Journal of Mathematical Psychology **44**, 3–19 (2000)
doi:10.1006/jmps.1999.1274, available online at http://www.idealibrary.com on IDEAL®

Accuracy, Scope, and Flexibility of Models

James E. Cutting

Cornell University

Traditionally, models are compared on the basis of their accuracy, their scope, and their simplicity. Simplicity is often represented by parameter counts; the fewer the parameters, the simpler the model. Arguments are presented here suggesting that simplicity has little place in discussions of modeling; instead, the concept of flexibility should be substituted. When comparing two models one should be wary of the possibility of their differential flexibility. Several methods for assessing relative flexibility are possible, as represented in this special issue of the *Journal of Mathematical Psychology*. Here, the method of cross-validation is applied in the comparison of two models, a linear integration model (LIM) and the fuzzy-logical model of perception (FLMP), in the fitting of 44 data sets concerning the perception of layout seen among three panels with the presence or absence of four sources of information for depth. Prior to cross-validation the two models performed about equally well; after cross-validation LIM was statistically superior to FLMP, but the overall pattern of fits remained nearly the same for both models.

THEORIES, MODELS, AND PERCEPTS: ACCURACY, SCOPE, AND SIMPLICITY

Across the practice of science in general, one finds suggestions concerning the goals of scientific theories. A good theory, among other things, should be more accurate, broader in scope, or simpler than its competitors (Kuhn, 1977; Thagard, 1990). Similarly, across cognitive science one finds suggestions that a good model should be descriptively adequate, general, and only as complex as is necessary (Jacobs Grainger, 1994). Finally, to complete a set of near parallels, in the field of perception a perceptual system should be well attuned to external affairs, it should operate smoothly in different environments with different stimuli, and it has often been said that it should prefer "goodness" in a form over less simple alternatives. Thus, in all three domains there is a concern with *accuracy*, which seems to have a well-defined basis, or at least a well-measured one; with *scope*, which would seem

I thank Michael Browne, Malcolm Forster, Dominic Massaro, In Jae Myung, Mark Pitt, and two anonymous reviewers for their comments. Correspondence and reprint requests should be addressed to Department of Psychology, Uris Hall, Cornell University, Ithaca, NY 14853-7601. E-mail: jec7-cornell.edu. Fax: (607)255–8433.

0022-2496/00 $35.00

to have a logical basis, but is nonetheless ill-defined; and with *simplicity*, which often seems to have little more than an aesthetic basis. If one were to rank the importance of these goals, I am sure most theorists and researchers in each domain would agree that accuracy is most important, then scope, then simplicity. The extent of discussion of these in each field, however, probably follows the reverse order: Accuracy raises few hackles; scope and simplicity, on the other hand, are often fraught with conflict.

Scope and the Problem of Demarcation

The major problem with the idea of scope is that of demarcation: What lies within the legitimate *domain* of a theory, a model, or anything else, even perception? What lies outside? If one could find firm boundaries to such a domain, one might calculate area or volume or perhaps simply count the number of entities within it. Any of these could serve as measure of scope and then be used to compare theories, models, or whatever one was concerned with. However straightforward this idea may seem, it does not work well in practice. The structure of conceptual domains, of whatever kind they be, has been the major focus in a field of study called *categorization*. Traditional approaches to categorization have looked to the idea of necessary and sufficient features, defining the boundaries of domains or categories. If such features exist, an item lies within the category of interest; if not, it does not. Within psychology over the past 20 years (Rosch & Lloyd, 1978) and within philosophy over the past 45 years (Wittgenstein, 1953), however, there has been growing dissatisfaction with this approach. Instead, boundaries are now best thought to be fuzzy, even indeterminant. The foci of inquiry have been on examples within a domain, sometimes called exemplars; on domain centers, sometimes called prototypes; and on the shifting nature of similarity within a category, sometimes called family resemblance (see Smith, 1990, for a review). With the idea of fuzzy boundaries the notions of demarcation and scope become problematic. It may be that the only way to compare relative scopes is in situations where the domain of one category appears to lie entirely within the domain of another.

Given that this special issue of the *Journal of Mathematical Psychology* is on model comparison, let me give a hypothetical example concerning models. The comparison of the scope of two models involves those data sets that two models do and do not fit well. If Models A and B fit several sets about equally well, but Model B fits other data sets better, then Model B can be said to have more scope. Such a situation is suggested in the left panel of Fig. 1, where the scope of the reasonable fits by Model A is nested within that of Model B. Thus, Model B is the broader model. If, on the other hand, the data sets fit well by each model are not nested, as suggested in the right panel of Fig. 1, then comparison becomes much more difficult. The region of fit where A is superior to B may be smaller than that where B is superior to A, but it might be judged as more important. Moreover, even the situation of nested domains is subject to revision; before declaring Model B the superior model, the researcher must try to search for cases in which Model A is superior to Model B. This is probably best done by simulation.

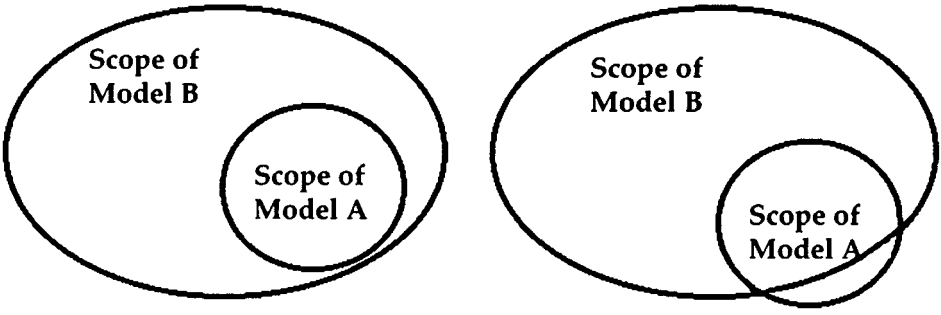

FIG. 1. Venn diagrams of the hypothetical scope of two models. The left panel shows nested domains for the two models, and the right panel shows nonnested domains. In the context of this discussion the relationships of the models LIM and FLMP might be analogous to those of Models A and B.

Simplicity, the Problem of Counting, and Other Issues

The concept of simplicity is even more slippery than that of scope. Interestingly, its discussion is much older in the field of perception than in mathematical modeling. Moreover, its problems may serve to elucidate the issues at stake. Mach (1906) discussed percepts as following a principle of economy, and Wertheimer (1920/1967) and Koffka (1935) gave us the idea of *pragnanz*, or of the goodness and simplicity of form. Midcentury interest in information theory led Hochberg and McAlister (1953) and Attneave (1954, 1982) to try to quantify goodness and simplicity. This line of inquiry culminated in the work of Leeuwenberg (e.g., 1971, 1982; see also Restle, 1979; Cutting, 1981). Leeuwenberg applied some ideas from algorithmic information theory (e.g., Chaitin, 1977; Komolgorov, 1968; see also Grunwald, 2000) to fashion his notion of a simplicity metric; that is, the representation of a preferred percept among many corresponds to a statement of the shortest length.

Consider the top figure in the left panel of Fig. 2. How does one perceive it? The interpretation in Configuration A of two similar panels, one behind the other, is almost universally preferred over the two alternative interpretations shown. Indeed, in a brief demonstration study done for this article, given the top figure 89 of 112 subjects chose Configuration A. It is also simpler than the others. In the spirit of Leeuwenberg, one coding might be

$$I_A = 4(S, \phi_{90}) \text{ applied to } [x_1 y_1, x_2 y_2]\ \theta_{90} = 8, \tag{1}$$

where I_A is the information load of Configuration A (its simplicity metric); (S, ϕ_{90}) is the sequence of operations for drawing a line segment S and then rotating the direction of the pen's path through ϕ_{90} (or a right angle drawn to the right); 4 is a statement of iteration that this set of operations is done four times, closing the figure; $[x_1 y_1]$ is the starting location for drawing the first square from the upper-left corner; $[x_2 y_2]$ is that for the second square (again, upper-left corner); and θ_{90} is the initial angle for drawing the first pen stroke (directly to the right). By Leeuwenberg's approach the information load for this array might be 1

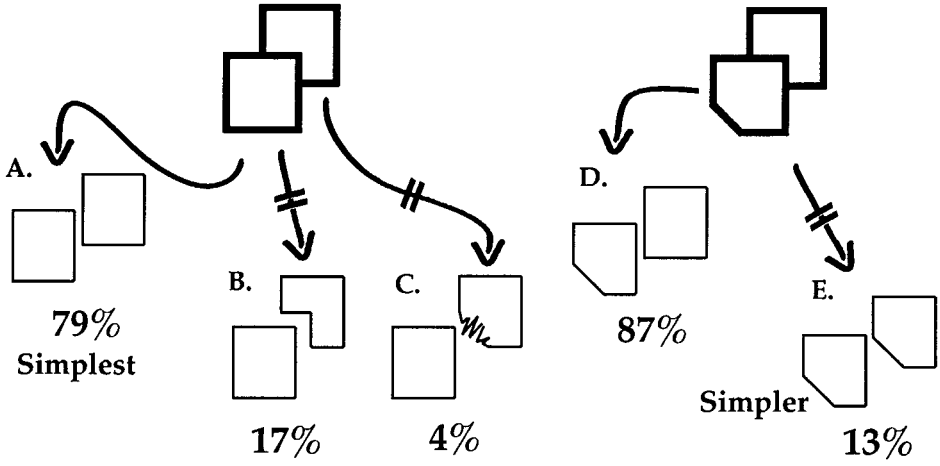

FIG. 2. One hundred twelve undergraduates were shown the ambiguous figures at the top of both panels and asked to interpret what they saw. Given the three options at the left, 89 chose Configuration A, 19 chose Configuration B, and only 4 chose something like Configuration C. Leeuwenberg (1971), among others, would claim this is because Configuration A is simpler. But simplicity metrics are problematic. These viewers preferred Configuration D (97 of 112) over Configuration E (15 of 112), even though the latter can be coded as simpler.

(the statement of iterations) + 2 (the drawing primitives) + 5 (the four initial coordinates and one initial direction for the drawings) = 8.

One alternative interpretation of the display shown in Configuration B is the concatenation of a square with an inverted L-shaped hexagon. Such a figure might be coded:

$$I_B = 4(S, \phi_{90}) \text{ applied to } [x_1 y_1 \theta_{90}],$$
$$\text{and } [(2(S, \phi_{90}), 2(s, \phi_{90}), (s, \phi_{270}), (s, \phi_{90})] \text{ applied to } [x_2 y_2 \theta_{90}] = 19 \quad (2)$$

where the first line of (2) is almost the same as in (1) and the second line contains the new symbol s for a shorter segment half the length of S and ϕ_{270} for a right angle drawn to the left. The information load for this display might be 6 (for the square in the first line) + 13 (for the hexagon) = 19. Clever choices of primitives and clever schemes of rewriting this statement might reduce the load, but because the redundancy of squares creates a relatively compact code, it seems unlikely that the load could be reduced to 8. And indeed only 19 of 112 subjects chose this configuration. Finally, another interpretation of this figure is as the occlusion by the front square of a figure with a serrated edge beneath the occluding contour shown in Configuration C. I will not attempt to code this interpretation, but simply rely on intuition that it would not have an information load as low as 8 or even 19; and only 4 of the 112 subjects chose this version.

On other theoretical accounts, these two nonpreferred alternatives would be discussed in terms of *accidental* and *nonaccidental properties* (Witkin & Tenenbaum, 1983). That is, the abutment of edges in the first nonperceived case would seem to be an accident of the selection of forms and their position, not likely to occur through any means except by special intervention of a designer. Similarly, in the

second, the concealment of differences between the two objects and the nonconcealment of their similarities also would seem to be the result of intervention rather than accident. This kind of explanation invokes a likelihood principle (Helmholtz, 1867/1925; Hochberg, 1982) rather than a simplicity principle (Leeuwenberg & Boselie, 1988). Leeuwenberg's approach, however, only takes into account the length of the logical statement necessary to write a code of the figure. Notice that these codes represent *models* of what could be perceived and the model for two squares has the fewest parameters.

Work in philosophy and in perception, however, has shown that simplicity metrics are problematic. Goodman's (1972) critique focused on the universe of primitives chosen to code a verbal statement. Here we can generalize to visual forms as in Fig. 2. The problem is that there is no universal alphabet of form components in vision. Thus, there are no domain-specific constraints on the choice of primitives. With a cagey choice of primitives one can show equal simplicity for any pair of figures or even greater simplicity for almost any figure over another. For example, if one allowed squares and L-shaped hexagons as primitives, the first two interpretations in Fig. 2 would have the same information load and would be equally simple. Thus, the outcome of any information load (or counting of primitives) is dependent on what primitives are allowed. The most principled applications of this approach to simplicity to perception are those with primitives chosen from another field, such as physics (Restle, 1979).

Most critically, however, Hochberg (1986) argued that simplicity often does not even predict what is seen. Consider the figure in the top-right panel of Fig. 2. Configuration D might be coded

$$I_{\mathrm{D}} = [\,2(S, \phi_{90}), 2(s,\phi_{45}), (s, \phi_{90})\,] \text{ applied to } [\,x_1 y_1 \theta_{90}\,] \text{ and } 4(S, \phi_{90}) \text{ applied to } [\,x_2 y_2 \theta_{90}\,] = 17, \tag{3}$$

where the length s has been adjusted a bit. Configuration E might be coded

$$I_{\mathrm{E}} = [\,2(S, \phi_{90}), 2(s, \phi_{45}), (s, \phi_{90})\,] \text{ applied to } [\,x_1 y_1, x_2 y_2\,]\, \theta_{90} = 13. \tag{4}$$

Although, by this construction, Configuration E is simpler, only 15 chose E while 97 chose D. One might quibble with this analysis and suggest that each perceived figure must be written separately. This would readjust some of the information loads: $I_{\mathrm{A}} = 12$ and $I_{\mathrm{B}} = 19$, with Configuration A retaining its advantage over B and $I_{\mathrm{D}} = 22$ and $I_{\mathrm{E}} = 17$, with Configuration E now being the simpler. But this kind of patchwork will not solve all problems.

For example, imagine a rotating wire-frame cube seen in perspective. It is multistable. One interpretation is the simplest, a true cube in rotation. One can easily imagine a Leeuwenberg code for such a figure. But often perceivers, particularly more experienced ones, prefer to see a prismatic six-sided figure that continually changes shape as it rotates. No simplicity metric can account for this phenomenon; it seems to be a clear exception to the notion that perception prefers simpler objects. Moreover, it provides fodder for a discussion of whether we should even prefer simpler theories or simpler models.

With these ideas about scope and simplicity from outside the considerations of modeling, let me now turn to the task at hand–using this larger framework for understanding some tensions in choosing among models. In experimental psychology and in many other fields one's style of experimentation and evaluation can proceed on two fronts. The first and most common is simple hypothesis testing. Data are generated and patterns are tested against the null hypothesis. Although it is standard fare in our field, not everyone is pleased with this approach (see, for example, Shrout, 1997); many claim that the null hypothesis is always wrong, regardless of what experimental hypothesis one might cherish. Second, and much more pertinent to this special issue, one can compare two models. This frees one from the vagaries of null hypothesis testing, but the inferential road is not always smooth.

THE FLEXIBILITY OF A MODEL

The standard conceptualization in our field, I would claim, is that shown in the top of Fig. 3. We start with the data given to us from our experiments. Conceptually we divide the data into a patterned portion and a portion that is residual of that pattern. We typically call the latter *noise*. The pattern is carved out of the data by the model, and the residual noise is measured as the lack-of-fit. When two models, say Models A and B, are fit to the data, the model with the smallest residual is typically declared the winner. Thus, if $noise_{\beta}$ is smaller than $noise_{\alpha}$, then we say that Model B fits better. Typically, care is taken that the two models are comparable. In particular, it is often said one should compare models with the same number of parameters. Counting parameters is typically thought to yield something akin to a measure of simplicity; the simpler model is the one with fewer parameters.

Given this view, there is an inherent tension between scope and simplicity. One the one hand, the experimenter wants his or her model to fit, and fit well, as many different data sets as possible. This, of course, could be done simply by adding more and more parameters. On the other hand, the experimenter wants his or her model

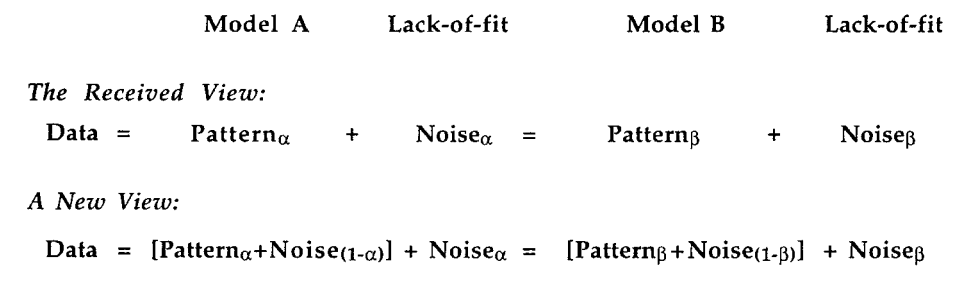

FIG. 3. The top panel shows a simplified view of a problem of induction in science, partitioning data into pattern and noise. We measure the goodness of fit of the model by the size of the noise; the less noise, the better the fit. However, a more complicated view proposed here suggests that all models incorporate some of the noise into the fit, leaving less noise as residual. When two models are compared, the classical approach only measures the residual noise not fit by the models. I claim that the goal of this special issue of the *Journal of Mathematical Psychology* is the discussion methods for considering both types of noise. (This figure shows the two types of noise to have a zero-sum relation, but this is a simplification that may not be the case.)

to be as simple as possible. This typically means that the number of parameters should be limited. Conceived in this manner, scope will always trade off with simplicity.

I claim, however, that a concern with this scope/simplicity trade-off is ill-considered. Simplicity is not a useful issue in this context; instead we should consider *flexibility*. Models can be differentially flexible, one accommodating more noise in data when it should not. The classic example of this concerns nested models, where the parameters of one model are a subset of those in another. The model with the greater number of parameters will necessarily fit data better than its simpler, derivative mate. The case I will consider later in this article, however, concerns nonnested models with the same number of parameters.

How I conceive of flexibility is shown at the bottom of Fig. 3, an unofficial elaboration of the received view. When, for example, Model A is fit to a set of data, I claim it partitions the noise. Some of the noise ($noise_{(1-\alpha)}$) will be fit by the model, incorporated into its pattern; other noise ($noise_{\alpha}$) will be left as residual. Since different models carve different patterns out of the data, the noise will be partitioned differently by the two models. The more flexible model will absorb more noise into its pattern and fit the data better. Thus, if Model B is more flexible than Model A, Model B will always have an advantage in model comparisons because $noise_{(1-\beta)}$, a measure of its flexibility in a particular context, is larger than $noise_{(1-\alpha)}$. The goal of this special issue is to present techniques developed in various disciplines to penalize a more flexible model appropriately, in proportion to its flexibility relative to the other model. According to Fig. 3, this should be something like $[noise_{(1-\beta)} - noise_{(1-\alpha)}]$ for a given data set.

A CASE STUDY IN THE COMPARATIVE ACCURACY, SCOPE, AND FLEXIBILITY OF TWO NONNESTED MODELS

There is a sense in which this symposium was my fault, or at least my fault that it occurred in the context of psychology. Thus, let me provide some background. I am a part-time psychophysicist interested in the visual perception of real-world phenomena. This makes me a global psychophysicist in the tradition of Gibson (1950). By habit, I manipulate relatively complex variables in both factorial and regression designs, and I search for the relative utility of multiple sources of information in a single perceptual task. Although the study of the comparative utility of multiple sources of information is common in the fields of decision making (Kahneman, Slovic, Tversky, 1982), social psychology (Nisbett & Ross, 1980), and speech perception (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967 Massaro, 1988), it has until relatively recently not been common in the field of visual perception. The reason for this may be the influence of the later work of Gibson (1979), with his emphasis on invariants, or one-to-one correspondences between information in the visual array and states of the physical world. Gibson proposed that single sources of information (invariants) could account for each percept. If this were true there would be little need to consider multiple sources and how they are integrated. On this view, they simply are not. But it seems to me this

idea is wrong; perception normally appears to function on the basis of multiple sources of information. To be sure, invariants have been found, but they probably number less than two dozen. With an army of researchers focused on their discovery for 20 years one might have thought they would be easier to find. Thus, I claim invariants are likely to be insufficient in number to account for the bulk of perception (Cutting, 1993).

This current state in the field of visual perception, and its focus on unitary sources of information, is a bit ironic. The oldest endeavors in psychology, and indeed among the oldest endeavors in all of science, are in trying to understand how we perceive depth and the relative layout of objects around us. Euclid (Burton, 1945), Leonardo (Richter, 1883), Helmholtz (1867/1925), Carr (1935), and the earlier Gibson (1950) are still very much worth reading in this regard and all listed many contributing sources of information. Moreover, virtually every contemporary textbook dealing with the perception of space, be it a general introduction to psychology or one specific to the field of perception, lists five or more sources, typically called cues to depth. One such list (Cutting Vishton, 1995; Cutting, 1997) includes occlusion, height in the visual field, relative size, relative density, aerial perspective, accommodation, convergence, binocular disparities, and motion perspective. Others include shadows, texture gradients, linear perspective, accretion and deletion of texture, kinetic depth, and gravity. Thus, if all of these were taken as independent, there are at least 15 sources of visual information about depth.

When lecturing on this material, I have been struck by questions from the more discerning students: "Which of these sources is more important?" and "How do these sources go together?" Despite millennia of thinking about these issues, centuries of listing information sources, and decades of intensive research, these two questions were largely unaddressed. Thus, through surveying the literature, we made some attempt at answering the first query (Cutting & Vishton, 1995), and through research we began investigating the second (Bruno & Cutting, 1988; Cutting, Bruno, Brady & Moore, 1992; see also Landy, Maloney, Johnston, & Young, 1995).

Addressing the issue of information integration is not particularly easy in the area of depth and layout perception. For example, if as many as 15 information sources are considered, all listed above, a complete factorial design would necessitate at least 2^{15} (or about 32,000) stimuli. Moreover, if one were really interested in potential interactions one should use 3^{15} (or about 14 million) stimuli. In the latter case, viewing one token of each of these stimuli would take about 30 years. And even if the complete design were run with a number of observers and an analysis of variance carried out, there would be 70,000 interactions to inspect, some 3,500 of which might be statistically reliable by chance.

Thus, we began less grandly. Bruno and Cutting (1988) generated an array of stimuli containing three square panels in depth. We then orthogonally manipulated four sources of layout information–relative size, height in the visual field, occlusion, and motion perspective among those panels–yielding 24, or 16 stimuli. That is, each source of information either was or was not present (but see Massaro, 1988). Employing 10 naive observers we conducted three experiments, two of indirect

scaling (where observers compared all possible pairs of stimuli, and the half-matrix of those relative judgments was then used as input to a scaling algorithm) and one of direct scaling (where observers produced numbers for each stimulus they felt corresponded to the magnitude of the variable investigated, in this case depth). The indirect scaling involved a preference judgment task (which of the two stimuli revealed more exocentric depth or distance among the three panels) and a dissimilarity judgment task (how dissimilar the two stimuli were in the depth that they revealed). Case V Thurstonian scaling (e.g., Dunn-Rankin, 1983) takes preference judgments and produces a one-dimensional solution measured in terms of standard deviations, and the scaling results reveal a good fit. Metric and nonmetric multidimensional scaling (e.g., Shepard, 1980) procedures take the dissimilarity ratings and convert them into solutions of N dimensions, as chosen by the investigator. The better the fit in the fewer dimensions, the more confident one can be that the solution captures the patterns in the data. We found that a one-dimensional solution fit the data quite well. These scale values for the 16 stimuli were then used as dependent measures in a regression analysis, and we found that the presence or absence of the four sources of information accounted for 92% of the variance in the scaled data. Finally, the direct scaling task asked observers to rate the depth seen in each of the stimuli, using a scale from 1 to 99. Regression analyses on the values of the 16 stimuli yielded results similar to those of the indirect methods; results were linear and apparently additive.

Massaro (1988) was unconvinced by our claims of additivity and he asked us for our direct scaling data. We obliged, and he tested the linear integration model, or LIM against his fuzzy-logical model of perception (FLMP, Massaro, 1987). He found that the data of five of our observers were better fit by LIM, and five were better fit by FLMP. Cutting and Bruno (1988) replied, suggesting in part that the data of the two indirect scaling tasks revealed different patterns. However, later analyses proved to generalize Massaro's (1988) claim: Regardless of which of the three tasks was examined, the data of half of the subjects were better fit by LIM and half by FLMP (although not always the same halves).

This lack of resolution festered in our minds. We were interested, in large part, because this was the first case we knew of in which two nonnested models with the same number of parameters were compared with an indeterminate outcome. Cutting *et al.* (1992) then reported a new set of studies focused on trying to falsify the notion of additive integration (LIM). First, we argued that the original stimulus set offered four uncorrelated information sources, all combinations equal in frequency, and that the lack of correlation within the set might have fostered perceptual independence (Ashby & Townsend, 1986; Garner, 1966; Nosofsky, 1991) and hence additivity in judgments. Thus, we created several sets of stimuli with correlated information sources. Second, we argued that linearity in judgments could have been an artifact of the range and frequency of the stimuli (Parducci, 1964, 1975). Were the stimulus set to be anchored with many more stimuli at one end than the other, judgments might reflect range and frequency considerations. Thus, we created stimulus sets with prominent anchors. Interestingly, neither of these manipulations was effective in changing direct scaling judgments, which remained quite linear.

More importantly, we also fit the data of 44 subjects to the two models, LIM and FLMP. The linear model seems like simplicity itself,

$$D = S + H + O + P + B, \tag{5}$$

where D equals perceived distance, in this case a number (a direct scaling judgment) from 1 to 99 representing the amount of depth seen in a given display; S, H, O, and P are the relative contributions (weights) of relative size, height in the visual field, occlusion, and motion perspective to the judgments, when present in the stimulus; and B is a background variable that includes the flatness of the computer monitor, reflections from it, and anything else in the experimental situation. Thus, LIM is a simple statement that perceived depth is the sum of the contributions of each source of information, with no interactions. FLMP, on the other hand, is a multiplicative model and does not appear nearly so simple,

$$D = SHOPB/[SHOPB + (1-S)(1-H)(1-O)(1-P)(1-B)], \tag{6}$$

where the parameters have exactly the same meaning as in LIM, but with additional constraints: The data are transformed and the weights constrained so that $0.0 < D, S, H, O, P, B < 1.0$. In addition (1-$S$) rather than S, for example, is the weight when relative size information is absent from a particular stimulus. Thus, FLMP in this context is a statement that perceived depth is enhanced by each source to the degree that the other sources do not contribute. Although this model can approximate additivity quite well, it really shines when certain nonlinearities occur. Massaro and Friedman (1990) have shown that this model is isomorphic with a Bayesian model. Notice that these two models are nonnested (that is, one is not contained within the other) and that both have five parameters—S, H, O, P, B.

Cutting *et al.* (1992) found that, of the 44 participants, the data of 23 were better fit by LIM and the data of the other 21 were better fit by FLMP. Since the general approach of Massaro has been to assume that the data of all subjects should be fit by the same model, we sought some resolution. But a clear resolution was still not at hand, if one was possible. This general deadlock provoked us to investigate the properties of the models and how they fit data. At issue were our impressions that, despite the fact that both models had the same number of parameters, somehow FLMP was more flexible. Thus, we undertook a series of simulation studies.

We first generated two families of data sets. One expressed exponential functions, $D = n^a/4^a$, where a could take values from 0.01 to 100, n is the number of sources in a given stimulus, and 4 is the number of possible sources of information present. The other expressed psychometric functions, $D = e^{(a+nb)}/[1 + e^{(a+nb)}]$, where $e = 2.718$, n is the number of sources of information in a given stimulus, and a and b are the two parameters for a logit function. We fit both models to each data set and found that within the area covered by the exponential functions, FLMP fit the data better than LIM in 80% of the cases and that within the area covered by the psychometric functions, FLMP fit better in 96%of the cases. Moreover, when LIM provided a better fit it did so by a narrow margin, often only in the third significant figure of the loss function but when FLMP provided a better fit it typically did so

by a very wide margin, often in the first significant figure. This is a powerful set of results in favor of FLMP; exponential and psychometric functions are exactly the kinds of functions that one would generally wish to capture in these types of experiments. Harkening back to the earlier discussion, FLMP clearly has more scope than LIM and the pattern shown in the left panel of Fig. 1 seemed applicable. But we worried; FLMP still did not seem as simple as LIM, and it seemed too flexible.

Next, we fit the two models to random data, a procedure commonly used to benchmark multidimensional scaling algorithms (De Leeuw & Stoop, 1984 Klahr, 1969; Levine, 1978; Wagenaar & Padmos, 1971). However good the first set of results seemed in support of FLMP, the second set seemed equally bad: FLMP fit 61% of the random data sets better than LIM. Although Myung and Pitt (1997) seemed to endorse the spirit of this analysis, Massaro and Cohen (1992) suggested that these "fits" were irrelevant and uninterpretable since most were unacceptably large, and Li, Lewandowski, and DeBrunner (1996, p. 361) regarded this procedure as "inelegant." I agree with both criticisms. Nonetheless, this result seemed important in reflecting the relative flexibility of FLMP compared to LIM. We felt that a model that performed better than another in fitting random data, however badly, was doing something beyond what a psychological model should do. Here was what we thought to be unwarranted scope due to excess flexibility. If LIM was considered Model A in Fig. 3, and FLMP was Model B, what were the relative sizes of $noise_{(1-\alpha)}$ and $noise_{(1-\beta)}$?

We struggled with how to account for this phenomenon and conducted a third set of simulations. Taking a linear set of data, we added increments of noise. Of course, with perfectly linear data LIM will always fit better, but with increases in noise added to the data FLMP began to fit a larger and larger proportion of the simulated data sets. Thus, in the context of our particular design of 16 stimuli, when a value between ± 0.11 randomly generated in a flat distribution was added to linear scores between 0.30 and 0.70, FLMP fit the data better than LIM more than 50%of the time. Myung and Pitt (1997) performed the reverse operation, adding random noise to functions generated from FLMP, and LIM never overtook FLMP. Such results suggest that FLMP is considerably more flexible than LIM.

What to do? Cutting *et al.* (1992) made two suggestions. First, we suggested that FLMP should be penalized by the degree to which it fit random data. Thus, we found that the data of 23 viewers were better fit by LIM by consideration of the random simulations and fits of the two models, we would have expected only 39%, or the data of 17.2 viewers, to have a better fit by LIM. By binomial expansion, this still did not yield a reliable difference ($z = 1.63$, $p > 0.10$). Gratifyingly, the core idea of a flexibility penalty turned out to be correct, as shown in these proceedings of the symposium.

Second, we thought that FLMP's greater flexibility might derive from its lack of simplicity, measured by description length. We had borrowed this idea from the interesting, if not completely convincing, analyses of Leeuwenberg (1971, 1982) as discussed above. Our understanding of minimum description length as it has developed in theoretical computer science, however, was flawed; Grunwald (2000) gives a tutorial. Moreover, Massaro and Cohen (1992) pointed out that if all of the

data were *z*-transformed, the appearance of FLMP would be much simpler than LIM. Not knowing of developments in the methodology of model comparison, this stopped us dead in our tracks.

REVISITING THE CASE STUDY

The symposium proved enormously stimulating. Four routes to the general comparison of models were considered. Grunwald (2000), aside from correcting my misunderstanding of the principle of minimum description length, further convinced me that in the context of comparing psychological data this method is still developing and may not soon prove fruitful. The other three approaches, however, seemed to offer more hope and avoid Goodman's (1972) problem of measuring simplicity by counting primitives. First, with respect to how one might correctly generate an appropriate penalty term for a more flexible model, Bozdogan (2000) was perhaps most convincing in his consideration of his informational complexity criterion (ICOMP) and the correlational structure of parameter estimates. However, because time was short in preparing this article, it was necessary to forego this type of analysis of our data. Second, Wasserman (2000) presented the outlines of the Bayesian approach to model comparison, amplifying the work of Myung and Pitt (1997). This, too, initially seemed promising because I had thought that our simulations of model fits to random data could be used as a priori probabilities. However, pursuing the logic of the analysis, I soon realized that this too would not be straightforward.

I found great solace, however, in the ideas of cross-validation presented by Browne (2000). Here was a scheme that seemed useful and also extremely similar to what I had already done with my colleagues. Rather than have FLMP and LIM fit all of our viewers' data, I would fit them to half the data, then use the weights obtained to fit the other half of the data with no free parameters. If FLMP were an overly flexible model, its fits to the residual half of the data should be relatively worse than those of LIM. This method effectively pits the magnitudes of $noise_{(1-\alpha)}$ against $noise_{\alpha}$ and $noise_{(1-\beta)}$ against $noise_{\beta}$.

Approach and Method

Since we had generally obtained only 10 observations per stimulus per viewer, it initially seemed possible to compare all possible halves of the data, using the fixed order in which the data accumulated during the experimental session (there are only somewhat more than 200 possible data halves). However, the constraints of the statistical package most easily available to me (the NONLIN and DATA modules of SYSTAT; Wilkinson, 1990) proved too inflexible to let this be done easily. Thus, I separated the data for each viewer into first vs second halves by the order in which they accumulated. These were chosen, in part, to maximize differences in the half data sets due to any possible practice effects, although previous statistical analyses had shown no simple effects.

I next found asymmetries in the fits for the first half carried onto the second half and vice versa. That is, the sum of least squares (SOLS) of the fits of the parameter

values from the first halves of the data in fitting the second halves did not predict the fits of the values from the second halves of the data in fitting the first. Thus, I performed the analyses both ways, and I report the means of the two initial fits and the means of the two fits using parameter values from the other halves (see Browne, 2000). As was done by Cutting *et al.* (1992), I fit both models (LIM and FLMP) to the data from all 44 viewers. I first obtained fits to each half of the data, second used the values of those parameters to fit the other half of the data and then averaged the two original SOLS values to obtain a first summary fit and the two parameter-free SOLS values to obtain a second summary fit.

Results

The first set of results was consistent with results reported by Cutting *et al.* (1992): The mean of fits of the two halves of the data revealed that the data of 23 viewers were better fit by LIM and 21 by FLMP. This is the same proportion found by Cutting *et al.* (1992). The relative fits of the models to those data from 40 of the viewers retained the ordering reported by Cutting *et al.* (1992) for their entire data sets (LIM < FLMP or FLMP < LIM in both the original fits and in the new analyses conducted here); the fits to two of the subjects' data previously favoring LIM now favored FLMP, and the fits of two others previously favoring FLMP now favored LIM. The relative fits of the two models to the data sets are shown as a scatter plot in the left panel of Fig. 4. Those points above the diagonal represent those subjects' data better fit by LIM; those below it by FLMP. Notice

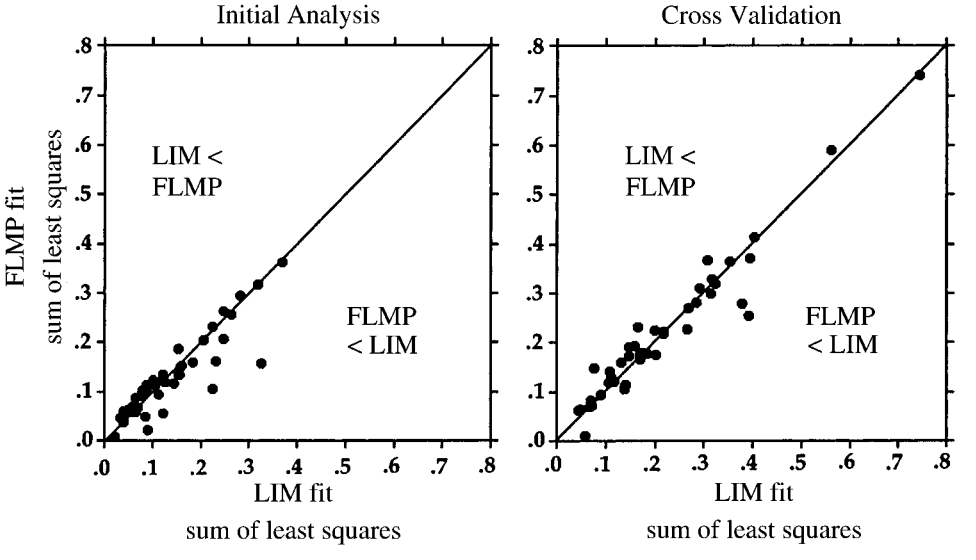

FIG. 4. The relative fits of two models, LIM and FLMP, to the depth-judgment data of 44 subjects reported by Bruno and Cutting (1988) and Cutting *et al.* (1992). The left panel shows the mean fits to the two halves of data those for 23 viewers were better fit by LIM and those of 21 by FLMP. The right panel shows the fits after cross-validation, using the parameter weights from one half of the data to fit the other. Although the data of 30 viewers were better fit by LIM and only 14 by FLMP, the general impression is that both models fit the data reasonably well and about equally well.

again that when FLMP provides a better fit, it often does so by a relatively wide margin, as reported by Cutting *et al.* (1992) and by Myung and Pitt (1997).

The second set of results involved the flexibility penalty assessed through cross validation. That is, if FLMP has the capability of absorbing variance in the data beyond merely fitting the data, then the fits with fixed weights to other halves of the data should be relatively worse than that for LIM. In general, this is what happened: After cross-validation, the data of 30 of the viewers were better fit by LIM and the data of only 14 by FLMP ($z = 2.265$, $p < .012$). The relative fits of the models to 35 of the viewers retained their ordering after cross-validation, the relative fits to the data of 8 viewers reversed from favoring FLMP to favoring LIM, and the fits for the other viewer reversed from that favoring LIM to FLMP. The comparative fits of the two models after cross-validation are shown in the right panel of Fig. 4. Notice that the cross-validation penalties seemed to make the relative fits more symmetric around the diagonal.

In keeping with this insight, consider an analysis of variance on the group of fits, before and after cross-validation. There was a reliable interaction of model fit before and after cross-validation ($F(1,43) = 11.9$, $p < .001$). That is, the mean SOLS values before cross-validation were 0.1358 and 0.1274 for LIM and FLMP, respectively; their values after cross-validation were 0.2136 and 0.2176, respectively. This shows that the cross-validation procedure reliably changed the relative fits of the two models, reinforcing the importance of this symposium to psychological modeling. Nonetheless, with differences only in the third significant digits of the loss function, the results of the cross validation analysis also suggest that the models fit the data about equally well.

CONCLUSION: ACCURACY, SCOPE, AND FLEXIBILITY OF MODELS

The purpose of this paper has been to offer new, empirical information about and perhaps a new perspective on modeling. Traditionally we have been interested in the accuracy, the scope, and the simplicity of models. Although accuracy is well measured, scope is less so, and simplicity may not even be a useful concept in modeling; just as it has proven problematic in perception. Thus, instead of considering simplicity, I suggest we focus on flexibility.

This new focus is warranted, I claim, by the results presented here. Two non-nested models, FLMP and LIM, were fit to 44 psychological data sets; 23 of these comparative fits favored LIM and 21 favored FLMP. Thus, there was no decisive winner; both had roughly equal accuracy. These models also have the same number of parameters, and thus by traditional standards they are equally simple. However, FLMP proved to have considerably larger scope than LIM in fitting psychometric and exponential functions. Thus, with essentially equal accuracy, equal simplicity, and unequal scope, one might prefer FLMP. Yet one factor seemed to mitigate this conclusion. FLMP fit random data better than LIM, suggesting that FLMP might have unwarranted flexibility. How might one compare models, properly penalizing a more flexible model? This symposium offered examples, and these methodological techniques for comparing models should become much more widely known, understood, and used by psychologists. There are now several ways in which a more

flexible model can be properly penalized. Here, I used cross–validation, but ICOMP (Bozdogan, 2000) or a Bayesian approach (Wasserman, 2000) could also have been used.

The outcome here was that, quantitatively, FLMP was sufficiently penalized that LIM now yielded a statistical but small superiority in its fits to the individual data sets. But this result raises a new tension concerning the measurement of accuracy. Through the methods of cross-validation, the accuracy of fits for LIM to these data sets is superior to that of FLMP. I am reluctant, however, to claim success in breaking the deadlock between LIM and FLMP. Scrutiny of the results shown in Fig. 4 after cross-validation reveals that they are essentially the same. From the residuals it seems clear that both models fit the data reasonably equally and reasonably well (see Massaro, 1998, for further discussion). Thus, when differential flexibility of two models has been neutralized (by cross-validation or any other technique), and the accuracy of one model shown to be reliably greater than the other (LIM better than FLMP), the outcome can still be indecisive.

We are left, then, with a choice between the two situations suggested in Fig. 1. If one interprets the outcome of this exercise in cross-validation as is apparent in the right panel of Fig.–4—that the sets of fits are essentially equal and equally good—then the situation shown in the left panel of Fig. 1 holds. That is, one should conclude that FLMP is the better model because, given equal accuracy in this domain and neutralized differences in flexibility, FLMP has greater scope than LIM. However, if one accepts the statistical outcome of the cross-validation—that the data of 30 of the 44 subjects are better fit by LIM—then the situation shown in the right panel of Fig. 1 holds, with the area of concern the small section of the Venn diagram where Model A is superior to Model B. That is, one should conclude that, after differential flexibility is neutralized, LIM is preferred to FLMP in this particular case, despite the greater scope of FLMP.

REFERENCES

Ashby, G., & Townsend, J. (1986). Varieties of perceptual independence. *Psychological Review*, **93**, 154–179.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, **61**, 183–193.

Attneave, F. (1982). Pragnanz and soap bubble systems: A theoretical explanation. In J. Beck (Ed.), *Organization and representation in perception* (pp. 11–29). Hillsdale, N. J.: Erlbaum.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in informational complexity. *Journal of Mathematical Psychology*, **44**, 62–91.

Browne, M. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, **44**, 108–132.

Bruno, N., & Cutting, J. E. (1988). Minimodularity and the perception of layout. *Journal of Experimental Psychology: General*, **117**, 161–170.

Burton, H. (1945). The optics of Euclid. *Journal of the Optical Society of America*, **35**, 357–372.

Carr, H. (1935). *An introduction to space perception*. New York: Longmans.

Chaitin, G. J. (1977). Algorithmic information theory. *IBM Journal of Research and Development*, **21**, 350–359.

Cutting, J. E. (1981). Coding theory adapted to gait perception. *Journal of Experimental Psychology: Human Perception and Performance*, **7**, 71–87.

Cutting, J. E. (1993). Perceptual artifacts and phenomena: Gibson's role in the 20th century. In S. Masin (Ed.), *Foundations of perceptual theory* (pp. 231–260). Amsterdam: North-Holland.

Cutting, J. E. (1997). How the eye measures reality and virtual reality. *Behavior Research Methods, Instruments, & Computers*, **29**, 27–36.

Cutting, J. E., & Bruno, N. (1988). Additivity, subadditivity, and the use of visual information: A reply to Massaro (1988). *Journal of Experimental Psychology: General*, **117**, 422–424.

Cutting, J. E., Bruno, N., Brady, N., & Moore, C. (1992). Selectivity, scope, and simplicity of models: A lesson from fitting judgments of perceived depth. *Journal of Experimental Psychology: General*, **121**, 364–381.

Cutting, J. E., & Vishton, P. M. (1995). Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In W. Epstein & S. Rogers (Eds.) *Perception and space and motion* (pp. 69–117). San Diego: Academic Press.

De Leeuw, J., & Stoop, I. (1984). Upper bounds for Kruskal's stress. *Psychometrika*, **49**, 391–402.

Dunn-Rankin, P. (1983). *Scaling methods.* Hillsdale, N. J.: Erlbaum.

Garner, W. R. (1966). To perceive is to know. *American Psychologist*, **21**, 11–19.

Gibson, J. J. (1950). *Perception of the visual world.* Boston: Houghton Mifflin.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston: Houghton Mifflin.

Goodman, N. (1972). *Problems and projects.* Indianapolis, IN: Bobbs-Merrill.

Grunwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, **44**, 133–152.

Helmholtz, H. von (1925). *Physiological optics* (3rd ed., Vol. 3; J. P. C. Southall, trans.) Menasha, WI: The Optical Society of America (original work published in German in 1867).

Hochberg J., & McAlister, E. (1953). A quantitative approach to figural "goodness." *Journal of Experimental Psychology*, **46**, 361–364.

Hochberg, J. (1982). How big is a stimulus? In J. Beck (Ed.) *Organization and representation in perception* (pp 191–218). Hillsdale, N. J.: Erlbaum.

Hochberg, J. (1986). Visual perception. In R. Atkinson, R. Herrnstein, G. Lindsay, & R. D. Luce (Eds.) *Stevens's handbook of experimental psychology* (pp. 195–276). New York: Wiley.

Jacobs, A. M., & Grainger, J. (1994). Models of visual word recognition: Sampling the state of the art. *Journal of Experimental Psychology: Human Perception and Performance*, **20**, 1311–1334.

Kahneman, D., Slovic, P., & Tversky, A., Eds, (1982). *Judgment under uncertainty: Heuristics and biases.* New York: Cambridge University Press.

Klahr, D. A. (1969). Monte Carlo investigations of the statistical significance of Kruskal's scaling procedure. *Psychometrika*, **34**, 319–330.

Koffka, K. (1935). *Principles of Gestalt psychology.* New York: Harcourt.

Komolgorov, A. N. (1968). Logical basis for information theory and probability theory. *IEEE Transactions of Information Theory*, **14**, 662–664.

Kuhn, T. S. (1977). Objectivity, value judgment, and theory choice. In T. S. Kuhn (Ed.), *The essential tension* (pp. 320–339). Chicago: University of Chicago Press..

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measurement and modeling of depth cue combination. *Vision Research*, **35**, 389–412.

Leeuwenberg, E. (1971). A perceptual coding language for visual and auditory patterns. *American Journal of Psychology*, **84**, 307–349.

Leeuwenberg, E. (1982). Metrical aspects of patterns and structural information theory. In J. Beck (Ed.), *Organization and representation in perception* (pp. 57–71). Hillsdale, NJ: Erlbaum.

Leeuwenberg, E., & Boselie, F. (1988). Against the likelihood principle in visual form perception. *Psychological Review*, **95**, 485–491.

Levine, D. M. (1978). A Monte Carlo study of Kruskal's variance-based measure of stress. *Psychometrika*, **42**, 307–315.

Li, S.-C., Lewandowski, S., & DeBrunner, V. E. (1996). Using parameter sensitivity and interdependence to predict model scope and falsifiability. *Journal of Experimental Psychology*: *General*, **125**, 360–369.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, **74**, 431–461.

Mach, E. (1906). *The analysis of sensations and the relation of the physical to the psychical* (5th ed., S. Waterlow, trans.). New York: Dover.

Massaro, D. W. (1988). Ambiguity and perception in experimentation. *Journal of Experimental Psychology*: *General*, **117**, 417–421.

Massaro, D. W. (1998). *Perceiving talking faces*: *From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

Massaro, D. W., & Cohen, M. M. (1992). The paradigm and the fuzzy-logical model of perception are alive and well. *Journal of Experimental Psychology*: *General*, **122**, 115–124.

Massaro, D. W., & Friedman, D. (1990). Models of integration given multiple sources of information. *Psychological Review*, **97**, 225–252.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, **4**, 79–95.

Nisbett, R., & Ross, L. (1980). *Human inference*: *Strategies and shortcomings of social judgment*. Englewood Cliffs, N.J.: Prentice–Hall.

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology*: *Human Perception and Performance*, **17**, 3–28.

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review*, **72**, 407–418.

Parducci, A. (1974). Contextual effects: A range-frequency analysis. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. 2, pp. 127–141). San Diego: Academic Press.

Restle, F. (1979). Coding theory of the perception of motion configurations. *Psychological Review*, **86**, 1–24.

Richter, J. P. (1883). *The notebooks of Leonardo da Vinci*. Reprinted, New York: Dover, 1970.

Rosch, E., & Lloyd, B. B. (1978). *Cognition and categorization*. Hillsdale, NJ: Erlbaum.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, **210**, 390–398.

Shrout, P. E. (1997). Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychological Science*, **8**, 1–2.

Smith, E. E. (1990). Concepts and induction. In M. I. Posner (Ed.), *Foundations of cognitive science* (pp. 501–526). Cambridge, MA: MIT Press.

Thagard, P. (1990). *Conceptual revolutions*. Princeton, NJ: Princeton University Press.

Wagenaar, W. A., & Padmos, P. (1971). Quantitative interpretation of stress in Kruskal's MDS technique. *British Journal of Mathematical and Statistical Psychology*, **24**, 101–110.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, **44**, 92–107.

Wertheimer, M. (1967). Laws of organization in perceptual forms. In W. D. Ellis (trans. and Ed.), *A sourcebook of Gestalt psychology* (pp. 17–54). New York: Humanities Press, 1967. (Originally published in German in 1920).

Wilkinson, L. (1990). *SYSTAT*: *The system for statistics*. Version 5.2 [Computer program]. Evanston, IL: SYSTAT, Inc.

Witkin, A. P., & Tenenbaum, J. M. (1983). On the role of structure in vision. In J. Beck, B. Hope, & A. Rosenfeld (Eds.), *Human and machine vision* (pp. 481–543). New York: Academic Press.

Wittgenstein, L. (1953). *Philosophical investigations* (G. E. Anscombe, trans.). London: Basil Blackwell & Mott.

Received: September 17, 1999; revised: March 31, 1998