
Linear	189	22	60	.56	.81	.80	.82	.03	.68	.09
Nonlinear	56	14	37	.55	.77	.84 *.79 (13)*	.73	.08*	.58	.12
Cue weights in the ecology										
Compensatory	97	19	58	.55	.76	.85	.84	.05	.65	.09
Noncompensatory	53	20	45	.54	.83 *.78 (13)*	.85	.77	−.04*	.67	.15
Equal weighting	33	32	62	.67	.91	.84	.79	−.01*	.74	.10
Type of study										
Laboratory	183	21	49	.60 *.54 (33)*	.84	.84	.79	.05	.69	.11
Field	65	19	85	.45	.69	.73	.83 *.75 (37)*	.03	.59	.07
Context										
Abstract	84	21	48	.56	.84 *.80 (14)*	.83	.76	.02*	.67	.13
Concrete	163	20	59	.56	.79	.80	.82	.07	.66	.08
Expertise										
Novice	204	21	48	.58	.83	.83 *.79 (36)*	.79 *.75 (42)*	.03	.68	.12
Some training	15	22	115	.51	.77	.66	.84	.08	.67	.03
Expert	29	14	104	.47	.60	.77	.85	.09	.54	.02
Learning										
One shot	49	20	101	.41	.63	.71	.80	.07	.52	.05
Multiple blocks	199	21	48	.60	.85	.84 *.80 (40)*	.80 *.76 (41)*	.03	.70	.11

Note. See "The Mathematical Formulation of Brunswik's Lens Model" section of the text for a description of the lens model indices. Correlations adjusted for publication/availability bias are shown in italics. The trim-and-fill method was used; the number of filled studies is given in parentheses. High heterogeneity remained within all of the subgroups.
*null effect size ($p > .05$).

.14 ($p < .05$).[13] However, the correlation between R_e and matching (G) was not significant (.10, *ns*), suggesting that decision makers can match environmental models equally well in noisy and predictable environments. Similarly, linear cognitive ability (GR_s) did not correlate significantly with R_e.

For all reported mean (weighted) indices, the estimates of heterogeneity (Q) were large and significant ($p < .01$), indicating that the variability across the data points surpassed within-study variability (the left-hand part of Table 1). This conclusion was confirmed by the estimates of I^2 percentages. For all lens model indices, almost 100% of variation in the meta-data was due to between-study heterogeneity. We provide DerSimonian and Laird estimates of between-study variance (τ^2) in the last column of the left-hand part of Table 1. The estimates are the largest for achievement (r_a), matching (G), and the composite statistics (GR_s and GR_e). The considerable between-study variability in the data has two implications. First, random- instead of fixed-effects models should be used. Second, the next step should be to determine moderator variables that explain this variability.

What Factors Affect the Accuracy of Human Judgment?

Table 2 classifies the lens model indices according to the task and individual characteristics enumerated above (see *Coding Studies* in the Method section). The numbers of studies in each category, average numbers of judges per study, average numbers of judgments made by each judge, and mean (weighted) lens model indices are specified. Table 3 presents the proportion of variance explained by each variable. All variables were significant in explaining the variance between the subgroups ($p < .05$) except for one entry marked with an asterisk. This means that the data across subgroups differed by more than sampling error.

Before providing an analysis of the results presented in Tables 2 and 3, we emphasize three points. First, the categories used, such as laboratory versus field, multiple blocks versus one shot, and levels of expertise, were not independent. In particular, in the laboratory compared with the field category, there were proportionally more learning studies (95% vs. 40%) and studies involving novice judges (94% vs. 49%) and fewer studies with concrete context (54% vs. 100%).

[13] It is interesting to note that mean R_e and mean R_s were close in value. However, as noted above, mean R_e is a "true mean," whereas mean R_s is a "mean of means." Thus, within any given study, values of R_s at the individual level can vary quite a lot for fixed levels of R_e.

Table 3
Heterogeneity Explained by Different Study Characteristics

Variable	Q_{bw}								Proportion of Q_{bw} to total Q, %							
	r_a	G	R_e	R_s	C	GR_s	GR_e	$GR_e - r_a$	r_a	G	R_e	R_s	C	GR_s	GR_e	$GR_e - r_a$
No. cues	555	2,185	216	862	522	1,921	1,268	252	3.2	11.0	2.0	8.0	8.4	11.0	6.1	10.2
Achieved/given cues	1,054	453	795	25	92	420	1,156	20	6.1	2.3	7.4	0.2	1.5	2.4	5.6	0.8
Cue redundancy	2,852	2,339	1,699	627	314	2,243	3,797	152	16.5	11.8	15.9	5.9	5.0	12.8	18.4	6.2
Function form in the ecology	1,279	1,276	1,154	25	33	847	2,111	165	7.4	6.4	10.8	0.2	0.5	4.9	10.2	6.7
Cue weights in the ecology	2,311	2,870	1,600	965	535	2,466	3,936	385	13.3	14.5	14.9	9.0	8.6	14.1	19.0	15.7
Laboratory/field study	1,974	1,939	1,573	105	4 (*ns*)	875	3,354	217	11.4	9.8	14.7	1.0	0.1	5.0	16.2	8.8
Context	1,633	1,802	1,183	409	184	1,477	2,547	182	9.4	9.1	11.1	3.8	2.9	8.5	12.3	7.4
Expertise	871	1,722	965	251	328	792	2,245	550	5.0	8.7	9.0	2.3	5.2	4.5	10.9	22.3
Learning	1,128	1,989	841	19	198	1,239	2,247	245	6.5	10.0	7.9	0.2	3.2	7.1	10.9	9.9

Note. See "The Mathematical Formulation of Brunswik's Lens Model" section of the text for a description of the lens model indices. All between-subgroup heterogeneity statistics (Q_{bw}) are significant ($p < .05$), except where denoted nonsignificant (*ns*).

Second, we assessed the presence of publication bias in the collected values of r_a, G, R_e, R_s, and C within all subgroups of studies. In the subgroups where publication bias was detected by the Begg and Egger procedures, we recalculated combined effects, incorporating the "missing" studies according to Duval and Tweedie's (2000) trim-and–fill method (see above). These adjusted values are shown in italics within the relevant cells in Table 2. The average downward adjustment was .04. Notably, funnel plots of the relevant subgroups did not contain signs of underrepresenting small studies with reported small correlations. Therefore, the adjusted values should be interpreted with caution and not taken at face value.

Third, considerable heterogeneity remained in all subgroups (Q: $p < .05$; I^2 greater than 50%). This implies that no single study-level variable can account for the variability in the meta-data and that the influence of several such variables should be examined simultaneously.

Effects of Specific Moderators

We next analyzed the effect of each moderator separately, commenting here on the largest proportions of variance explained by each. We further complemented the variable-by-variable analysis with the meta-regression technique.

Number of cues. In these data, the number of cues explained 11% of the heterogeneity of G, 8% of the heterogeneity of R_s, 8.4% of the heterogeneity of C, and 11% of the heterogeneity of GR_s. Human achievement, r_a, was negatively affected by the number of cues. In particular, mean r_a was only .52 in environments with more than three cues, whereas it reached .55 and .63 in the environments with three and two cues, respectively. However, the number of cues explained only 3.2% of heterogeneity of r_a. Matching (G) decreased when the number of cues increased. The lowest value of mean G corresponded to the subgroup with more than three cues (.70; Table 2). Linear cognitive ability (GR_s) was also negatively affected by a greater number of cues. However, the effect for judgmental consistency (R_s) seemed to be positive. As for the correlation between residuals (C), it was the highest for the simplest case of two cues.

Given–achieved cues. The variable of given versus achieved cues alone explained 6.1% of heterogeneity of achievement (r_a) and lower proportions of variability of R_s, G, and C. The data in Table 2 suggest that r_a was slightly larger in the studies where cues were achieved by judges (.58) than in the studies where the cues were provided directly by experimenters (.56). Nevertheless, there was a significant overlap of the confidence intervals for the population mean (.52–.59 for given cues vs. .51–.65 for achieved cues). Overall, there were no large differences between studies with achieved and given cues when this task variable was analyzed alone.

Cue redundancy. When analyzed alone, cue redundancy negatively affected the average level of matching, G (.89 with none vs. .71 with some and .75 with high redundancy), explaining 11.8% of the heterogeneity. As a consequence, linear cognitive ability (GR_s) was also negatively affected by the presence of cue redundancy in the ecology, with 12.8% of heterogeneity explained. On the contrary, mean response consistency (R_s) did not differ much across the subgroups, and the confidence intervals overlapped substantially. The overall influence of redundancy on achievement (r_a) was negative, with 16.5% of variability explained by redundancy. However, the possible influence of environmental predictability (lower in studies with higher redundancy, *ceteris paribus*) should be filtered out before drawing any conclusion.

Although surprising at first, the negative effect of redundancy on G may be due (at least partially) to a positive relation with the number of cues. In particular, only 16% of studies with more than three cues contain no inter-cue redundancy, which is well below the 85% of environments with two cues and 61% of environments with three cues. (In addition, among the environments with three cues, none is classified as containing high redundancy.)

To separate the effects of inter-cue redundancy, number of cues, and linear predictability of the environment, we fit a meta-regression model, with G as the dependent variable and number of cues, cue redundancy, and R_e as predictors. The results showed significant negative effects of both cue redundancy ($B = -.06$,

$SE_B = .03$, $p < .05$)[14] and the number of cues ($B = -.07$, $SE_B = .02$, $p < .01$). The regression explained 19% of the variability of G. A similar meta-regression explained 28% of the variability of achievement (r_a), with negative coefficients for both redundancy ($B = -.06$, $SE_B = .02$, $p < .05$) and number of cues ($B = -.04$, $SE_B = .02$, $p < .05$). That is, people match the environmental model better when cue redundancy and the number of cues are smaller, controlling for linear environmental predictability.

Function form in the ecology. Studies with nonlinear cue–criterion relations, as compared with studies with linear functions in the ecology, reported lower matching, G (0.77 vs. 0.81), and judgmental consistency, R_s (0.73 vs. 0.82). However, function form alone explained only 6.4% of variability of G and 0.2% of variability of R_s. Of the 56 studies with nonlinear cue–criterion relations, 50 studies involved nonmonotonic forms (e.g., an inverted U-shape).

Cue weights in the ecology. The three-level categorical variable defining the distribution of cue weights explained 13.3% of the variability of r_a, 14.5% of that of G, and 9% of the variability of R_s. Mean achievement (r_a) was highest in equal-weighting environments (.67). Judgmental consistency (R_s) was smallest in noncompensatory environments (.77 vs. .79 in equal-weighting and .84 in compensatory environments). Moreover, the confidence intervals corresponding to the means of noncompensatory and compensatory environments did not overlap (.72–.80 and .83–.86).

In matching the environmental model, participants did the best job, on average, in equal-weighting environments (average G of .91 vs. .83 in compensatory and .76 in noncompensatory environments; however, the confidence intervals corresponding to the equal-weighting and noncompensatory environments overlapped, .84–.98 and .74–.91).

Laboratory–field. We compared laboratory and field studies on two dimensions. First, how does performance compare in the two kinds of environments? Second, do the conditions of laboratory studies mirror those of field studies?

The dummy variable, laboratory–field, explained 11.4% of heterogeneity of r_a, 9.8% of the heterogeneity of G, and 14.7% of the heterogeneity of R_e. In our data, field studies contained more noise (i.e., smaller R_e) than did laboratory studies (R_e of .73 vs. .84; no overlap between confidence intervals: .69–.77 vs. .81–.87). Matching (G) was smaller in field than in laboratory studies (.69 vs. .84; no overlap between confidence intervals: .63–.75 vs. .80–.88). The same pattern was observed for achievement, r_a (.45 in field studies vs. .60 in laboratory studies; .54, adjusted for a possible availability bias). However, high remaining heterogeneity within the subgroups suggested that other moderator variables should be sought. In fact, all field studies contained three or more cues, whereas 72 of the 183 laboratory studies (39%) contained only two. Moreover, studies classified as field studies contained, on average, more inter-cue redundancy than did laboratory studies. In particular, none of our field studies lacked redundancy, whereas most laboratory studies (68%) had no redundancy.

To separate the effects of these three study-level variables, we fit meta-regression models with r_a and G as dependent variables (separately) and the laboratory–field dummy variable, number of cues, level of cue redundancy, and linear predictability of environments (R_e) as predictors. The results showed that the laboratory–field dummy variable explained no variance in either r_a ($B = -.01$, $SE_B = .04$, *ns*) or G ($B = -.02$, $SE_B = .05$, *ns*). The meta-regressions explained 28.3% of the variability of r_a and 18.7% of the variability of G. Thus, similar levels of achievement and matching can be reached in laboratory and field studies, which are comparable in terms of numbers of cues, inter-cue redundancy, and overall linear predictability of the environment.

Does environmental predictability (R_e) have the same importance for human judgment in laboratory and field studies? To answer this question, we fit meta-regression models separately for field and laboratory studies, with either G or R_s as dependent variables and R_e as the predictor. Because the laboratory–field variable was confounded with learning (1 if multiple blocks), expertise (two dummy variables), and context (1 if concrete), we added these predictors to the models. The coefficient for R_e was not significant in the regressions run on the subgroup of field studies (G: $B = .18$, $SE_B = .21$, *ns*; R_s: $B = .06$, $SE_B = .08$, *ns*). In laboratory studies, however, when environments were more predictable, participants were more consistent in their judgments: The coefficient for R_e was positive (.14; $SE_B = .06$, $p < .05$). When explaining the variance of matching (G) in laboratory studies, R_e was not significant ($B = -.08$, $SE_B = 10$, *ns*).

Context. Studies involving abstract tasks were characterized by a lower overall level of judgmental consistency than studies with concrete, contextually situated tasks (R_s of .76 vs. .82). The context dummy variable explained 3.8% of the variability of R_s. However, matching (G) was larger in abstract than concrete tasks (.84 vs. .79). The difference decreased if the figure for abstract tasks was corrected (until .80) for possible availability bias.

Expertise. Initial level of expertise (as opposed to expertise acquired through learning across experimental trials) was negatively related to achievement, r_a (5.0% of variability explained by the three-level categorical variable of expertise) and matching, G (8.7% of variability explained), but positively related to judgmental consistency, R_s (2.3% of variability explained). Notably, studies involving participants with some training and experts contained more noise than studies involving novices (R_e of .83 in studies with novices vs. .66 and .77 in studies with somewhat experienced participants and experts, respectively). In addition, linear cognitive ability (GR_s) and the validity of linear bootstrapping models (GR_e) decreased with increasing expertise (4.5% and 10.9% of the variability explained, respectively), but the nonlinearity/configurality coefficient (C) tended to increase in value with increasing expertise (albeit at an overall low level; 5.2% of variability explained).

To separate the effects of R_e and expertise, we fit meta-regression models, with achievement, matching, response consistency, and residual correlation as dependent variables and with R_e, two dummy variables characterizing levels of expertise, and the laboratory–field and learning variables (confounded with the level of expertise) as predictors. The results revealed that experts were more consistent in applying their linear judgmental policies ($B = .07$, $SE_B = .03$, $p = .05$, $n = 233$, in the regression with R_s as the dependent variable; the five variables and the intercept jointly explained 7% of the variability of R_s). However, there were no other significant effects. Matching, residual correlation, and achievement were not related to the level of expertise in these data.

[14] B refers to model coefficients; SE_B refers to standard errors of the coefficients.

Table 4
Meta-Regression Models of Lens Model Indices

	Dependent variable													
	r_a		G		R_s		C		GR_s		$GR_e - r_a$ (1)		$GR_e - r_a$ (2)	
Predictor	B	SE_B	B	SE_B	B	SE_B	B	SE_B	B	SE_B	B	SE_B	B	SE_B
Intercept	.16	.16	.76**	.17	.77**	.10	.25	.14	.58**	.17	−.19**	.07	.02	.08
R_e	.69**	.10	−.02	.11	.03	.07	−.17	.10	.04	.11	.16**	.04	.16**	.05
R_s													−.25**	.05
C													−.23**	.04
No. cues	−.02**	.01	−.03**	.01	.00	.00	.00	.01	−.03**	.01	.00	.00	.00	.00
Achieved cues	−.13**	.05	−.13*	.05	−.12**	.03	−.07	.05	−.17**	.05	.00	.02	−.03	.02
Low-medium cue redundancy	−.13**	.05	−.19**	.05	−.06	.03	.05	.05	−.18**	.06	−.03	.02	−.03	.02
High cue redundancy	−.10	.07	−.22**	.07	−.12**	.04	.13*	.06	−.24**	.07	−.07**	.03	−.07**	.03
Cue weights: noncompensatory	−.05	.04	−.05	.05	−.07*	.03	−.07	.04	−.08	.05	.01	.02	−.02	.02
Cue weights: equal	.09	.05	.02	.06	−.06	.04	.05	.05	−.03	.06	−.06*	.02	−.06*	.03
Nonlinear function	−.12*	.05	−.07	.05	−.04	.03	.03	.04	−.11	.06	.01	.02	.01	.02
Context: concrete	.05	.04	.06	.05	.01	.03	.12**	.04	.03	.05	−.01	.02	.01	.02
Some training	.12	.07	.04	.07	.01	.04	.03	.07	.06	.08	−.06*	.03	−.07*	.03
Expert	.08	.06	.01	.06	.04	.04	.08	.05	.06	.07	−.08**	.03	−.04	.02
Multiple blocks	.02	.05	.09	.06	.06	.03	−.09	.05	.13*	.06	.02	.02	.01	.02
Field study	.06	.05	.13*	.05	.04	.03	−.13**	.05	.13*	.06	.04*	.02	.03	.02
Log (judges × judgments)	−.02	.04	.06	.05	.01	.03	−.03	.04	.06	.05	.07**	.02	.07**	.02
τ^2 (unexplained variability)	.036		.040		.013		.024		.043		.005		.003	
% of variability explained	37.6%		41.3%		28.3%		—		28.5%		36.8%		55.3%	
n	173		166		166		141		167		166		140	

Note. See "The Mathematical Formulation of Brunswik's Lens Model" section of the text for a description of the lens model indices.
* $p < .05$. ** $p < .01$.

Learning. The possibility to learn over multiple trials substantially increased achievement, r_a (6.5% of variability explained by the two-level categorical variable one shot–multiple blocks); matching, G (10.0% of variability explained); and linear cognitive ability, GR_s (7.1% of variability explained). Interestingly, there was no difference in the values of judgmental consistency (R_s) between one-shot and multiple-block studies (0.2% of variability explained). Note, however, that multiple-block studies contained less noise than did one-shot studies (R_e of .84 and .71, respectively). Meta-regression models with R_e, two dummy variables characterizing levels of expertise, and the laboratory–field and learning variables as predictors (see above) revealed significant positive effects of learning on matching, G ($B = .15$, $SE_B = .05$, $p < .01$, $n = 233$) and on linear cognitive ability, GR_s ($B = .17$, $SE_B = .05$, $p < .01$, $n = 233$).

Simultaneous Effects of Moderators

To complement the above variable-by-variable analysis, we fit meta-regression models of judgmental achievement, its components, and composite lens statistics with all available moderators included (Table 4). There were 11 dummy predictors in the models: achieved cues; low–medium redundancy and high redundancy (for inter-cue redundancy); noncompensatory weights and equal weights (for cue weighting scheme); nonlinear function; concrete context; some training and expert (for expertise); multiple blocks (for learning); and field studies. The models also contained three continuous variables: number of cues, environmental predictability (R_e), and a measure ofy sample size: the product of number of judges and number of judgments made. A $\log_{10}$ transformation was applied to this variable because it had a pronounced positive skew.[15] The results of the meta-regressions can be summarized as follows.

First, achievement (r_a) was lower when there were more cues, when cues were achieved rather than given, when there was some redundancy between cues (vs. none), and when the function form in the ecology was nonlinear (37.6% of the variance explained by all included predictors).

Second, matching (G) was lower when there were more cues, when the cues were achieved, and when the cues were correlated (vs. not correlated at all). In addition, field studies reported higher levels of matching than did laboratory studies with the same characteristics on the dimensions included in the model. The model explained 41.3% of variance of G.

Third, judgmental consistency (R_s) was lower when cues were achieved, when inter-cue redundancy was high, and when the cue-weighting scheme in the ecology was noncompensatory (28.3% of the variance explained by all predictors).

[15] Exactly the same results occurred when this explanatory variable entered the models of r_a, R_s, G, C, and GR_s as a continuous untransformed variable, that is, the proportion of explained variance remained the same, and the same explanatory variables were or were not significant (the only exception was the nonlinear function dummy that became significant in the model explaining GR_s, $B = -.12$, $SE_B = .05$, $p < .05$). The differences among the models in the advantage of bootstrapping ($GR_s - r_a$) are discussed in the Results section, *What Factors Affect the Accuracy of Bootstrapping Models?*

Fourth, residual correlation (C) was higher (i.e., judges use omitted cues and/or integrate cues in a nonlinear or configural manner) when inter-cue redundancy was large and the task was context-specific. Field studies reported lower levels of C than did comparable laboratory studies. The model, however, hardly explained any variance of C.

Fifth, linear cognitive ability, GR_s, was lower under the same conditions as G, that is, under conditions of more cues, achieved cues, and correlated cues. This composite statistic was higher in studies that examined judgment over multiple blocks (vs. one-shot studies with comparable characteristics) and field studies (vs. laboratory studies with comparable characteristics). The model explained 28.5% of the variability of GR_s.

In all five models, the unaccounted residual variance exceeded 50%, suggesting that important moderator variables may be missing from this meta-analysis. Interestingly, the variables that define the level of expertise (some training and expert) were not significant in any of the models. Therefore, experts apparently do not perform better than less experienced judges on comparable tasks. We further discuss below the definition of expertise used to code the studies in this meta-analysis and the importance of distinguishing between experience and expertise.

How Effective Is Learning? The Role of Feedback

We next examined the effect of experience acquired through learning across experimental trials in studies with multiple blocks of judgments. Our objective was to quantify learning and to compare effects of different types of feedback. In our sample, 199 studies explicitly addressed the question of learning over multiple blocks of judgments (Table 2).

Main Effects

To quantify the magnitude of learning, we calculated the shifts in achievement, matching, judgmental consistency, residual correlation, and linear cognitive ability by subtracting the results observed in the first block of trials from those of the last block. Among 199 learning studies, 96 to 139 studies contained both pre-learning (i.e., the first block) and post-learning (i.e., the last block) data of various lens model statistics. On average, there were 176 learning trials between the first and last blocks of judgments in these studies (interquartile range = 40 to 125). Table 5 contained weighted means (calculated with random-effects models), 95% confidence intervals for the means, and various measures of between-study variability.

All lens model statistics improved over multiple experimental trials. Mean improvement of achievement (r_a) and matching (G) was .20, whereas judgmental consistency (R_s) and residual correlation (C) increased by .08. None of the confidence intervals contained zero. To calibrate these results, we compared them with average (weighted) values of lens model statistics corresponding to the first block of judgments (i.e., the first block of trials): on average, r_a increased by 48%, G by 31%, R_s by 11%, and GR_s by 42%. As for residual correlation, C, its pre-learning average value was .00 ($z = .01$). Notably, however, significant between-study heterogeneity remained in the average shifts of the lens model correlations after correcting for sampling error, that is, large and significant estimates of heterogeneity (Q); high percentages of variation attributable to between-study heterogeneity (I^2); and especially large DerSimonian and Laird estimates of between-study variance (τ^2), for achievement (r_a), matching (G) and linear cognitive ability (GR_s).

Types of feedback. Different types of feedback available to participants in the learning studies were examined as potential moderators of changes in the lens model indices. Outcome feedback on learning trials was provided in 94 studies (with both pre- and post-learning data available for at least one lens model index), task information feedback in 55 studies, cognitive feedback in 26 studies, and other types of feedback in 16 studies. However, in many studies a combination of different types of feedback was provided. Thus, 69 studies had outcome feedback alone, 14 studies had only task information feedback, and 5 studies had only cognitive feedback. Both cognitive and task information feedback were provided in 17 studies, and outcome and task information feedback were available in 21 studies. All three types of feedback were present in 3 studies, and, finally, participants in 1 study received both outcome and cognitive feedback.

Meta-regression models were fit to the shifts in the lens model indices (Table 6). Predictors included (a) pre-learning levels of human performance, as these can limit the space for learning; (b) the number of cues and predictability of the environment (R_e) as two measures of task difficulty (e.g., it could be more difficult to learn in noisy environments); (c) a $\log_{10}$-transformed measure of sample size (i.e., the product of number of judges and number of judgments per judge); and (d) the $\log_{10}$-transformed number of

Table 5
Descriptive Statistics of Learning Effects

Shift of:	M (weighted)	95% confidence interval	n	Q	I^2 (%)	τ^2
r_a	.20	.16–.25	139	5,968*	98	.070
G	.20	.15–.26	128	8,178*	98	.105
R_s	.08	.06–.11	127	1,303*	90	.017
C	.08	.04–.13	96	2,181*	96	.045
GR_s	.21	.16–.27	129	7,621*	98	.096

Note. See "The Mathematical Formulation of Brunswik's Lens Model" section of the text for a description of the lens model indices. Q represents within-group heterogeneity; I^2 is the percentage of variation attributable to between-study heterogeneity; τ^2 is the DerSimonian and Laird (1986) estimate of between-study variance.
* $p < .05$.

Table 6
Meta-Regression Models of Learning Effects

	Dependent variable—Shift of:																			
	r_a (M1)		r_a (M2)		R_s (M1)		R_s (M2)		R_s (M3)		G (M1)		G (M2)		C (M1)		C (M2)		GR_s (M1)		GR_s (M2)	
Predictor	B	SE_B	B	SE_B	B	SE_B	B	SE_B	B	SE_B	B	SE_B	B	SE_B	B	SE_B	B	SE_B	B	SE_B	B	SE_B
Intercept	−.07	.17	−.11	.17	.50**	.11	.44**	.11	.38**	.11	.47*	.19	.45*	.19	.28	.26	.12	.19	.19	.19	.14	.18
Prelearning level of the index	−.45**	.05	−.44**	.06	−.55**	.05	−.55**	.05	−.54**	.05	−.64**	.04	−.63**	.05	−.40**	.10	−.33**	.08	−.50**	.05	−.49**	.05
R_e	.30**	.07	.29**	.07	.09*	.04	.10*	.04	.10*	.04	−.02	.09	−.02	.09	−.19	.12	−.25**	.09	.09	.08	.10	.09
Number of cues	−.01	.01	−.01	.01	−.01	.01	.00	.01	−.01	.01	−.01	.01	−.01	.01	.00	.02	.01	.01	−.01	.01	−.01	.01
OF			.02	.04			−.02	.02	−.06	.03			.06	.04			.00	.05			.01	.04
TIF			.08*	.04			.07**	.02	.06**	.02			.09**	.04			−.03	.04			.13**	.04
CF			.01	.05			−.01	.03	−.02	.02			.02	.05			−.03	.05			.02	.05
Other type of feedback			−.05	.05			.05	.03	.04	.02			−.02	.05			−.11*	.05			.00	.05
OF prelearning block									.09**	.02												
OF last block									−.05	.03												
OF only	.00	.06			−.09**	.03					.00	.06			.06	.08			−.06	.06		
TIF only	.05	.06			.00	.04					.03	.06			−.05	.10			.04	.06		
CF only	−.06	.09			.01	.05					−.01	.08			−.14	.12			−.02	.09		
OF and TIF	.08	.06			.02	.04					.12	.06			−.09	.09			.10	.07		
TIF and CF	.07	.07			−.01	.04					.06	.06			−.06	.09			.07	.06		
Log(learning trials)	.07**	.02	.07**	.02	.00	.02	.00		.01	.01	.05	.03	.04	.03	.04	.03	.11**	.03	.06*	.03	.05	.03
Log(judges × judgments)	−.03	.04	−.03	.04	−.01	.03	−.02	.03	−.01	.02	−.02	.04	−.02	.05	−.08	.08	−.12*	.06	−.01	.04	−.01	.05
τ^2 (unexplained variability)	.024		.024		.006		.006		.004		.023		.024		.041		.022		.023		.023	
% of variability explained	65.3%		65.1%		66.9%		67.5%		74.0%		78.3%		77.5%		9.8%		51.0%		76.4%		76.3%	
n	135		131		124		120		120		125		119		95		93		124		120	

Note. See "The Mathematical Formulation of Brunswik's Lens Model" section of the text for a description of the lens model indices. OF = outcome feedback; TIF = task information feedback; CF = cognitive feedback. M1, M2, and M3 represent Models 1, 2, and 3, respectively.
* $p < .05$. ** $p < .01$.

learning trials (the untransformed distribution was heavily positively skewed).

For each lens model statistic, we fit two models. The first model, (indicated by M1) included three dummy variables identifying the exclusive availability of outcome feedback, task information feedback, or cognitive feedback. Two additional dummies identified studies with two available types of feedback: outcome and task information feedback, and cognitive and task information feedback.[16] The second model (indicated by M2) included four dummy variables that identified the availability (either exclusive or not) of outcome feedback, task information feedback, cognitive feedback, and other type of feedback.

Achievement. The results of the meta-regressions revealed that judgmental achievement (r_a) improved more in the studies that contained less noise in the environmental model (i.e., higher R_e) and a longer sequence of learning trials. In addition, achievement improved more in the studies where task information feedback was available, either alone or in combination with other types of feedback, than in the comparable studies (i.e., same number of learning trials, number of cues, and R_e) without such information: r_a (M2). The two models explained about 65% of the variability of shifts of r_a that occurred across learning trials.

Consistency. Similarly, judgmental consistency (R_s) increased more when R_e was higher and task information feedback (either alone or in any combination) was available: R_s (M2). There was no effect of the number of learning trials: R_s (M1) and (M2). Moreover, judgmental consistency decreased across learning trials in the studies that had only outcome feedback: R_s (M1). Interestingly, the intercepts were large and significant in both R_s (M1) and R_s (M2), indicating that judgmental consistency increased with experience, regardless of the type of feedback (if any). The models explained about 67% of the variability in shifts of R_s.

Overall, outcome feedback played a negative role in the dynamics of judgmental consistency (R_s). However, does it matter when outcome feedback is administered? There were 72 studies that provided outcome feedback both on the first (pre-learning) and last (post-learning) blocks of trials. In addition, in 8 studies outcome feedback was provided on pre-learning but not on post-learning trials, and in 19 studies this type of feedback was available on post-learning but not on pre-learning trials.

To understand the role of outcome feedback better, we added two new dummy variables to R_s (M2): one for studies with outcome feedback available on the first block of trials (outcome feedback, pre-learning block) and another for studies with outcome feedback available on the last block of trials (outcome feedback, last block). This new meta-model, Model 3, R_s (M3), showed that the moment at which outcome feedback becomes available is important. In particular, outcome feedback improved judgmental consistency when it was available on the pre-learning block, that is, when participants had the least knowledge about the task ($B = .09$). The coefficients for the dummies defining outcome feedback available on learning trials (outcome feedback) and on the last block of trials (outcome feedback, last block) were negative ($-.06$ and $-.05$) but only showed a trend toward significance ($p < .10$). Adding two dummies related to outcome feedback increased the proportion of variance explained by 6.5%.[17]

Matching. The models of the matching index (G) indicated that people were better able to acquire linear task-specific knowledge when task information feedback was available—either alone or in any combination— G (M2), 77.5% of variance explained. No other type of feedback was helpful. Moreover, noise in the environment was not a determinant of the magnitude of changes in G. Large and significant intercepts in the models of shifts of G indicated that experience improved matching across the conditions examined here, in a manner similar to consistency.[18]

Residual correlation. The increase due to learning of nonlinear or configural knowledge (C) was naturally stronger in environments with lower linear predictability: GR_s (M2), 51% of variance explained. No specific type of feedback was important for acquiring such knowledge; in fact, other types of feedback (e.g., relations between the ecology and the judge's strategy) were detrimental ($B = -.11$). Interestingly, the length of the learning period was positively related to improvements in C, but larger studies (as defined by the product of judges by the number of judgments per judge) reported smaller shifts.

Linear cognitive ability. Finally, the models explaining shifts of linear cognitive ability (GR_s) reflected a positive effect of task information feedback but no effects of R_e or of any other type of feedback: GR_s (M2), 76.3% of variance explained.

Summary. We found that, first, achievement, matching, consistency, and linear cognitive ability all increased with experience. Judgmental consistency was the least sensitive to learning. Second, less noise in the ecology facilitated improvements in judgmental consistency and achievement but was irrelevant for matching. Third, the availability of task information magnified the positive effect of (linear) learning. Fourth, outcome feedback was beneficial for judgmental consistency when the judge was unfamiliar with the task but detrimental when some familiarity had already been acquired. In addition, outcome feedback did not affect achievement, matching, or linear cognitive ability.

Other Effects

The data showed that some kinds of feedback can be helpful. However, is feedback helpful only in simpler tasks, for example, more predictable environments, situations with fewer cues? Does inter-cue redundancy help or hurt? And do experts learn from feedback better than novices? We next examined the role of each of these variables.

Environmental predictability and the number of cues. Because the median R_e in the studies with multiple blocks of trials was .91, we fit meta-regression models similar to those presented in Table 6 separately on the subgroups of studies with $R_e < .91$ and $R_e \geq .91$. The results of these models showed that task information feedback was only beneficial for achievement (r_a) and judgmental consistency (R_s) in the studies with $R_e \geq .91$ ($B = .15$, $SE_B = .07$, $p < .05$, $n = 69$, and $B = .11$, $SE_B = .04$, $p < .01$, $n = 59$,

[16] Dummies for identifying other combinations of feedback (i.e., outcome and cognitive; outcome, cognitive, and task information) were not included because of insufficient numbers of studies with such characteristics (1 and 3, respectively).

[17] A similar split in Model 2 of the other dependent variables had no effects on coefficients or on the explanatory power of the models.

[18] It should be noted, however, that higher pre-learning levels of matching were stronger impediments to learning than were higher pre-learning levels of consistency ($B = -0.64$ and $B = -0.55$ in G (M1) and R_s (M1), respectively).

respectively). In contrast, for matching (G), this was only the case in the studies with R_e <.91 ($B = .07$, $SE_B = .03$, $p < .05$, $n = 60$). Using the same .91 split, we found that outcome feedback provided on the first block of trials was beneficial for R_s only in the subgroup with $R_e \geq .91$ ($B = .11$, $SE_B = .05$, $p < .05$, $n = 59$). In short, there is some evidence that feedback is more beneficial when the ecology contains less noise. However, given the high level of R_e observed in our data, we believe that this should be considered more a hypothesis than a conclusion.

Among studies with multiple blocks of trials, there were 51 with two cues, 56 with three cues, and 32 with more than three cues. We fit meta-regression models, such as those presented in Table 6, separately on the subgroups of studies with two or three cues, on the one hand, and with more than three cues, on the other. These models showed that in the studies with more than three cues ($n = 30$), task information feedback had positive effects on r_a ($B = .16$, $SE_B = .07$, $p < .05$), G ($B = .17$, $SE_B = .08$, $p < .05$), and R_s ($B = 0.12$, $SE_B = .05$, $p < .01$). However, in the studies with fewer cues ($n = 90$), the only significant effect of task information feedback was on R_s ($B = .05$, $SE_B = .02$, $p < .05$). As for outcome feedback, its effect was significant only in the studies with fewer cues (positive if available in the first block of trials, $B = .08$, $SE_B = .02$, $p < .01$, and negative if available in the last block, $B = -.07$, $SE_B = .03$, $p < .05$).

Redundancy. Cue redundancy was absent in 74 studies with multiple blocks of trials, 45 studies contained some redundancy, and 8 had a lot of redundancy. Meta-regression models were fit separately on the subgroup of studies with and without redundancy. The results revealed that task information feedback improved r_a and G only when redundancy was present ($B = 0.12$, $SE_B = .06$, $p < .05$, $n = 45$, and $B = .15$, $SE_B = .06$, $p < .05$, $n = 39$, respectively) and improved R_s only when it was absent ($B = .10$, $SE_B = .03$, $p < .01$, $n = 69$). Similarly, the effect of outcome feedback on R_s was significant only in the subgroup of studies with no redundancy (positive if available in the first block of trials, $B = .12$, $SE_B = .04$, $p < .01$).

In brief, it seems that judges can use outcome feedback, either for the good or the bad of their judgmental consistency, only in tasks with few cues. Task information feedback, on the contrary, improved judgmental consistency, regardless of the number of cues. Moreover, inter-cue redundancy reduced the sensitivity of judgmental consistency to feedback. Providing task-relevant information increased linear task-specific knowledge (G) and overall judgmental achievement (r_a) only in tasks with more cues and more redundancy.

Expertise. Most studies with multiple blocks of trials were done with novices ($n = 127$). Only 12 studies involved somewhat experienced judges and experts. Therefore, one should be cautious in making general statements from comparing these two subgroups. The regressions showed that whereas task information feedback increased G in subgroups of novices as well as those consisting of more experienced judges ($B = .11$, $SE_B = .04$, $p < .05$, $n = 108$, and $B = .13$, $SE_B = .06$, $p < .05$, $n = 11$, respectively), its beneficial effect on R_s and r_a was significant only in the subgroup involving novices ($B = .07$, $SE_B = .02$, $p < .01$, $n = 109$, and $B = .10$, $SE_B = .04$, $p < .05$, $n = 120$, for R_s and r_a, respectively). Outcome feedback was irrelevant for the R_s of somewhat experienced judges, beneficial for the R_s of novices ($n = 109$) when available on the first block ($B = .08$, $SE_B = .02$, $p < .01$), and detrimental for the R_s of novices when available on the last block ($B = -.07$, $SE_B = .03$, $p < .05$).

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 is instructive to analyze the potential sizes of effects due to different variables on the basis of past samples of data. Our objective, therefore, is to isolate the conditions under which the application of bootstrapping is likely to be advantageous.

The average (weighted) advantage of bootstrapping models, found as the difference between linear achievement with perfect consistency and human achievement (i.e., $GR_e - r_a$) was .10, with a tight 95% confidence interval of .09 − .11 (Table 1). As might be expected, the effect of eliminating judgmental inconsistency more than outweighed the advantage of any residual correlation (C), and bootstrapping was, on average, effective (Camerer, 1981; Dawes et al., 1989; Goldberg, 1970). However, in 30 of 236 studies for which the bootstrapping advantage could be measured, GR_e was inferior to r_a, which suggests limitations in applying bootstrapping. High heterogeneity in the values of $GR_e - r_a$ (i.e., high Q and I^2 in Table 1) further highlights the importance of identifying the task and judge characteristics that favor bootstrapping.

Variable-by-variable analysis (Tables 2 and 3, last columns) revealed substantial differences in bootstrapping advantage between studies of different characteristics. Study-level variables that explained higher proportions of variance between the subgroups were the following: expertise (22.3% of variance explained), cue weights in the ecology (15%), number of cues (10.2%), and learning (9.9%). The average bootstrapping advantage differed from zero in all subgroups ($p < .05$). High heterogeneity is present in all entries of Table 2 and thus variable-by-variable analysis is insufficient.

A meta-regression model, similar to those described above for the lens model components, was fit with $GR_e - r_a$ as the dependent variable: Table 4, $GR_e - r_a$ (M1), with 14 predictors and an intercept. The model explained 36.8% of the variance in the values of bootstrapping advantage and showed that bootstrapping models are differentially advantageous in tasks with less noise (or larger R_e, $B = .16$). The advantage of bootstrapping is smaller when cues are highly correlated ($B = -0.07$) or equally weighted ($B = -.06$) and when judges have some experience ($B = -.06$ for some training and $B = -.08$ for expert). In addition, there was a positive effect for field as opposed to laboratory studies ($B = .04$).

The predictive power of the model was substantially improved by adding two additional predictors: judgmental consistency (R_s) and residual correlation (C): Table 4, $GR_e - r_a$ (M2), 55.3% of variable explained. Now it is apparent that the bootstrapping advantage is smaller when judges are more consistent in applying their linear policies ($B = -.25$), use cues not included in the model, and/or aggregate cues in an appropriate (i.e., corresponding to the ecology) nonlinear or configural manner ($B = -.23$). The

field study and expert variables, however, proved to be insignificant in this second model.[19]

Discussion

Our first goal in conducting this meta-analysis was to quantify levels of human judgmental achievement, as captured by lens model indices. Our second goal was to account for variability in performance due to individual and task characteristics (i.e., moderator variables). Our third goal was to illuminate the conditions that facilitate or impede learning, as studied within the lens model paradigm (i.e., by examining effects of different types of feedback). Our fourth goal was to illuminate the conditions under which bootstrapping models are more likely to be effective than unaided human judgment. Below we discuss conclusions reached, limitations of this particular meta-analysis, and the implications of our work.

Major Conclusions

Although subject to limitations (of generalization), the studies we examined essentially demonstrate three important findings: (a) People are capable of achieving high levels of judgmental performance, (b) people learn best from feedback that instructs them about the characteristics of the tasks they face, and (c) the inconsistency that people exhibit in making judgments is sufficient for models of their judgments to be more accurate than they are themselves (i.e., eliminating inconsistency outweighs the benefits of idiosyncratic knowledge that is not captured by linear models).

Our meta-analysis shows that evidence accumulated over the past 5 decades is consistent with the conclusion that linear models can provide good higher level representations of both human judgment and task environments (Einhorn et al., 1979; Payne et al., 1993). Linear models capture similar proportions of variance in environmental outcomes and human judgments (around 64% on average). However, there are clearly situations where human decisions are better described by nonlinear processes. For example, under time pressure, experts may rely more on intuitive judgment, consider few (if more than one) alternatives, simulate scenarios using imagination, and engage in experience-based pattern matching (Hogarth, 2001). The extent to which such tacit decision processes can be represented by linear models remains an open question.

In our sample, task environments in laboratory studies generally contained less noise than those in field studies that represent decision makers' natural ecologies. Moreover, laboratory—but not field—studies with high environmental linear predictability tend to report higher judgmental linear predictability (i.e., consistency). Parenthetically, we note that some 30 years ago, Brehmer (1976) showed, in a small sample, that environmental and judgmental linear predictabilities were positively correlated in field studies and claimed that this relation was also observed in laboratory studies. Our larger sample does not support the former claim.

We identify other differences between laboratory and field studies. For example, laboratory studies focus more on learning, use mostly novice (naïve) participants, and involve fewer cues with less inter-cue redundancy. We find that, even after controlling for task differences between laboratory and field studies, field studies tend to report higher linear matching, obtain lower residual correlation between the linear models of the judge and the environment, and suggest greater advantages of bootstrapping models over unaided human judgment. Therefore, extrapolating the results observed in laboratory studies to people's natural environments implies underestimating human linear knowledge and the advantage of bootstrapping models and overestimating nonlinear (configural) knowledge.

Consistent with the Brunswikian tradition, we identified several task characteristics that affect judgmental performance. First, when the number of cues is large, judges are less effective at matching environmental models. Consequently, levels of performance are lower (cf. Einhorn, 1971; Payne et al., 1993). Second, inter-cue redundancy also makes it more difficult to match environmental models and thereby reduces achievement (but see our arguments below as to why the lens model methodology may not be appropriate for capturing effects of redundancy). We also find that people respond to feedback more when redundancy is low. Third, identifying cues and quantifying their values is a hard task, and accuracy is better when cue values are directly supplied by investigators as opposed to being "achieved" by judges. Whereas there may be a relation between whether the cues are given or achieved and intrinsic interest in the task (Michael Doherty, personal communication, February 23, 2007), motivation to exert additional effort does not necessarily lead to better performance (Camerer & Hogarth, 1999; Pelham & Neter, 1995). Fourth, human achievement is lower when there are nonlinearities in the ecology. In addition, in environments that involve additive noncompensatory cue weighting (i.e., differentially weighted cues), consistency in applying judgmental policies is less than in environments with compensatory weighting schemes.

These results suggest that individuals may have preconceived, simplified expectations of decision environments and try to apply decision strategies that are 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. Perhaps, within the class of linear strategies, equal weighting is most attractive psychologically because it guarantees that the judge considers all information. It is also possible that the judge gives equal weights to the cues when he or she lacks knowledge about their differential validities. Equal-weighting strategies generally provide a good default (Dawes & Corrigan, 1974).

The presence of redundancy and differential cue weights creates an imbalance between individual expectations and environmental structure and, therefore, hurts performance. In the presence of such imbalance, many learning trials are needed for improvement to occur. The correct application of decision strategies that rely heavily on a single cue or a few cues (e.g., availability, Tversky & Kahneman, 1973; "take-the-best", Gigerenzer & Goldstein, 1996) requires a certain level of expertise (Hogarth & Karelaia, 2007).

[19] When the $\log_{10}$ transformation was not applied to the product of number of judges and number of judgments, the proportion of explained variance in both models remained the same. However, the significance of some coefficients was affected. In particular, in both models, the equal cue weights variable became marginally significant ($p = .057$); the variable, some training, gained in significance ($p < .01$); and the field study variable became significant ($B = .05$, $SE_B = .02$, $p < .05$).

Indeed, some studies report that experts use surprisingly little information (Goldberg, 1970; Shanteau, 1992b). Performance levels that can be achieved by novices when applying such strategies are limited, especially in environments where cue redundancy, and therefore cue interchangeability, is low.

With regard to expertise, our data do not suggest that experts match environmental models better than novices (cf. Shanteau, 1992b), nor do experts rely more than novices on configural, nonlinear judgmental strategies. Thus, both expert and novice judgment can be well described by simple linear models, although we do find that experts are more consistent than novices in applying their decision policies (cf. quasi-rationality hypothesis; Brehmer, 1994). However, the effect disappears when we control for a wider range of study-level characteristics. This may suggest that the conditions under which novice and expert judgments are studied are different and that the results of novices and experts are not directly comparable. The finding that more expertise does not lead to better judgmental achievement echoes a review by Dawes et al. (1989), in which the authors suggest that in some domains, expertise can be difficult to develop and is therefore only weakly related to performance criteria (see also Shanteau, 1992a). It is important to note that we coded as expert all studies where participants were described as being familiar with the task to be performed and/or to have made similar judgments repeatedly before. Therefore, although our expert judges were experienced, the extent to which they had developed domain knowledge—or expertise—is unclear. Moreover, if expertise is defined ex post, on the basis of judgmental performance, the participants classified as experts in our data should be referred to as "experienced" rather than "expert."

With regard to learning, we find that decision makers are capable of learning when they repeat a task over multiple trials. Judgmental consistency, however, is the least sensitive component of the lens model indices to learning. Our results show that more learning occurs in less noisy environments where there is greater consistency in applying individual decision policies. The learning of domain-specific linear knowledge, however, is not related in our data to the predictability of the task environment. With regard to feedback, it is task information that improves learning, whereas cognitive feedback does not help. Outcome feedback, when provided alone, has a negative effect on judgmental consistency and no effect on either matching or overall judgmental accuracy. The effectiveness of task information feedback has been emphasized previously in the literature (e.g., Balzer et al., 1989), whereas the effectiveness of outcome feedback has also been questioned (e.g., Brehmer, 1980; Hammond et al., 1973; Hoffman et al., 1981). However, we find that outcome feedback can be beneficial for learning (i.e., it improves judgmental consistency) when judges are unfamiliar with the task. Notably, none of the types of feedback considered in this analysis improves nonlinear (configural) knowledge.

Our meta-sample contains some evidence that decision makers are only able to use outcome feedback in tasks with few (and uncorrelated) cues. The effects of outcome feedback on judgmental consistency can be sometimes positive and sometimes negative. Task information feedback, on the contrary, improves judgmental consistency whatever the number of cues. We also hypothesized, on the basis of limited data, that task information feedback and outcome feedback affect only novice (naïve) judges. We do not find any evidence that more experienced judges benefit more from cognitive or task information feedback than do novices (Steinmann, 1976).

We identified several relevant task and judge characteristics that delimit the conditions under which bootstrapping human judgment is effective. The advantage of bootstrapping models over unaided human judgment is larger when there is less noise in the ecology; cue redundancy is low; cues are differentially weighted in the ecology; and the judge is less experienced, is less consistent in applying his or her linear policies, only uses the same cues as in the linear model, and does not rely on any nonlinear (configural) component. The latter, however, comes as no surprise, as bootstrapping models are linear models of judges.

What explains these findings? Following Camerer's (1981) analysis of the situations in which bootstrapping will be more effective than unaided human judgment, the roles of noise in the ecology, judgmental inconsistency, and any residual correlation are quite clear. However, the "vicarious" role of cue redundancy (Brunswik, 1952) can explain why bootstrapping models lose their advantage when redundancy increases. Cue redundancy reduces the importance of accurately detecting the most valid cues in the ecology, thereby making the cues interchangeable. As for cue weights in the ecology, the greater advantage of bootstrapping in environments where cues are differentially valid echoes our result that judgmental consistency is lower when these are noncompensatory. Finally, experienced judges may rely more on tacit, intuitive decision schemes that are difficult to capture by a statistical model such that their bootstrapping models are merely imperfect mirrors of the judges "corrected" for inconsistency.

In terms of more detailed findings, it is useful to consider what the data reveal about the effects of the different lens model statistics represented in Equation 3. The pattern of correlations between the various indices (Table 1) reveals that judgmental achievement is more strongly correlated with judgmental consistency than with environmental linear predictability (.56 vs. .43) and that whereas judgmental consistency is correlated with matching, environmental predictability is not. Thus, although environmental predictability necessarily limits achievement, in our data it explains little variance in differential achievement. Instead, this variance is more adequately captured by the human side of the lens model, that is, by the particular strategy the judge uses and how consistently this is executed. The result that matching and environmental predictability do not correlate echoes the Brunswikian idea of the overall ecological adaptability of human judgment processes (Brunswik, 1952).

Limitations and Further Research

Our meta-analysis was limited in that it did not consider several factors that could potentially affect judgmental accuracy. Indeed, we know there is variance in judgmental performance that is unexplained even after accounting for the factors coded in this article. For example, we did not consider possible effects of cue reliability as defined by variability in multiple observations of cues; York et al. (1987), for example, found that people weighted more reliable cues more heavily regardless of their validities. Nor did we consider the sign of cue validities (e.g., Lafon et al., 2004).

However, important limitations on our conclusions result from the fact that, taken as a whole, the 249 studies included in this analysis could hardly be described as being generated by principles

of representative design (Brunswik, 1955; Dhami, Hertwig, & Hoffrage, 2004). For example, most of the laboratory studies in our sample had little or no inter-cue redundancy, an important component of realistic task environments that was present in the field studies. Care should be exercised in generalizing from a research sample, as studies are often crafted to focus on specific cues, values, and ranges.

Interestingly, the presence of redundancy is an important component of Brunswik's (1952, 1955) 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—and assuming that people are always applying the same weights—the linear lens model methodology does not capture this aspect of how people may be processing information (cf. Ullman & Doherty, 1984). Thus, there is a need to develop methodology that can capture this dimension of behavior within a lens model framework and thereby lead to a better understanding of the effects of redundancy (for a discussion of alternatives to the lens model framework, see Cooksey, 1996, p. 331).

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 2 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.[20]

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 —this assumption is not appropriate (see also Holzworth & Doherty, 1976). For example, consider the relative importance of the two types of errors in medical decision making (e.g., Einhorn, 1972), predicting violent behavior of newly admitted inmates (Cooper & Werner, 1990), or predicting the number of annual advertising pages in a journal (e.g., Ashton, 1982). 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 average lens model statistics. The effect of some task variables, for example, inter-cue redundancy, could be more accurately assessed through within-judge analyses and, therefore, the conclusions reached from the aggregated data should be interpreted with caution. The data describing the "average judge" clearly limited our ability to comment on individual variability but reflect 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 also be useful, for example, to utilize multilevel hierarchical techniques (see, e.g., Raudenbush & Bryk, 2002) to understand simultaneous group- and individual-level behavior in lens model studies (Schilling & Hogge, 2001).

Castellan (1973, 1992) has provided an illuminating critique of the meaning of the matching index, *G*, in lens model studies, pointing out limitations in its interpretation due to mathematical constraints. In our 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), which 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 conditions favorable to human judgment. For example, the data we examined suggest that feedback is better assimilated in environments with less redundancy. It would also be interesting to investigate, in more detail, the interaction between the effects of feedback and expertise. The data in our sample are insufficient to draw any definite conclusion, although our preliminary investigation shows that experts may be more insensitive to outcome or task information feedback than novices. As another example, consider our finding 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 handling redundancy?

Parenthetically, we note that advances in technology can help greatly in collecting data within Brunswik's (1955) paradigm of representative design and in linking this with lens model analysis. Hogarth (2006) and Hogarth, Portell, and Cuxart (2007), for example, have exploited the messaging capacity of cell telephones to conduct experience sampling method studies of decision making and the perception of risk. In a similar vein, Mehl et al. (2006) have pioneered the use of the electronically activated recorder 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, researchers would be hard pressed not to find reasons to be optimistic about harnessing these and related technological developments.

Concluding Remarks

Experimental sciences—like psychology—advance in incremental fashion. New studies appear each year, often as a response to immediately preceding articles and what might be referred to as "local" issues (i.e., those that mark certain points in time). One can understand, therefore, why—at the level of individual studies—researchers have often adopted simple research designs involving only a few orthogonal cues. Thus, it is interesting to ask how studies might have been planned some 50 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 249 environments in advance?

[20] From our perspective, however, a problem with this work is that the unit of analysis has predominantly been that of aggregate (mean) judgments.

This question cannot be answered, of course, unless one first decides on the appropriate research questions. In broad terms, therefore, and drawing on hindsight, we observe that the main questions that dominated the research program centered not so much on "how good" people are at making judgments per se but on defining the task conditions that lead to differential levels of judgment, which, of course, includes 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, given that the lens model paradigm lends itself so well to the methodology, we hope that future research can take our analysis as a starting point.

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. Despite many lens model studies, it is not at all clear to which populations results should or could be generalized. For example, we found that laboratory studies differed from field studies on several dimensions and, therefore, laboratory results should not be blindly extrapolated to judges' natural habitats. More generally, this point also speaks to 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 has the potential to unlock many further insights about human judgmental processes.

References

*References marked with an asterisk indicate studies included in the meta-analysis.

*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.

Ader, H. J., & Mellenbergh, G. J. (Eds.). (1999). *Research methodology in the life, behavioural and social sciences.* London: Sage.

Anderson, N. H. (1981). *Foundations of information integration theory.* New York: Academic Press.

*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,* 385–392.

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

*Ashton, A. H. (1982). An empirical study of budget-related predictions of corporate executives. *Journal of Accounting Research, 20,* 440–449.

*Athanasou, J. A., & Cooksey, R. W. (2001). Judgment of factors influencing interest: An Australian study. *Journal of Vocational Education Research, 26,* 77–96. Retrieved from http://scholar.lib.vt.edu/ejournals/JVER/v26n1/athanasou.html

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, or functional validity information: Which components of cognitive feedback affect performance? *Organizational Behavior and Human Decision Processes, 53,* 35–54.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50,* 1088–1101.

*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,* 605–616.

*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,* 266–280.

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

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

*Brehmer, B. (1973b). Effects of task predictability and cue validity on interpersonal learning of inference tasks involving both linear and non-linear 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,* 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, 38,* 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, 26,* 54–64.

Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychological Review, 50,* 255–272.

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

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62,* 193–217.

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

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

Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty, 19,* 7–42.

Camerer, C. F., & Johnson, E. J. (1997). The process–performance paradox in expert judgment: How can experts know so much and predict so badly? In W. M. Goldstein & R. M. Hogarth (Eds.), *Research on judgment and decision making: Currents, connections, and controversies* (pp. 342–364). Cambridge, United Kingdom: Cambridge University Press.

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

Castellan, N. J., Jr. (1992). Relations between linear models: Implications for the lens model. *Organizational Behavior and Human Decision Processes, 51,* 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, 103,* 229–238.

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

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *The Behavioral and Brain Sciences, 4,* 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: Academic Press.

*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., 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,* 41–64.

*Cooksey, R. W., Freebody, P., & Wyatt-Smith, C. (2007). Assessment as judgment-in-context: Analysing how teachers evaluate students' writing. *Educational Research and Evaluation. 13,* 401–434.

*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,* 431–447.

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

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

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243,* 1668–1674.

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

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7,* 177–188.

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

*Djamasbi, S., Remus, W., & O'Connor, M. (2004). Does mood influence judgment accuracy? In Proceedings of the 2004 IFIP TC8/WG8.3 International Conference: Decision support in an uncertain and complex world (pp. 213–222). Retrieved from http://users.wpi.edu/~djamasbi/Djamasbi%20et%20al%20(DSS-Conference%202004).pdf

*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, OH: Bowling Green State University.

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

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56,* 455–463.

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

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315,* 629–634.

Einhorn, H. J. (1971). Use of non-linear, 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., & Hogarth, R. M. (1975). Unit weighting schemes for decision making. *Organizational Behavior and Human Performance, 13,* 171–192.

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

Field, A. P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods, 6,* 161–180.

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,* 398–412.

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,* 422–432.

Goldberg, L. R. (1976). Man versus model of man: Just how conflicting is that evidence? *Organizational Behavior and Human Performance, 16,* 13–22.

*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,* 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,* 127–132.

Hammond, K. R. (1955). Probabilistic functioning and the clinical method. *Psychological Review, 62,* 255–262.

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice.* Oxford, United Kingdom: Oxford University Press.

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 non-linear 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.

*Harvey, N., & Harries, C. (2004). Effects of judges' forecasting on their later combination of forecasts for the same outcomes. *International Journal of Forecasting, 20,* 391–409.

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

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3,* 486–504.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21,* 1539–1558.

*Hirst, M. K., & Luckett, P. F. (1992). The relative effectiveness of different types of feedback in performance evaluation. *Behavioral Research in Accounting, 4,* 1–22.

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. (2001). *Educating intuition.* Chicago: University of Chicago Press.

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, United Kingdom: Cambridge University Press.

Hogarth, R. M., & Karelaia, N. (2005). Simple models for multi-attribute choice with many alternatives: When it does and does not pay to face trade-offs with binary attributes. *Management Science, 51,* 1860–1872.

Hogarth, R. M., & Karelaia, N. (2007). Heuristic and linear models of judgment: Matching rules and environments. *Psychological Review, 114,* 733–758.

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). *Risk Analysis, 27,* 1427–1439.

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

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings.* Thousand Oaks, CA: Sage.

Hursch, C. J., Hammond, K. R., & Hursch, J. L. (1964). Some methodological considerations in multiple-probability studies. *Psychological Review, 71,* 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. (1996). On the reality of cognitive illusions. *Psychological Review, 103,* 582–591.

Kaufmann, E., & Athanasou, J. A. (2007). *A meta-analysis of judgment achievement in the lens model equation* [Working paper]. Mannheim, Germany: University of Mannheim; Sydney, Australia: University of Technology.

Kenny, D. A., & Albright, L. (1987). Accuracy in interpersonal perception: A social relations analysis. *Psychological Bulletin, 102,* 390–402.

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

Kleinmuntz, B. (1990). Why we still use our heads instead of formulas: Toward an integrative approach. *Psychological Bulletin, 107,* 296–310.

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

Kuo, Y.-Y., & Liang, K. Y. (2004). Human judgment in New York state sales and use tax forecasting. *Journal of Forecasting, 23,* 297–314.

*LaDuca, A., Engel, J. D., & Chovan, J. D. (1988). An exploratory study of physicians' clinical judgment: An application of social judgment theory. *Evaluation and the Health Professions, 11,* 178–200.

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

Law, K. S. (1995). The use of Fisher's *Z* in Schmidt-Hunter type meta-analyses. *Journal of Educational and Behavioral Statistics, 20,* 287–306.

*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,* 363–377.

*Levi, K. (1989). Expert systems should be more accurate than human experts: Evaluation procedures from human judgment and decision making. *IEEE Transactions on Systems, Man, and Cybernetics, 19,* 647–657.

*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,* 739–745.

Lindell, M. K., & Stewart, T. R. (1974). The effects of redundancy in multiple-cue probability learning. *American Journal of Psychology, 87,* 393–398.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

*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,* 561–587.

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

*Maras, M. (2007). *The disposition effect in the venture capital decision-making process: An experimental approach* [Working paper]. Universitat Pompeu Fabra, Barcelona, Spain.

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: Oxford University Press.

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

*Mear, R., & Firth, M. (1987). Assessing the accuracy of financial analyst security return predictions. *Accounting, Organizations and Society, 12,* 331–340.

Mehl, M. R. (2006). The lay assessment of subclinical depression in daily life. *Psychological Assessment, 18,* 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, 90,* 862–877.

Miklich, D. R., & Gillis, J. S. (1975). Interaction of age and cue validities in multiple-cue probability learning by children. *Psychological Reports, 37,* 235–240.

Miller, P. M. (1971). Do labels mislead? A multiple cue study, within the framework of Brunswik's probabilistic functionalism. *Organizational Behavior and Human Performance, 6,* 480–500.

*Muchinsky, P. M., & Dudycha, A. L. (1975). Human inference behavior in abstract and meaningful environments. *Organizational Behavior and Human Performance, 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, 3,* 47–61.

*Newton, J. R. (1965). Judgment and feedback in a quasi-clinical situation. *Journal of Personality and Social Psychology, 1,* 336–342.

*Nystedt, L., & Magnusson, D. (1973). Cue relevance and feedback in a clinical prediction task. *Organizational Behavior and Human Performance, 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.

Pelham, B. W., & Neter, E. (1995). The effect of motivation on judgment depends on the difficulty of the judgment. *Journal of Personality and Social Psychology, 68,* 581–594.

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, 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,* 22–30.

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

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52,* 59–82.

*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,* 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.

Russo, J. E., & Schoemaker, P. J. H. (2002). *Winning decisions.* New York: Doubleday.

Schilling, S. G., & Hogge, J. H. (2001). Hierarchical linear models for the nomothetic aggregation of idiographic descriptions of judgment. In K. R. Hammond & T. R. Stewart (Eds.), *The essential Brunswik: Beginnings, explications, applications* (pp. 332–341). New York: Oxford University Press.

*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.

Shanteau, J. (1992a). Competence in experts: The role of task characteristics. *Organizational Behavior and Human Decision Processes, 53,* 252–266.

Shanteau, J. (1992b). How much information does an expert use? Is it relevant? *Acta Psychologica, 81,* 75–86.

*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.

Sterne, J. A. C., Bradburn, M. J., & Egger, M. (2001). Meta-analysis in Stata™. In M. Egger, G. Davey Smith & D. G. Altmann (Eds.), *Systematic reviews in health care: Meta-analysis in context* (pp. 347–369). London, United Kingdom: British Medical Journal Publication Group.

*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,* 24–34.

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

*Strauss, R., & Kirlik, A. (2003). *A systems perspective on situation awareness II: Experimental evaluation of a modeling and measurement technique* [Tech. Report AHFD-03–13/NTSC-03–3]. Savoy Institute of Aviation, University of Illinois, Urbana Champaign.

Strube, M. J. (1988). Averaging correlation coefficients: Influence of heterogeneity and set size. *Journal of Applied Psychology, 73,* 559–568.

*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,* 203–211.

*Szucko, J. J., & Kleinmuntz, B. (1981). Statistical versus clinical lie detection. *American Psychologist, 36,* 488–496.

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

*Todd, F. J. (1954). *A methodological analysis of clinical judgment.* Unpublished doctoral dissertation, University of Colorado..

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.

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

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

Ullman, D. G., & Doherty, M. E. (1984). Two determinants of the diagnosis of hyperactivity: The child and the clinician. In M. Wolraich & D. K. Routh (Eds.), *Advances in behavioral pediatrics* (Vol. 5, pp. 167–219). Greenwich, CT: JAI Press.

Vazire, S., & Gosling, S. D. (2004). E-perceptions: Personality impressions based on personal Websites. *Journal of Personality and Social Psychology, 87,* 123–132.

*Werner, P. D., Rose, T. L., Murdach, A. D., & Yesavage, J. A. (1989). Social workers' decision making about the violent client. *Social Work Research & Abstracts, 25,* 17–20.

*Werner, P. D., Rose, T. L., & Yesavage, J. A. (1983). Reliability, accuracy, and decision-making strategy in clinical predictions of imminent dangerousness. *Journal of Consulting and Clinical Psychology, 51,* 815–825.

*Wiggins, N., Gregory, S., & Diller, R. (1974). [Unpublished raw data].

*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,* 100–106.

*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,* S5–S6.

*Wright, W. F. (1977). Financial information processing models: An empirical study. *The Accounting Review, 52,* 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 thy own self be true: Finding the utility of cognitive information feedback. *Journal of Behavioral Decision Making, 18,* 319–341.

Received February 1, 2007
Revision received December 28, 2007
Accepted December 28, 2007 ■