Journal of Experimental Psychology: General
1992, Vol. 121, No. 3, 364-381

0096-3445/92/$3.00

Selectivity, Scope, and Simplicity of Models: A Lesson From Fitting Judgments of Perceived Depth

James E. Cutting
Cornell University

Nicola Bruno
University of Trieste
Trieste, Italy

Nuala P. Brady
Cornell University

Cassandra Moore
Columbia University

When comparing psychological models a researcher should assess their relative selectivity, scope, and simplicity. The third of these considerations can be measured by the models' parameter counts or equation length, the second by their ability to fit random data, and the first by their differential ability to fit patterned data over random data. These conclusions are based on exploration of integration models reflecting depth judgments. Replication of Massaro's (1988a) results revealed an additive model (Bruno & Cutting, 1988), and Massaro's fuzzy-logical model of perception (FLMP) fit data equally well, but further exploration showed that the FLMP fit random data better. The FLMP's successes may reflect not its sensitivity in capturing psychological process but its scope in fitting any data and its complexity as measured by equation length.

Good scientific theories are usually thought to have several properties: They are accurate, simple, broad in scope, internally consistent, and have the ability to generate new research (Kuhn, 1977). When models can be used to instantiate theories, they might reflect these same properties. For our purposes the two key concepts in this set are *simplicity* and *scope.* Simplicity can be measured in several ways. We measure it in two: by the number of parameters in a model and, in a way not customary to experimental psychology, by the length of the equation that instantiates a model. Scope can also be measured in various ways, but here we consider how theory or model accounts for all possible data functions, where those functions are generated by a reasonably large sample of random data sets. Under this construal, broad scope is a mixed blessing. A model with greater scope than another may fit more data functions of interest to the researcher, but simultaneously it may also fit more functions of no interest. Thus, we propose a new criterion for testing and comparing models: *selectivity.* We define selectivity as the relative ability of a model to fit data functions of interest with its ability to fit random data factored out.

This perspective on modeling arises out of our struggles with three different sources of evidence: first, our continuing empirical study of how individuals use multiple sources of information about the perception of objects laid out in depth (see also Bruno & Cutting, 1988); second, our study of the properties of the models used to fit those data; and third, our investigation of why those models have the data-fitting properties they do. Thus, our presentation is divided into these three parts, followed by a set of suggestions about how future research with psychological models might be conducted.

Models of Information Integration and Their Fits to Human Judgments of Depth

How do we perceive the layout of objects in depth? This is among the oldest questions in psychology, and answers to it have been myriad. One reason for persistent interest in, and debate over, this query is the existence of multiple sources of information in any scene, which can contribute to perceived depth. One list, reworded and reorganized from Gibson (1950, pp. 71-73), includes binocular disparity, convergence, accommodation, linear perspective, apparent size, relative motion, occlusion, aerial perspective, height in plane, shading, and texture gradients. To be sure, some theorists deny the existence of multiple sources of information (e.g., Burton & Turvey, 1990), based largely on Gibson's later thoughts about invariance and one-to-one mappings between information and objects or events (see Cutting, 1986, 1991a, 1991b). However, for those who accept their existence for the perception of objects in depth (and for the perception of many other properties) a major question arises: How is all this information used?

Two general possibilities about information use emerge: Either perceptual information is selected, one source from

This research was supported by National Science Foundation Grant BNS-8818971 to James E. Cutting. Results without modeling or simulations were reported briefly at the 28th annual meeting of the Psychonomic Society, Seattle, Washington, November 1987.

We thank Dominic W. Massaro for helping us understand and implement the fuzzy-logical model of perception; Michael S. Landy and Mark J. Young for insights into implementing other models; James L. McClelland for a general discussion about modeling; William Epstein, James A. Ferwerda, and Mary M. Hayhoe for random discussions related to the topics presented here; Carol L. Krumhansl, Michael S. Landy, Geoffrey R. Loftus, Dominic W. Massaro, and an anonymous reviewer for comments on previous versions of this article; and Nan E. Karwan for sustained interest in and discussions about the project.

Correspondence concerning this article should be addressed to James E. Cutting, Department of Psychology, Uris Hall, Cornell University, Ithaca, New York 14853-7601.

many, or many sources are integrated in some manner. Although there are examples in support of information selection (e.g., Cutting, 1986; Cutting & Millard, 1983; Knudsen & Konishi, 1979), information integration is probably more common. If integration of various sources of depth information occurs, a new question arises: By what rule are these sources combined? Many more possibilities emerge here, but we confine ourselves to three approaches—additive, multiplicative, and averaging—and four corresponding models from the literature.

Modularity, Paradigms, Models, and Depth

Results concerning how information is combined have been taken as evidence both for and against modularity (Fodor, 1983; Marr, 1982)—an idea about isolable and independent subfaculties of the mind that do not pass information freely among themselves. Cutting and Bruno (1988), for example, reported additive combination of visual information and took their data as evidence for separate "minimodules" within the visual system. Maloney and Landy (1989; Landy, Maloney, & Young, 1991) also espoused this point of view. In contrast, Massaro (1989) reported multiplicative combination of various kinds of information and took his data as evidence against modularity—interactions of data and similarities of processing strategies abound.

As a result of the set of investigations reported here, we now doubt that models of information combination can speak to the issue of modularity. We do not doubt, however, the importance of information integration to all areas in, and all modalities of, perception. Thus, we consider integration models in detail.

A Paradigm and the Fuzzy-Logical Model of Perception (FLMP)

Massaro (1987b, 1989) promoted a new paradigm for psychological research.[1] For our purposes, it has three parts. First, the paradigm embraces the existence of multiple sources of information and the problem of their integration in perception. We are pleased with this stance, in part because it dovetails nicely with *directed perception* (Cutting, 1986, 1991a, 1991b), which embraces multiple specification of perceived objects and events but makes no statement about the combination rule.[2] In both Massaro's view and in ours the perceptual world is a rich place, full of information to be picked up, gathered, and processed at every turn. Second, incorporating Platt's (1964) idea of *strong inference*, the paradigm proceeds by binary opposition, pitting two hypotheses—instantiated as models—against one another. We are not fans of strong inference (see Cutting & Millard, 1983, p. 207), but we recognize its attractiveness in dealing with competing hypotheses, unfettered by considerations of the null hypothesis. Third, the paradigm is idiographic; wherever possible it focuses on individual data. This focus is appropriate, but we offer a caveat against it when these data are the inputs to models with different scope.

Built on the work of Anderson (1981, 1982), Massaro's paradigm systematically explores information integration. Massaro found impressive support for a type of multiplicative combination, as captured by his FLMP. Many models can be cast in fuzzy-logical format, and thus it is not the format we are interested in by its multiplicative nature. The domains Massaro has studied are impressive in breadth and cover most of cognitive psychology; they include attention (Massaro, 1985), reading (Massaro, 1984, 1987a), letter recognition (Massaro & Friedman, 1990; Massaro & Hary, 1986), and speech perception (Massaro, 1987b, 1989), and they all support FLMP. Before Bruno and Cutting (1988), however, Massaro had not explored the perception of objects laid out in depth, and visually perceived depth is the domain of this article.

Additivity, Depth, and Previous Results

There are two empirical byproducts of adding sources of depth information to a display: The range and mean of observers' depth judgments increases, and the variability in their judgments decreases (Künnapas, 1968). Bruno and Cutting (1988) and Massaro (1988a) focused on increases in apparent depth; others, such as Maloney and Landy (1989), have focused on both. Here, we continue our focus on the former.

Bruno and Cutting (1988) reported three experiments on the perception of three panels laid out in depth, with variation in four sources of information: relative size (s), height in plane (h), occlusion (o), and motion parallax (p). Schematic versions of the stimuli are shown in Figure 1. The observer's simulated path for motion parallax stimuli is shown in Figure 2. Together, we call these sources of information *shop*, following a scheme suggested by Massaro (1988a), and we indicate the presence or absence of information about each source by a 1 or 0, respectively, as codes for each variable. Thus, Stimulus 0000 has no depth information. Stimuli 1000, 0100, 0010, and 0001 have only one source of information about size, height, occlusion, and motion parallax, respectively;

---

[1] Massaro (1987b, 1989) cast his paradigm in the shadow of Popper (1959) and falsificationism. However, there is an inherent conflict between the notion of falsification (an idea from Popper about the logic of the scientific method) and the notion of paradigm (an idea from Kuhn, 1970, against the logical standards of any scientific method). There are also abundant criticisms of falsificationism (see Schilpp, 1974; Suppe, 1977). In addition, Kuhn (1974) later recast his notion of paradigm as a *disciplinary matrix* consisting of several parts. He considered only three in detail: symbolic generalizations (equations with agreed-on variables), models (or, to avoid confusion with the mathematical models as used here, metaphors), and exemplars (or examples of how one should do science). We think this last meaning is closest to Massaro's intention in his use of the term *paradigm*; it is also this intention we question, at least in the paradigm's current instantiation.

[2] Massaro's approach also contrasts with directed perception in one other important way. That is, his approach partly follows Brunswik (1956) in that it embraces the idea of cue validity, the notion that sources of information are probabalistically related to objects and events in the world. Directed perception assumes multiple sources of information each specify (map uniquely back to) objects and events (Cutting, 1986, 1991a, 1991b).

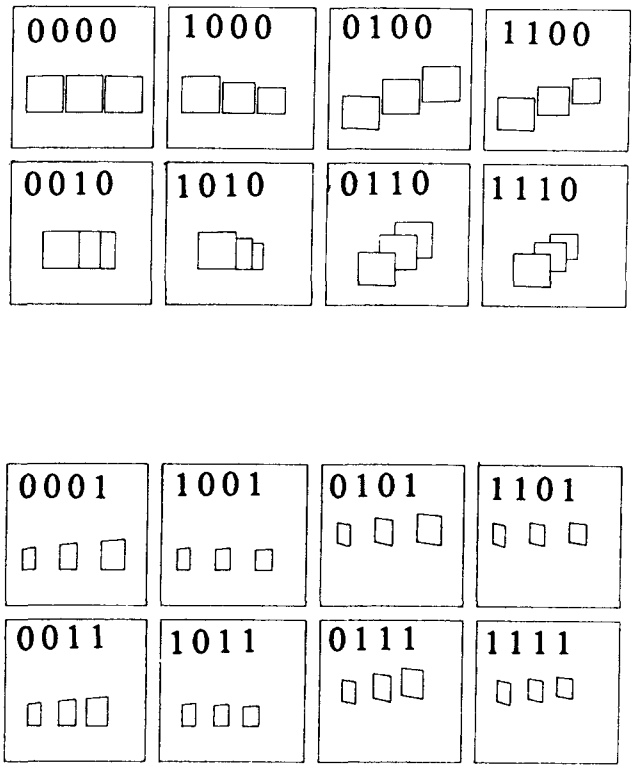

Figure 1. Representations of the 16 stimuli used in the two experiments reported here and in Bruno and Cutting (1988). (The upper panels show the eight static stimuli and the last frames of the eight moving stimuli; the lower panels show the first frames of the eight moving stimuli. The four places in each stimulus code correspond to the four sources of information: size, height in plane, occlusion, and motion parallax; 1 indicates the presence and 0 indicates the absence of information.)

Stimulus 1001 has size and parallax information; and so forth, until one reaches Stimulus 1111, which has all four sources. In this manner, across the 16-item set, a stimulus could have 0, 1, 2, 3, or 4 sources of information. Notice that all sources of information were either present (providing differential depth information) or absent (providing no differential depth information); nothing beyond binary oppositions was used.

Bruno and Cutting's (1988) first experiment was a direct-scaling task. Viewers indicated the degree of relative depth perceived among the panels on a scale of 0 to 99, with larger numbers indicating more depth. The data for the first experiment are shown in the left panel of Figure 3. Their second experiment was an indirect-scaling task using dissimilarity judgments ("how different are these stimuli in the relative depth they portray?"). Viewers rated differences among pairs of stimuli on a scale from 1 to 9 (with 9 indicating maximal difference). Results were forced (but with relatively little stress) into one dimension using multidimensional scaling. Their third experiment was also an indirect-scaling task, this time using preference judgments among pairs ("which stimulus reveals most depth?"). Results were scaled according to Thurston's Case V (e.g., see Dunn-Rankin, 1983).

In all three studies Bruno and Cutting (1988) claimed support for additive information integration. That is, ignoring differences in weights among separate sources of information (which are reported in detail in Bruno & Cutting, 1988, and in Massaro, 1988a), scaled judgments of stimuli with different numbers of information sources were generally linear. In addition, an expected set of inequalities held for judgments among stimulus classes with increasing numbers of information sources, with mean scale values in the order $0 < 1 < 2 < 3 < 4$. Moreover, the differences were roughly equal between means of stimulus classes differing by one source: $(1 - 0) \sim (2 - 1) \sim (3 - 2) \sim (4 - 3)$.

Massaro (1988a) was unconvinced by these claims; an additive model is only one of many classes of models that might fit the kinds of results Bruno and Cutting (1988) reported. Massaro then proposed two models of perceived layout and reanalyzed the individual data from Bruno and Cutting's first experiment, fitting models to data. We now consider these models.

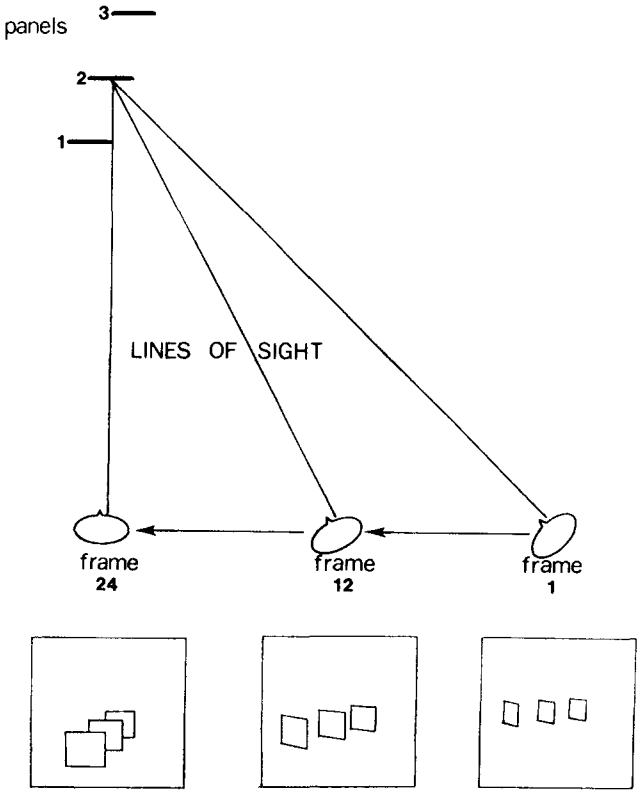

Figure 2. The setting simulating Stimulus 1111, suggesting the motion of the observer for those stimuli with motion parallax. (The lower panels show three frames taken out of the stimulus sequence. From "Minimodularity and the Perception of Layout," by N. Bruno and J. E. Cutting, 1988, *Journal of Experimental Psychology: General, 117,* p. 163. Copyright 1988 by the American Psychological Association. Adapted by permission.)

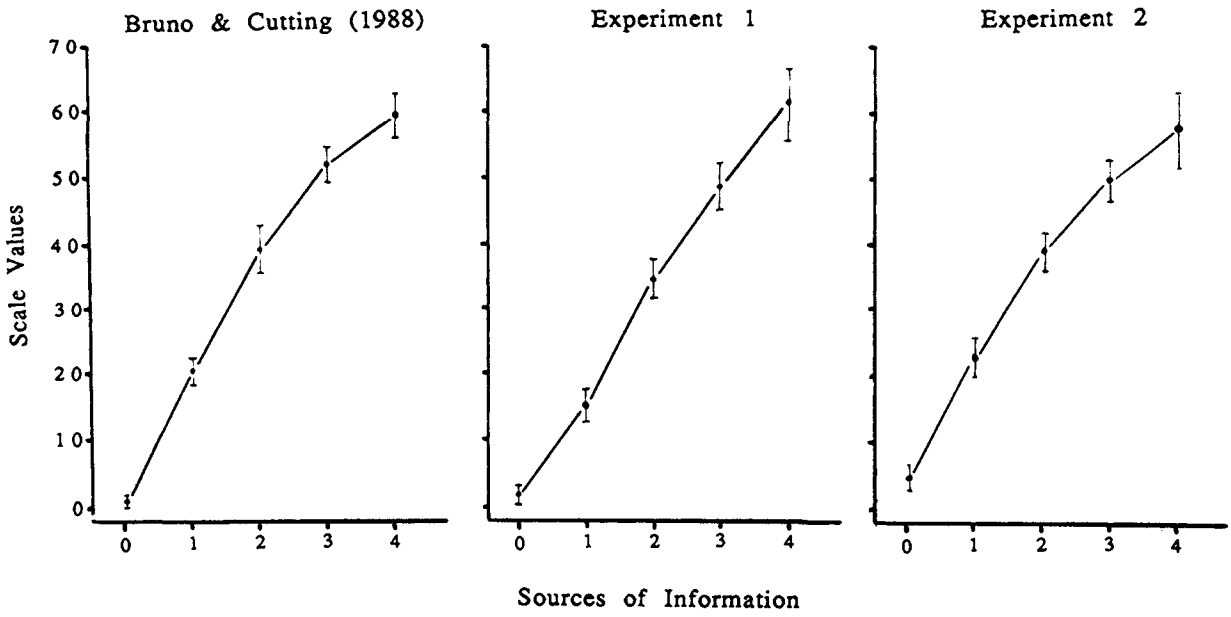

Figure 3. Rating scale data of Bruno and Cutting (1988, Experiment 1) and data of the two experiments reported here, as a function of the number of sources of information in a stimulus. (Error bars indicate ± 1 *SEM.*)

Additive and Multiplicative Models of Information Integration for Judgments of Depth

Before presenting these two integration models we need to establish a set of conventions. Lowercase italicized letters, *shop,* stand for the codes indicating the presence (1) or absence (0) of each source of information in the stimuli; uppercase italicized letters, *SHOP,* stand for the weights corresponding to the codes as they are to be determined by fitting the models to data; and uppercase bold letters, **SHOP,** stand for their respective parameters as implemented in the models.

The Additive Model

The first model is essentially an analysis of the data by multiple linear regression. Its full form is as follows:

$$R(\text{Depth}) = \mathbf{S} + \mathbf{H} + \mathbf{O} + \mathbf{P} + \mathbf{B},$$

$$\text{where } \left\{\begin{array}{l}\mathbf{S} = S \text{ when } s = 1 \text{ (present)} \\ \mathbf{S} = \text{zero when } s = \text{zero (absent)}\end{array}\right\} \text{ etc.,}$$

$$\text{for } \mathbf{H,O,P,B}, \quad (1)$$

and where R(Depth) is the rating for the amount of perceived depth in a given display. The new term, **B,** is the parameter for a background variable (b) not manipulated in the experiment but present on all trials and in all stimuli. The background variable (which includes the flatness of the screen) would usually suggest less depth than would otherwise be apparent in the display; it also serves, simply, as an intercept in the regression function. Given one-to-one mappings between codes and weights and between weights and parameters, the stipulations below the equation line are not needed. The full form is given here only so it can be compared with FLMP and, later, with other models.

FLMP

Massaro's model takes on different forms in different contexts for different purposes. Applied to our four-source situation, and with some clarifications, it is as follows:

$$R(\text{Depth}) = \mathbf{S} \times \mathbf{H} \times \mathbf{O} \times \mathbf{P} \times \mathbf{B} / [\mathbf{S} \times \mathbf{H} \times \mathbf{O} \times \mathbf{P} \times \mathbf{B} + (1 - \mathbf{S}) \times (1 - \mathbf{H}) \times (1 - \mathbf{O}) \times (1 - \mathbf{P}) \times (1 - \mathbf{B})],$$

$$\text{where } \left\{\begin{array}{l}\mathbf{S} = S \text{ when } s = 1, \\ \mathbf{S} = (1 - S) \text{ when } s = 0\end{array}\right\} \text{ etc., for } \mathbf{H,O,B}, \text{ and}$$

$$\text{where } \left\{\begin{array}{l}\mathbf{P} = P \text{ when } p = 1, \\ \mathbf{P} = 0.5 \text{ when } p = 0,\end{array}\right\} \text{ and}$$

$$\text{where } 0.0 < S,H,O,P,B < 1.0. \quad (2)$$

The difference between weights and parameters in FLMP is crucial: When relative size is present in a stimulus (e.g., as in Stimulus 1000), the parameter for Size **S** in Equation 2 is given the weight S. However, when it is absent (e.g., as in Stimulus 0000), all instances of **S** are given the weight $(1 - S)$. Notice then that the parameter string corresponding to size in the denominator, $(1 - \mathbf{S})$, is then given a weight of $[(1 - (1 - S)]$, or S, in the absence of relative size.[3]

[3] In addition, one can also recast the additive model into fuzzy-logical format as well. It becomes the following:

$$R(\text{Depth}) = \mathbf{S} + \mathbf{H} + \mathbf{O} + \mathbf{P} + \mathbf{X} / [\mathbf{S} + \mathbf{H} + \mathbf{O} + \mathbf{P} + \mathbf{X} + (1 - \mathbf{S}) + (1 - \mathbf{H}) + (1 - \mathbf{O}) + (1 - \mathbf{P}) + (1 - \mathbf{X})],$$

Three other aspects of FLMP are also important. First, the value of 0.50 is reserved for complete ambiguity, with values less than 0.50 indicating degrees of flatness and values above indicating degrees of depth. Moreover, the effect of presence and absence of a source of information is measured by the same-sized weights as they deviate positively or negatively from 0.50. Thus, when one source is present it might have a weight of 0.61, and when absent it might have a weight of 0.39.

Second, Massaro treated motion parallax (p) differently than the other sources. He argued that when p is absent it leaves depth ambiguous, and hence the variable should be removed from Equation 2. Giving it a fixed value of 0.50 effectively does this (see Massaro, 1987b, p. 167).

Finally, the difference between the additive model and FLMP (a multiplicative model) is less than might first be apparent. If an individual's judgments were fit to a normal distribution and then z transformed, FLMP simply adds the resulting z scores (see Massaro & Friedman, 1990). Thus, both models can be thought to be additive; the additive model simply adds evaluations of information among the five sources, whereas FLMP evaluates the normalized information from those judgments and then adds it. Thus, differences between the additive model and FLMP will manifest themselves most clearly in the use of the extremes of the scale range, where linear and logistics functions diverge.

Fits of the Two Models to Human Data

In his commentary on Bruno and Cutting (1988), Massaro (1988a) found mixed support for additive combination of information. Computing best fits to the data of 10 subjects who participated in Bruno and Cutting's Experiment 1, he found that the data from 5 were better fit by the additive model and the data from 5 others were better fit by FLMP. In response, Cutting and Bruno (1988) reanalyzed their original data, looking for a subadditive trend. Subadditivity is one of the possible fruits of multiplicative integration, where the addition of successive sources results in successively smaller increments in judgment values. Cutting and Bruno found reliable subadditivity in the results of their Experiment 1 but not in Experiments 2 and 3.

Two Methodological Qualms

Although the outcome of debate between Bruno and Cutting (1988) and Massaro (1988a) was inconclusive, two empirical leads held hope for differentiating the two positions. Both are based on a critique of the methods used by Bruno and Cutting (1988); one is a criticism of multiplicative integration based on assumptions underlying direct-scaling procedures, and the other is a criticism of additive integration based on the composition of stimulus sets.

The first qualm concerns multiplicativity. Despite finding results partly consistent with a multiplicative model, Cutting and Bruno (1988) questioned FLMP's efficacy in the direct-scaling task. The 16 *shop* stimuli seemed to have a strong anchor, as suggested by inspecting the panels in Figure 1 and the results in Figure 3. That is, the Stimulus 0000 garnered judgments of near zero with essentially no variance. On the other hand, stimuli with three (1110, 1101, 1011, and 0111) and four (1111) sources of information were all rated toward the upper middle of the scale. This relative isolation of one stimulus at one end of the continuum and clustering of stimuli at the other looked like a possible range-frequency effect (Parducci, 1965, 1974). In particular, it seemed as if viewers may not have been using the scale in a linear fashion. The use of the scale, then, might reflect not simply perceived differences in depth but also perceived differences in the spacing of stimuli along a continuum of depth and their consequent compensatory adjustments.

The second qualm concerns additivity. Bruno and Cutting (1988) combined the four sources of information orthogonally in their stimulus set. Although the results seemed generally to support additivity, this result may have been directly caused by the independent manipulation of the four sources. That is, uncorrelated information in a stimulus set may create uncorrelated information use; perhaps correlated information would create interactions implicating some other model.

Two experiments were conducted to resolve these two issues and then provide more grist for modeling. In one experiment we manipulated the shape of the frequency distribution of information in the stimulus set, and in the other we correlated selected sources across the stimulus set.

General Method

Stimuli were identical to those used by Bruno and Cutting (1988). Four sources of information were varied orthogonally across the set: relative size, height in plane, occlusion, and motion parallax. The eight static stimuli are shown in the top panels of Figure 1; the beginning frames of the motion stimuli are shown in bottom panels. The last frame of each motion stimulus was identical to its static counterpart. Again, Figure 2 shows the experimental situation generating motion parallax. Stimuli were generated on a Hewlett-Packard (HP) 1000L Series computer and displayed on an HP 1350s vector-plotting system with a P31 phosphor and 1,024 × 1,024 pixel resolution. They subtended about 8° and were seen binocularly in a moderately lit room, with the sides of the display clearly visible. Each stimulus was presented for about 2 s, and each frame of the motion sequences was 87 ms. (See Bruno and Cutting, 1988, for further details.) The two experiments reported here differed in how often the stimuli were selected from the population of 16. The distributions are shown in Table 1. In Experiment 1 the stimuli were selected by skewing the population of stimuli in the test sequence according to the number of sources represented, as shown in Table 2.

Thirty-four members of the Cornell University community were run individually, 16 in Experiment 1 and 18 in Experiment 2.

with the same provisos as Equation 2 for mappings between codes, weights, and parameters. Of course, this equation vastly simplifies to something very close to Equation 1. However, run this way the computational results of this fuzzy-logical equation and Equation 1 are different. That is, although for any given input data the sum of least squared deviations is the same as given by Equation 1, the weights (as deviations from 0.50) are closer to those given by the fuzzy-logical model of perception (FLMP) and Equation 2. Thus, we claim it is not the fuzzy-logical form of FLMP that is important, it is the Bayesian form and the multiplicative manner in which sources of information are combined.

Table 1
Stimulus Frequencies for Experiments 1 and 2

Stimulus (*shop*) code[a]	No. of sources of information	Experiment 1: Skew		Experiment 2: Correlation					
				sh		*so*		*sp*	
		Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
0000	0	10	50	10	2	10	2	10	2
1000	1	5	10	2	10	2	10	2	10
0100	1	5	10	2	10	10	2	10	2
1100	2	5	5	10	2	2	10	2	10
0010	1	5	10	10	2	2	10	10	2
1010	2	5	5	2	10	10	2	2	10
0110	2	5	5	2	10	2	10	10	2
1110	3	10	5	10	2	10	2	2	10
0001	1	5	10	10	2	10	2	2	10
1001	2	5	5	2	10	2	10	10	2
0101	2	5	5	2	10	10	2	2	10
1101	3	10	5	10	2	2	10	10	2
0011	2	5	5	10	2	2	10	2	10
1011	3	10	5	2	10	10	2	10	2
0111	3	10	5	2	10	2	10	2	10
1111	4	50	10	10	2	10	2	10	2

Note. pos = positive; neg = negative.
[a] Stimulus (*shop*) code refers to the presence (1) or absence (0) of information on relative size (*s*), height in plane (*h*), occlusion (*o*), and motion parallax (*p*).

Participants in Experiment 1 were paid $4 for about 25 min of viewing; those in Experiment 2 were paid $5 for about 35 min of viewing. Individuals rated the relative depth of the panels in the stimuli on a scale from 0 to 99. Each stimulus was presented once before the individual was to make his or her response, but the trial could be repeated if the viewer desired. As practice, Stimuli 0000 and 1111 were alternated several times to familiarize viewers with the range of stimuli.

Experiment 1. Bruno and Cutting's (1988) Experiment 1 had one stimulus with no sources of information about depth (0000), four stimuli with one source (1000, 0100, 0010, 0001), six stimuli with two sources (1100, 1010, 1001, 0110, 0101, 0011), four with three sources (1110, 1101, 1011, 0111), and one with four sources (1111). Because each stimulus was presented 10 times, the distribution of the number of sources in the test sequence—as measured by the second moment (variance, or spread), third moment (skewness, or asymmetry), and fourth moment (kurtosis, or peakedness)—was approximately normal (variance = 1.06 sources of information, skew = 0.0, and kurtosis = 2.2). In contrast, the distribution of stimuli in the two conditions in this study were equal to each other in variance (1.61) and kurtosis (2.1), but one had positive skew (0.56) and one had negative skew (−0.56). Two groups of 8 participants each viewed

Table 2
Distribution of Information Sources and Scale Values in Bruno and Cutting (1988) and in Experiment 1

Experiment	No. of sources of information[a]				
	0	1	2	3	4
Frequency distribution of stimuli (%)					
Bruno & Cutting (1988), Experiment 1	6	25	38	25	6
Experiment 1, negative skew[b]	7	13	20	27	33
Experiment 1, positive skew[b]	33	27	20	13	7
Mean normalized scale values[c]					
Bruno & Cutting (1988), Experiment 1	0	32	58	76	84
Experiment 1, negative skew	0	19	46	68	96
Experiment 1, positive skew	0	22	54	73	87

[a] How many of the sources of information—size, height in plane, occlusion, and motion parallax—are found in given stimuli. [b] Negatively skewed distributions have their long tail at the low end, and positively skewed distributions have their long tail at the high end. [c] The maximum mean scale value used by any subject for a particular stimulus was set at 100, and other values increased proportionately. Means were then taken across subjects.

sequences of 150 stimuli, randomized differently for each viewer. Thus, we tested possible range-frequency effects in a between-subjects design.

Experiment 2. The second study also molded stimulus populations, this time in six ways. Each selection involved *s* correlated with other sources of information. Two sequences correlated the occurrences of *s* and *h*. In one sequence stimuli were selected such that they were positively correlated ($r = .71$) across the test sequence, as shown in Table 3; in the second they were negatively correlated ($r = -.71$), but in both there were no correlations between all other sources (*so, sp, ho, hp,* and *op*), as can be computed from Table 1. In two other sequences *s* and *o* were correlated ($rs = .71$ and $-.71$), with other pairs uncorrelated, and in two others *s* and *p* were correlated, with other pairs uncorrelated. Each sequence consisted of 96 stimuli, randomly ordered for each participant. Three groups of 6 participants each viewed each pair of correlated sequences. Within groups, 3 viewed the positively correlated set first, then the negative set; the other 3 viewed the sequences in reverse order. Thus, we tested positive versus negative correlation effects within subjects, order effects between subjects, and source correlation effects between groups.

Experiment 1: Skewed Distributions of Information Do Not Affect the Use of the Response Scale

Massaro's (1987b, 1989) view of perception and pattern recognition entails three operations: evaluation of information, integration of information, and response classification. We concur with his claim that any complete approach to perception and the issue of multiple sources of information must have the formal equivalent of these three stages. In his application of FLMP to our stimulus situation (Bruno & Cutting, 1988), however, we worried that the classification process—rating (between 0 and 99) the perceived depth in the display—might have an additional classification component impeding the straightforward measurement of information integration. In particular, if the stimulus continuum used by Bruno and Cutting (1988) were perceived to be nonuniform, nonlinearities might result, and, thus, the data might be subject to range-frequency analysis (Parducci, 1965, 1974).

Range-frequency analysis starts with a consideration of two variables, a stimulus continuum and a response scale, and the relation between them. It proposes that the mappings between them are flexible and that participants adjust their use of a response scale (corresponding to Massaro's classification stage) according to the perceived distribution of stimuli along a continuum. One general prediction is that when the distribution of stimuli is skewed, the distribution of responses will be less skewed, and that individuals spread their responses more uniformly throughout the scale. These could create nonlinearities in the data.

Table 3
Correlation Matrix for Four Sources of Information in One of the Six Conditions in Experiment 2

Source of information	Height in plane (*h*)	Occlusion (*o*)	Motion parallax (*p*)
Relative size (*s*)	.71	.00	.00
Height in plane (*h*)	—	.00	.00
Occlusion (*o*)	—	—	.00

If range-frequency analysis applies here, the following patterns of shifts should occur in the response scale: A positively skewed stimulus continuum (the long tail at the upper end of the distribution, with more stimuli with 0 and 1 sources of information in them) should generally raise ratings for stimuli with 2 and 3 sources of information. A negatively skewed distribution (more stimuli with 3 and 4 sources of information) should lower ratings for stimuli with 1 and 2 sources. Thus, if range-frequency effects occur, the results should show a main effect of skewness (higher scores for positively skewed distributions than for negatively skewed distributions) and a possible interaction between skewness and number of sources (no difference between conditions at 0 and 4 sources, but with higher scores for the positively skewed distribution for stimuli with 1, 2, and 3 sources). Such a main effect and interaction would indicate that the classification stage can play a nonsignificant role in the results of Bruno and Cutting (1988) and perhaps create nonlinearities in the data; absence of such effects would suggest that the integration results are not contaminated by range-frequency-dependent classification effects.

The range of scale values was first normalized within individuals to correct for individual differences in scale use unrelated to the experimental manipulation. In particular, the lowest and highest mean values assigned to the 16 stimuli were set to 0 and 99, and intermediate values were linearly scaled between them. Analysis of variance was then performed on the data, looking at differences across groups (skewness of the stimulus distribution) and level (number of information sources). There was no reliable effect of skewness, $F(1, 14) = 0.128$, $p > .70$, and only a marginal interaction of Skewness × Number of Information Sources, $F(4, 56) = 2.35$, $p < .065$. Because of a cross-over in the data between stimuli with 3 and 4 sources (shown in Table 2), this second result is not easily interpretable as a range-frequency effect.

Thus, we conclude, contrary to the suggestion of Cutting and Bruno (1988), that the subadditivity seen in Experiment 1 of Bruno and Cutting (1988) is not due to a range-frequency, or classification, effect in the viewers' use of the scale. The lack of such an effect here may have been due to the orthogonality of information in the implied stimulus set (Garner, 1966; Pomerantz & Lockhead, 1991) rather than the actual stimulus sequence. That is, the mere presence of two *shop* stimuli, one with all sources of information (1111) and one with none (0000), may suggest to the viewer the 14 other possibilities; just as two dice, one with one pip and the other with six, imply the other four possibilities (two, three, four, and five), as well as their 20 other combinations. With such sets implied in the stimulus structure, little room may be left for any effect of a differential frequency manipulation.

Experiment 2: Correlated Information Sources Do Not Affect Responses

If uncorrelated information could create additive results, then correlated information might create differences across data sets revealed in any number of interactions. Candidates are between-groups interactions concerning which information sources are correlated, within-group interactions concern-

ing the order in which viewers participated, and within-subject interactions of polarity of correlation (positive or negative). More concretely, in a stimulus sequence within the positive *sh* correlation condition, Stimuli 11** and 00** (where * is a place holder for other sources of information, taking the value of 1 or 0) were frequent and Stimuli 01** and 10** were rare. The exemplar model of classification (Nosofsky, 1991) or any model based on the mere-exposure effect (Zajonc, 1968) would predict that due to their frequency, Stimuli 11** and 00** might garner higher than normal judgments, and due to their infrequency, Stimuli 10** and 01** might receive lower than normal judgments. These shifts should create a statistical interaction. When compared with the values for sequences with negative *sh* correlations, the interaction should be compounded, and so forth for other conditions.

Globally, there were no main effects of correlated sources of information; for *sh* vs. *so* vs. *sp*, $F(2, 12) = 0.005$, $p > .99$; means were 23.7, 24.7, and 25.1, respectively. There was no effect of polarity, with means of 24.0 and 25.0 for positive and negative correlations, $F(1, 12) = 1.306$, $p > .27$. There also was no effect of order of presentation; positive/negative = 24.3 and negative/positive = 24.7; $F(1, 12) = 0.116$, $p > .73$. Also, there were no simple interactions among polarity and order of presentation. Because there were seven factors in this experiment (Source Correlation × Order × Polarity × Presence/Absence of Four Sources of Information, *shop*), there are many other potential interactions to consider. Of the 122 interactions involving source correlation, order, and polarity, only 8 (6.5%) were reliable with an alpha level of .05, and only 1 (0.8%) with an alpha level of .01. None were interpretable, and we view the overall pattern of interactions as not deviating from random variation in the data.

Inspecting separately the data of the three source-correlation groups (*sh, so,* and *sp*) revealed only one reliable, predictable interaction (*sh* in the *sh* group). However, this interaction was also reliable in one of the other groups, and within the *sh* group there was no interaction of *sh* stimuli across positively and negatively correlated sets. More concretely, Stimulus 11** always garnered lower than expected judgments based on values given Stimuli 10**, 01**, and 00**. Thus, the overall *sh* interaction cannot be attributed to source correlations.

An Interaction and a Trend, Both Against Additivity

Because the two qualms motivating these studies were not borne out by the data, we focused our attention on information integration and pooled the data of the 34 viewers. Bruno and Cutting (1988) looked at the main effects and interactions in the analysis of variance as potential sources of additivity and nonadditivity. If all main effects and no interactions are reliable, additivity is indicated; that is what was found. In the combined data presented here, all four main effects (*s, h, o, p*) were significant, $Fs(1, 33) > 15.4$, $ps < .001$. Among the many possible interactions, only one first-order interaction (*sh,* reported earlier) and the one highest-order interaction (*shop*) were reliable, $Fs(1, 33) > 8.1$, $ps < .008$. The latter interaction would be expected if the data were subadditive or had other nonlinear trends. Bruno and Cutting (1988) found neither of these interactions, but then they used only 10 viewers rather than 34, and thus their tests had less statistical power.

In addition, following Cutting and Bruno (1988), we looked at the pattern of responses for stimuli with 0, 1, 2, 3, and 4 sources of information. In particular we looked at the differences in the individual data that accrued from successively adding information sources. Assuming scale use was linear (and the results of Experiment 1 suggest it was), we subtracted the value of the 0-source stimulus from the mean of the four 1-source stimuli, then the mean of the 1-source stimuli from the 2-source stimuli, then the mean of the 2-sources from the 3-sources, and finally the mean of the 3-sources from the 4-source stimulus. If information is added, these differences ought to be equal; if, on the other hand, it is subadditive, differences ought to be systematically smaller across comparisons: $(1 - 0) > (2 - 1) > (3 - 2) > (4 - 3)$. Indeed, the four differences were 16.6, 18.1, 12.0, and 10.2, respectively, $F(3, 99) = 5.87$, $p < .001$, generally consistent with a decreasing trend. Moreover, this is essentially the same result found by Cutting and Bruno (1988). The patterns of data for Experiments 1 and 2 here are shown in the middle and right panels of Figure 3.

Modeling Analyses

We then refocused on Massaro's (1988a) paradigm and considered the two competing hypotheses entertained by Massaro, instantiated as the additive model of information integration and FLMP (Equations 1 and 2). We also broadened our approach by looking at two other models. Consider the new models first.

A Partial-Cue Model

Our third model is adapted from Maloney and Landy (1989; Landy, Maloney, & Young, 1991). They presented an averaging model of information integration based on a logical analysis of sources of depth information.[4] For our purposes,

[4] Maloney and Landy (1989) regarded depth as being provided by four classes of information. The first class provides absolute information and includes motion parallax and binocular disparity. That is, with an individual's movement path known and with the distance between the eyes known, absolute depth between the observer and all objects can be computed, at least in principle. These sources of information can be represented by only one variable, their weight given in the equation of combination. The second class provides depth estimates up to a multiplicative scale factor. That is, a source of information might reveal that one object is twice as far away as another, but the two objects could be 1 and 2 m or 15 and 30 m away. Maloney and Landy listed texture gradients and linear perspective as examples of such sources; relative size and height in plane are others. Each of these sources needs at least two variables, one for weighting in perceptual combination and the other for scaling. The third class provides scaled depth information but is subject to reversals and includes kinetic depth (in parallel projection) and shading. These need three variables: one for weighting, one for scaling, and one for sign. Finally, and most important for our situation, the fourth class offers no depth information per se but can only be used to disam-

however, the important feature of their model is that they regard occlusion as giving no information about depth; it can only disambiguate other sources (such as shading) not present in our stimuli. Maloney and Landy devised their model to test perceived amount of depth, not rated depth, and the experimental procedures their model entails include variation of information beyond mere presence or absence. Nonetheless, by collapsing some parameters and assuming perceived depth maps linearly onto rated depth, it can be written as an additive model with four parameters:

$$R(\text{Depth}) = \mathbf{S} + \mathbf{H} + \mathbf{P} + \mathbf{B},$$

$$\text{where } \left\{\begin{array}{l}\mathbf{S} = S \text{ when } s = 1,\\ \mathbf{S} = \text{zero when } s = \text{zero}\end{array}\right\} \text{ etc., for } \mathbf{H}, \mathbf{P}, \mathbf{B}. \quad (3)$$

Three aspects of this application should be noted. First, Maloney and Landy (1989) did not include a background variable, but they did consider retinal disparity as a source of information. Because our stimuli were seen on a flat display scope, retinal disparity is always zero and can thus be included in the background variable. Second, our purpose here is not to try to do justice to the Maloney and Landy model (which also uses robust estimators to de-emphasize outliers in response distributions); instead, we only wish to step outside direct comparisons between FLMP and the additive model by presenting a model related to one in the existing literature. Third, because this model is a subset of the additive model, and because the additional parameter in the additive model is orthogonal to the others, the additive model will always fit data at least as well as this model, and usually better. This model is included here to indicate how much better and to provide additional comparisons with the other, nonadditive models.

A Full-Cue Weighted-Averaging Model

Our fourth model is patterned after one used by Massaro (1987b, pp. 182–183). Adapted to our purposes, it includes occlusion and is written as follows:

$$R(\text{Depth}) = \mathbf{S} + \mathbf{H} + \mathbf{O} + \mathbf{P} + \mathbf{B},$$

$$\text{where } \left\{\begin{array}{l}\mathbf{S} = S \text{ when } s = 1,\\ \mathbf{S} = \text{zero when } s = \text{zero}\end{array}\right\} \text{ etc., for } \mathbf{H}, \mathbf{O}, \mathbf{P}, \mathbf{B} \text{ and}$$

$$\text{where } 0.0 < S,H,O,P < 1.0 \text{ and}$$

$$\text{where } S + H + O + P = 1.0. \quad (4)$$

The major difference between this model and our additive model is the last conditional statement: All weights of manipulated variables sum to 1.0, making it a weighted-averaging model. Because of this constraint the model has only four free parameters. The parameter **P**, for example, becomes the complex of $(1 - S - H - O)$.

Implementing and Verifying the Models

The four hypotheses about information integration were instantiated in four models (Equations 1–4). To accommodate restrictions on FLMP and on the weighted-averaging model, all data were divided by 100, transforming the scale values to a range between 0.0 and 1.0. We implemented the four models using the NONLIN module of SYSTAT (Wilkinson, 1990), which allows specification of a model as an equation (computer subroutine) and iterates through it, minimizing deviations in a data set. We chose the sum of least squared deviations method to fit the data because of its consistency with standard statistical practice. Massaro (1987b, 1988a, 1988b; Massaro & Friedman, 1990) has used root mean squared deviations (RMSD) in a modification of the program STEPIT (Chandler, 1969). SYSTAT does not allow RMSD as an option, but because both methods rely on minimization of squared deviations, we felt our method would be a straightforward transformation of RMSD.

Data of Bruno and Cutting (1988). To insure our method captured the same results as Massaro's RMSD/STEPIT instantiation of FLMP, we ran our versions of the models on the data of Bruno and Cutting (1988, Experiment 1), which Massaro (1988b, Tables 2 and 3) analyzed. Results are shown in Table 4, and the relative deviations for the additive model and FLMP are reasonably well matched to those reported by Massaro: The data of the same 5 subjects are better fit by the additive model, and the data of the other 5 are better fit by FLMP. Means of individuals are shown in Table 4 and in Table 5.

Moreover, and more importantly, the FLMP sums of least square values for the 10 individuals are a straightforward transformation of their RMSD values as given by Massaro (1988a). Our measure simply sums squared deviations, whereas RMSD sums them, divides them by 16 (the number of stimuli), and then takes the square root. After transformation the FLMP residual values reported in Table 4 for Cutting and Bruno (1988, Experiment 1) are the same as those reported by Massaro (1988a), to three significant figures.

As expected, fits of the additive model were superior to the partial-cue model data in all cases, and the fits of FLMP were better in 9 of 10 cases ($p < .02$). In addition, fits of both the additive model and FLMP were superior to the weighted-averaging model in all cases.

Weights and other aspects of model fits. In our FLMP analysis, the weights of the four experimental parameters *SHOP* and the background variable, *B*, were identical to those reported by Massaro (1988a), to three significant figures, adding further evidence that we properly implemented FLMP. However, we had two concerns about Massaro's modeling.

First, as seen in Equation 2, Massaro (1988a) treated motion parallax (p) differently from the other three sources of information. That is, the absence of p is ambiguous information about depth (and should achieve a fuzzy-logical value of 0.50), whereas absence of *s*, *h*, or *o* is information about no depth (and achieves a value closer to 0.0). We wondered what effects this might have on the model fits. Empirically, we determined that this coding procedure had essentially no effect on either the sum of least squares measure or the weights of the other variables (both mean effects did not differ in the third significant decimal). Thus, we kept Massaro's scheme in latter modeling of human data.

Second, Massaro's (1988a) implementation of the additive model did not allow it to have negative weights, whereas the

biguate other information, such as kinetic depth or shading. Maloney and Landy regarded occlusion as one such information source.

analog of negativity (weights below 0.50) was allowed in FLMP. We worried that this might penalize our model unduly, because in our running of the model the data of 3 subjects indicated a negative effect of occlusion.[5] Empirically, however, we determined that allowing negative weights had a small effect on least square values everywhere except for Observer 7 and had essentially no effect on the weights of the other, positively weighted variables. In later modeling we allowed parameters to have negative weights, but because weights themselves are not pertinent to the rest of our discussion, we do not refer to them further.

Table 4
Sum of Least Square Fits for Four Models

Subj	ADD	FLMP	PCUE	WTAVE
Bruno & Cutting (1988), Experiment 1				
01	.103	.110	.179	.185
02	.031	.047	.054	.523
03	.212	.207	.227	.512
04	.119	.105	.144	.831
05	.150	.161	.190	.276
06	.082	.107	.097	.289
07	.148	.145	.315	.879
08	.021	.031	.063	.366
09	.214	.152	.215	.617
10	.082	.076	1.179	.413
Mean of individual fit	.116	.114	.266	.489
Fit to grouped data	.017	.034	.041	.358
Experiment 1				
11	.069	.084	.237	.741
12	.045	.059	.207	.618
13	.177	.142	.448	.288
14	.066	.032	.152	.511
15	.019	.007	.158	1.683
16	.084	.099	.757	.952
17	.047	.049	.086	1.015
18	.204	.221	.206	.327
19	.149	.145	.255	.914
20	.102	.031	.449	.120
21	.054	.053	.090	.229
22	.030	.038	.071	.034
23	.049	.066	.093	.226
24	.291	.145	.990	.941
25	.030	.045	.109	.556
26	.050	.057	.062	.701
Mean of individual fit	.092	.080	.273	.612
Fit to grouped data	.012	.024	.075	.537
Experiment 2				
27	.034	.035	.058	1.153
28	.121	.092	.604	.122
29	.053	.066	.299	1.210
30	.247	.251	.397	.320
31	.118	.152	.198	.690
32	.265	.280	.329	.601
33	.210	.087	.383	.799
34	.049	.078	.130	.734
35	.020	.043	.293	.514
36	.042	.070	.457	.300
37	.067	.066	.924	.088
38	.266	.265	.413	.601
39	.075	.083	.107	.188
40	.074	.018	.148	1.263
41	.099	.095	.104	1.122
42	.180	.177	.181	.719
43	.097	.095	.166	.100
44	.146	.140	.157	1.293
Mean of individual fit	.120	.116	.297	.656
Fit to grouped data	.015	.028	.052	.375

Note. ADD is the additive model (Equation 1), FLMP is Massaro's (1987b, 1988a) fuzzy-logical model of perception (Equation 2), PCUE is the partial-cue model without an occlusion parameter (Equation 3) adapted from Maloney and Landy (1989), and WTAVE is the weighted-averaging model (Equation 4) adapted from Massaro (1987b, pp. 178–183). Subj = subject.

Table 5
Mean Sum of Least Square Fits for the Additive Model and FLMP Across Experiments and Simulations

Data source	Additive	FLMP	FLMP advantage
Bruno & Cutting (1988) Experiment 1	.116	.114	.002
Experiments 1 & 2	.107	.099	.008
Simulation 2			
Random data	.932	.924	.009
Regressed to range of Experiments 1 & 2	.117	.098	.0198
Simulation 3: Added random error (%)			
0	.00000	.00066	−.00066
2.5	.00053	.00061	−.00008
5.0	.00218	.00224	−.00006
10.0	.00861	.00865	−.00004
20.0	.03447	.03446	.00001
40.0	.1380	.1386	.0004
60.0	.3136	.3120	.0016

Note. FLMP = fuzzy-logical model of perception.

Fitting Models to the New Individual and Group Mean Data

We next fit the four models to the individual data of Experiments 1 and 2. Results are also shown in Table 4. We were roundly thwarted in our efforts to differentiate the additive model and FLMP: Of the 34 subjects' data, 18 were better fit by the additive model, and the other 16 were better fit by FLMP. Across the 34 subjects the fits of FLMP were slightly but not significantly better, $t(33) = 1.143$, $p > .26$. The mean advantage of FLMP was 0.008, also shown in Table 5. Again, the additive model and FLMP were superior to the others. In all cases, both fit individual data better than the partial-cue model. The additive model bested the weighted-averaging model in all cases, and FLMP fit better in 33 of 34 cases.

Interestingly, the fit of the additive model to each of the three group data sets was somewhat better than FLMP, as shown at the bottom of Table 4. In each case the residuals for the additive model were about half those for FLMP. Moreover, across the group data of 44 subjects in the three experiments, the residuals of the additive model were only about one-third of those for FLMP (0.0091 vs. 0.0271, respectively).

[5] In addition, three other subjects' data showed negative effects of the background variable. A negative weight of occlusion, for example, might arise when Stimulus 1100 is seen as having more depth than Stimulus 1110, because occlusion is seen as indicating that the three panels lie on top of one another, reducing depth here and elsewhere that is otherwise seen when relative size and height in plane are present.

Again the additive model and FLMP fit the group data better than the other two models.

Preliminary Conclusions

Three conclusions can be drawn from these studies. First, the partial-cue and weighted-averaging models did not fare well against the additive model and FLMP. Insofar as the partial-cue result applies to the full Maloney and Landy (1989) model, we think their omission of occlusion may be a mistake. Second, with respect to the additive model and FLMP, the cluster of results is inconsistent. The modeling results of the group mean data favored the additive model, the modeling results of the individual data were indeterminate, and the results of analyses of variance on the raw data and on difference scores were against the additive model and therefore, it would seem, favored FLMP. Third, it seemed unlikely that we would replicate twice, once in each experiment reported here, the indecisiveness Massaro (1988a) found between models in fitting individual data. To try to make more sense of these antinomies, we investigated the numerical properties of the models.

Models of Information Integration and Their Fits to Simulated Data

As noted earlier, Cutting and Bruno (1988) interpreted FLMP as a model that captured *sub*additivity in data. From inspecting some of the model fits of individual data in Experiments 1 and 2, and from reading Massaro (1988b) and Massaro and Friedman (1990), it became clear that fitting subadditivity is not all that FLMP is good at. Thus, we ran three classes of simulations involving FLMP and the additive model.

Simulation 1: FLMP Fits Most Monotonic Functions

Exponentials

We created a series of 15 functions anchored at values of 0.001 and 0.999, as shown in the upper-left panel of Figure 4. Again, 16 stimuli with four orthogonal sources of information were used. All four sources were treated identically; that is, there was no analog to motion parallax here, which in its absence had been treated as noninformative (given a value of 0.50).

As with the real stimulus sets, there were 1 stimulus with no information, 4 with one source, 6 with two sources, 4 with three sources, and 1 with four sources. Stimuli were given values that fit exponentials, plotted on an abscissa with 0 through 4 sources of information. The 15 functions fit the following equation:

$$R(\text{depth}) = n^a / 4^a, \tag{5}$$

where n is the number of sources of information. The exponent a took the values 0.01, 0.15, 0.4, 0.6, 0.7, 0.8, 0.9, 1.0, 1.11, 1.25, 1.43, 1.66, 2.5, 6.67, and 100.0 for the 15 functions, respectively. Every stimulus with a given number of sources had the same value. Thus, there was no variance at any level on the abscissa.

The fits of the two models to the 15 sets of simulated data are shown in the upper-middle panel of Figure 4. Notice three results: First, the additive model was superior to FLMP for only five data sets, obviously those most nearly linear. Using only the sign and ignoring the magnitude of fit differences, the upper-right panel of Figure 4 shows that, within the region covered by possible exponential functions, FLMP fit better throughout 80.3% of the area. Second, even when the additive model was superior the two fits were quite similar. For these functions the additive model's advantage was never greater than 0.016; for the 44 subjects included in Table 4, only 5 of the 23 subjects favoring the additive model showed differences greater than this value. Third, all functions deviating significantly from linearity were fit extremely well by FLMP.

Psychometric Functions

We next generated a family of psychometric functions, from a linear function to a step function. These are shown in the bottom-left panel of Figure 4, again plotted for 0 through 4 sources of information. Again, there was no variance at any level on the abscissa, except for the step function (for which three sources were coded as having values of 0.999 and three others were coded with values of 0.001). The six functions fit the following logistics equation:

$$R(\text{depth}) = e^{(a + n \times b)} / [1 + e^{(a + n \times b)}], \tag{6}$$

where e is 2.718, n is the number of sources of information, and a and b are the two parameters for a logit function. Values for a across the first 5 functions were −2.7, −3.3, −4.0, −5.3, and −48.7, respectively; and values for b were 1.35, 1.63, 2.0, 2.65, and 24.4.

Again, the additive model fit only a few functions better than FLMP, and even when it did the difference between the two models was small. Ignoring the differences in magnitude of fits, FLMP fit better within 95.7% of the region covered by these functions. Moreover, FLMP fit all functions extremely well. It is clear FLMP has more scope than the additive model; indeed, it appears to be a very powerful model. How powerful is it?

Simulation 2: FLMP Fits Random Data

From Simulation 1 we learned that FLMP fits many more types of functions than does the additive model. We wanted to explore this idea further. By generating many random data sets, we hoped to sample the population of all possible data functions. Random numbers between 0.001 and 0.999 were generated for each of the 16 stimuli, 1,000 times each. Of the 1,000 simulated data sets, 608 were better fit by FLMP; only 392 were better fit by the additive model. Across the set FLMP's mean sum of least squares advantage was 0.009, $F(1,999) = 104.6$, $p < .0001$, as shown in Table 5. As seen in Table 5, this difference is at the high end of the range of the fits to the subject data in Experiments 1 and 2 and in Bruno and Cutting (1988, Experiment 1).

One might suggest, however, that FLMP's superiority in this domain could be ignored. For example, the magnitudes of least square values in the random simulations are about eight times those in Experiments 1 and 2, and one might

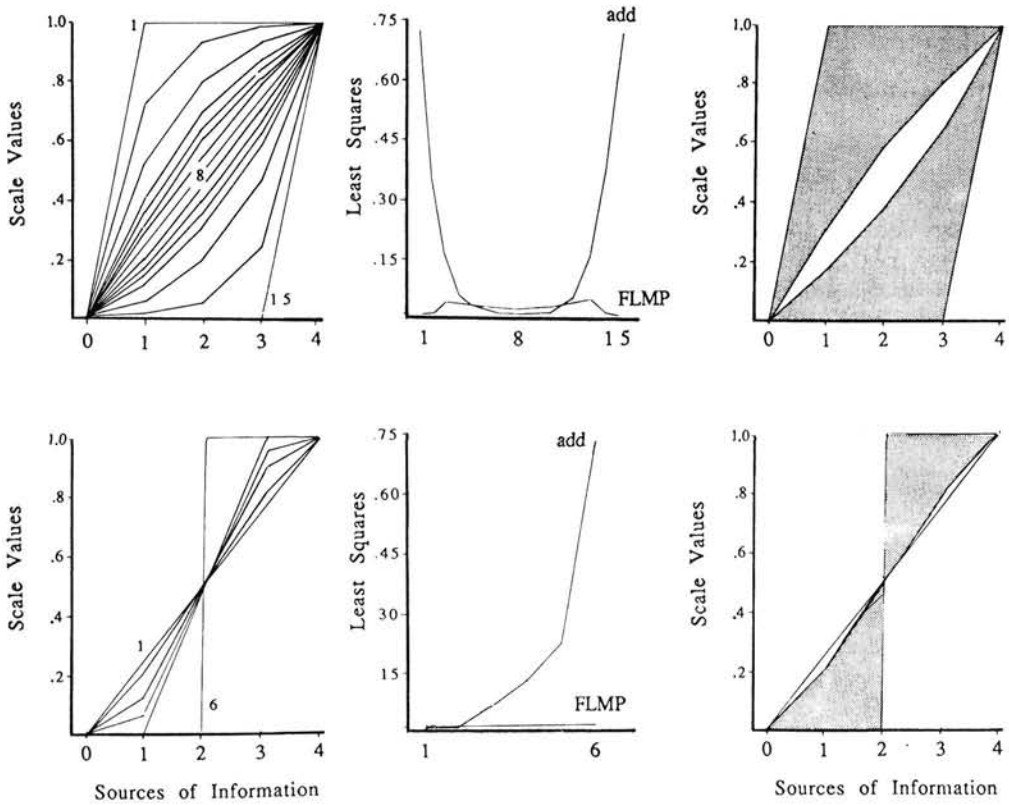

Figure 4. A comparison of the additive model and fuzzy-logical model of perception (FLMP) in fitting exponential and psychometric functions. (The upper panels show the exponential functions and their measures of sum of least squares fit by the additive model [add] and FLMP and the area of the functions better fit by FLMP. The lower panels show approximations to psychometric functions, the fits, and the area of the functions better fit by FLMP. Functions were generated by Equations 5 and 6.)

expect the difference between the two models to inflate with the magnitude of the least squares fit. However, the reverse was true: The difference between model fits was slightly negatively correlated with the magnitude of least square values; $r = -.16$, $t(998) = 5.07$, $p < .0001$, as shown in Figure 5. This means larger differences between model fits occur with smaller sum of least square (and hence RMSD) values. As shown in Table 5, when regressed back to the range of fits of Experiments 1 and 2, the difference between the models in favor of FLMP is 0.0198. This value is more than double the difference found in any of the experiments. Thus, it is clear from the first 2 simulation studies that FLMP is at an advantage compared with the additive model.

Simulation 3: FLMP Absorbs Random Error

If FLMP accrues an advantage in random data, it ought to begin to reveal it in data that start out as linear (additive) but then have increasing amounts of random error added to them. To explore this idea we started with a perfectly additive function. Stimulus 0000 was fixed at a value 0.30; all four stimuli with one source were fixed at 0.40, the six stimuli with two sources were fixed at 0.50, the four stimuli with three sources were fixed at 0.60, and Stimulus 1111 was fixed at 0.70. An additive model fits these data perfectly (with a sum of least squares residual of 0.00000); FLMP also fits them very well (0.00066).

Adding Random Error to an Additive Function

What happens when random error is added? To explore this idea we generated random values to add or subtract from the baseline values for each of the 16 stimuli, a method similar to that used by Collier (1985). Each deviation was constrained within a range, and the proportion added or subtracted was yoked across three conditions. Thus, the depth rating is generated as follows:

$$R(\text{depth}) = 0.30 + 0.10 \times n + r, \qquad (7)$$

where n is again the number of sources of information and r is a random number generated within a rectangular distribution. Across three conditions these distributions spanned ±0.05, 0.10, and 0.20—10%, 20%, and 40% of the range of values, respectively. These bounds correspond to grand mean RMSD values of about 0.029, 0.058, and 0.115 from the linear trend.[6] The yoking occurred as follows: if, in the 10%

[6] Grand mean RMSD values were computed by creating 20 equal-sized bins within the range of random values, squaring the mean for each bin, averaging those squared values, and taking the square root of the average.

condition, a value of 0.035 was added to one stimulus, a value of 0.070 was added to it in the 20% condition, and a value of 0.140 was added to it in the 40% condition. If a different, negative value were added to one stimulus, the resulting additions might be −0.024, −0.048, and −0.096, respectively.

Across 200 triads of simulated data, the patterns of results were straightforward, as shown in Table 5. With increasing amounts of random error added to the data, FLMP begins to accrue its advantage over the additive model. In particular, the interaction between the two models and the amount of random error added to the baseline data were reliable, $F(2, 398) = 3.782$, $p < .024$. This interaction was not due simply to increases in the residual value of the sum of least squares. When the differences are rescaled for increases in variance (dividing the 20% variability data by 4, dividing the 40% variability data by 16, then comparing both with the 10% condition), the effect is still reliable, $F(2, 398) = 3.702$, $p < .026$. More concretely, in the 10% variation condition 110 of 200 data sets favored the additive model, but in the 20% condition this dropped to 103 of 200, and in the 40% condition it dropped further to 87 of 200.

Different Additions of Random Error

The previous result seemed worth replicating with different amounts of random error. We added three more data points on a possible function between 0% and 100% added error, two smaller (2.5% and 5%, or adding ±0.0125 and 0.025 to the linear function in Equation 5) and one larger (60%, or adding ±0.30). These correspond to grand mean RMSD values of about 0.007, 0.014, and 0.173 from the linear trend, respectively.

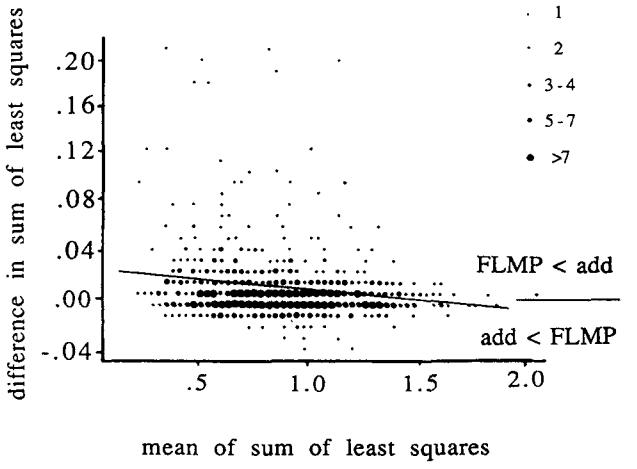

Figure 5. A scatterplot of the 1,000 comparisons of model fits in Simulation 2. (The difference in sum of least square residuals of fuzzy-logical model of perception [FLMP] and the additive model to random data sets is plotted as a function of the magnitude of the mean sum of least square residual of the two models. The size of the dot indicates the number of comparisons at each point on the graph. FLMP < add indicates that FLMP fit better than the additive model. That most values are positive shows that FLMP is at an advantage compared with the additive model; that the slope is negative shows that FLMP is of a particular advantage when residuals are smallest.)

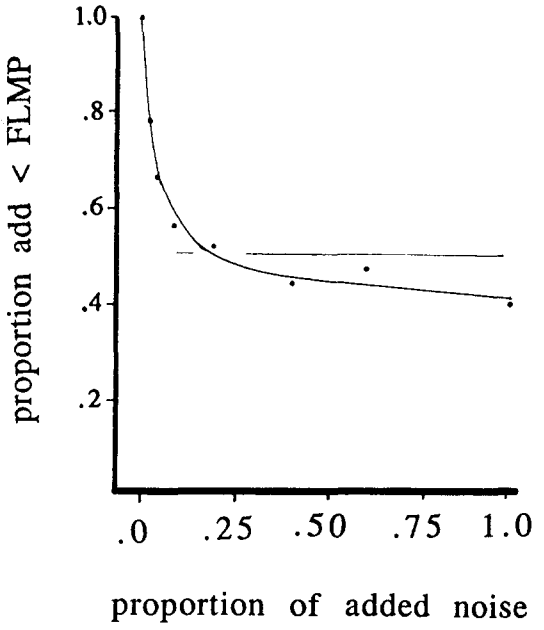

Figure 6. The proportion of simulated trials in which the additive model (add) fits better than fuzzy-logical model of perception (FLMP) as a function of how much random error is added to a linear function.

Across 200 more triads of simulated data, results were similar, as shown in Table 5. Again, with increasing amounts of error, FLMP accrues an advantage, both in the raw data, $F(2, 398) = 8.11$, $p < .001$, and in the variance scale-transformed data, $F(2, 398) = 13.3$, $p < .001$. The strength of the effect is greater here due to the increased range of error added. Again, and more concretely, the additive model was superior to FLMP for 149 of 200 data sets in the 2.5% condition and 129 of 200 data sets in the 5% condition, but only 91 of 200 in the 60% condition.

The overall results are combined in Figure 6, along with the logical consideration of 0% random error and the results of Simulation 2. The pattern shows how, by adding random error to a linear function, FLMP attains its advantage over the additive model. The data are well fit by a negative exponential; the more error that is added, the more often FLMP will be superior. In particular, when the error reaches 22%, or a grand mean RMSD value of about 0.063, from the linear trend, FLMP becomes superior even when the mean data are linear.

Discussion

When compared with the additive model, FLMP appears to capitalize on fitting data by three methods. First, as shown in Simulation 1, even when the additive model fits better, FLMP is not far behind. Second, also shown in Simulation 1, FLMP fits certain nonlinearities in data, accruing great advantage over the additive model. Thus, in truly linear domains, both models fit well; in certain nonlinear domains, only FLMP fits well. Neither of these methods, however, is able to account for the increasing proportion of data sets fit better by FLMP as variability increases. Thus, the third way

it accrues its advantage, shown in Simulation 3, is simply to absorb increasing variability. The last method shows how a model with broader scope is at an advantage in fitting relatively noisy individual data.

Simulation 4: Comparative Fits of the Other Models to Random Data

In our earlier modeling we found that the partial-cue and weighted-averaging models fit the data of our human subjects quite poorly compared to both FLMP and the additive model. Thus, it seemed worth investigating how they would fit random data. As in Simulation 2, we generated random data sets with entries between 0.001 and 0.999 for each of 16 stimuli. Again, the stimuli corresponded to a $2 \times 2 \times 2 \times 2$ set, with each comparison representing the presence or absence of four sources of information. We generated 300 random data sets and then fit them with the four models: additive, FLMP, partial cue, and weighted averaging.

Median fits of the four models were 0.934, 0.946, 1.028, and 1.872, respectively. Again FLMP fit the random data better than the additive model, $F(1, 299) = 25.4$, $p < .001$, this time in 181 of 300 sets (60.3%), a result similar to that of Simulation 2. FLMP was also superior to the other two models, besting the partial-cue model in 268 of 300 comparisons and the weighted-averaging model in 287 of 300.[7] As expected, the additive model fit the random data better than the partial-cue model in all 300 cases, and the additive model bested the weighted-averaging model in 287 of 300 cases. Finally, the partial-cue model fit the random data better than the weighted-averaging model in 282 of 300 cases.

Measuring Simplicity and Discounting Scope

Tensions Between Simplicity and Scope

Given two self-consistent and fruitful scientific theories of roughly equal scope, the standard view in science is to prefer the more parsimonious, or simpler, theory. When these theories can be instantiated as models, there is a preference for the simpler model (Reichenbach, 1949). Ockham's razor dictates as much. But how do we measure simplicity? There are many difficulties in and few agreements about its measure (e.g., Goodman, 1972; Kuhn, 1977; Popper, 1959; Quine, 1976). Nonetheless, we consider three general proposals for arbitrating simplicity and their relation to the concept of scope.

Parameter Count, Simplicity, and Scope

First, Jeffreys (1957, 1961), among others, proposed that to compare models or theories one should sum the number of parameters in the equation, including the degree of an exponent or derivative that may be contained in it. The lower the count, the simpler the model or theory. Most views of modeling in contemporary psychology follow Jeffrey's dictum. In this view simplicity and scope are independent attributes of a theory.

Any concern about the number of parameters in a model is well justified. Models with more free parameters are likely to fit data better than those with fewer free parameters. At the limit, considering a set of data with n observations, any model with $n - 1$ parameters should fit perfectly. Collier (1985) and many others have noted that when building models, researchers face a trade-off between the number of parameters and closeness of fit to the data. Thus, when comparing different models researchers often try to make comparisons between those with equal numbers of free parameters.

FLMP and the additive model have the same free parameter count—5. This measure of simplicity aside, then, one would ordinarily turn to considerations of scope. In our analyses we found FLMP superior to the additive model in fitting exponentials, psychometric functions, and random data and about equal to the additive model in fitting human data. These considerations and results would imply that FLMP and the additive model are equally simple but that FLMP has more scope. Thus, under this construal of simplicity and scope, FLMP is the better model and represents the better theory.

Falsifiability, Simplicity, and Scope

A second and perhaps less standard approach to simplicity is Popper's (1959): Competing theories should be compared on grounds of falsifiability. Thus, given two theories in the same domain, we should prefer the simpler, but where simplicity is defined as a property that places the greatest restrictions on the world. Thus, we should prefer the theory, or its model, that is more easily proved wrong (falsified). Kemeny (1955) proposed a similar idea. Notice that this view contrasts with the previous one in that scope and simplicity are correlated, rather than independent; simple theories by definition have less scope, and are preferable to, more complex theories.

Popper's (1959) idea can be applied to the current situation as follows: Given the results of Simulation 1 showing the superiority of FLMP over the additive model in fitting exponential and psychometric functions, and given the results of Experiments 1 and 2 here and those of Bruno and Cutting (1988, Experiment 1) showing the veritable equality of the two models to fit human depth-judgment data, one should prefer the additive model over FLMP because it is more falsifiable. Thus, under this construal of simplicity and scope, the additive model is better and represents a better theory.

Given opposite conclusions from these two approaches to modeling from the perspective of the philosophy of science, we searched for an alternative. Consistent with the second approach, we think broad scope is not always a positive quality in a model or theory, but consistent with the first approach we think the concept of simplicity ought to be measured independently of scope. To discuss our view we need first a new measure of simplicity.

---

[7] In certain speech perception tasks, Massaro (1987b, pp. 178–183) found the FLMP to fit human data much better than a weighted-averaging model. Given that FLMP fits random data so much better than a weighted-averaging model, we find it unsurprising that it also fits human data better.

Equation Length and Simplicity

A third approach to measuring simplicity comes from information theory and the economics of transmitting data. In general, the data could be auditory signals, visual images, or simply strings of numbers. One area within this field is concerned with predicting the next datum or a missing datum in a finite data sequence. Within this area algorithmic information theory (Chaitin, 1977; see also Komolgorov, 1968) is concerned about the minimal length of the algorithm, equation, or computer program for making this prediction. The shortest such program is deemed the simplest, and longer programs are deemed more complex.[8]

There are two important corollaries of this information-theoretic approach. First, once programs have been set to their minimal form, longer programs are deemed more complex and are expected to represent more data sets (have greater scope) than shorter ones. Second, given equally good predictions for a particular sequence of data, a longer program is considered more complex than a shorter one.

Notice that the relation between an information-theoretic algorithm and the data set it represents parallels the relation between a psychological model and a subject's data. In both cases the model must predict the data and do so economically. Thus, if this approach is applied to psychological modeling, equation length might be a predictor of how well models fit data. Notice further that, in contrast to parameter counting, length considerations will include all elements in the algorithm, not simply the free parameters. In this manner, comparing Equations 1 and 2, one can see that FLMP is more complex than the additive model, despite the fact both have five free parameters. In addition, the additive model and the other two models would be about equally complex despite their varying numbers of parameters.

But how is equation length measured? As a first approximation program complexity can be measured as the number of ASCII characters needed to run the program unambiguously in a general-programming language on a general-purpose computer. Such an approximation works well for equations but has potential difficulties in coding other logical and extralogical operations. To accommodate this problem we follow the general lead of Goodman (1972) and his predicate calculus for simplicity; we represent each possible operation or predicate with a single symbol. We also count each syntactic marker as a single symbol. To be more concrete, a comma, a bracket, a *zero*, a *where*, and an *etc.* statement each count as a single symbol.

Measuring Relative Effects of Parameter Count and Equation Length

The most straightforward way to apply the criterion of equation length to our situation is to count the elements in the right-hand side of Equations 1 through 4, representing the four models. Two counts will be considered, one without and one with the conditional statements beneath each equation line. The counts without conditionals are simply the total number of the ASCII characters; those with conditionals are ASCII-based except where words are used to set up restrictions on the calculations. Each word is counted as a single symbol (or character), as it might be represented as such in the compiled form of an algorithm. We prefer the measure without conditionals, because conditionals are not strictly involved in the computation; they only clarify or set bounds on it. The relative equation lengths of each model with and without conditionals are shown in Table 6, along with their numbers of free parameters.

To determine the effect of equation length in fitting data we used multiple regression on difference scores. The two independent variables were the differences in length of the models considered two at a time and the differences in the number of free parameters in those models considered two at a time. The dependent variable was the difference in least-squared fits of the two models to the data of each of the 44 subjects shown in Table 4. Given four models to consider, six pairwise model comparisons are made for each of the 44 subjects, yielding a total of 264 comparisons.

Consider first the comparison between parameter count and the measure of equation length without conditionals. As expected, the multiple correlation using both as independent variables was statistically reliable; $R = .26$, $F(2, 261) = 9.37$, $p < .0001$. However, the partial correlations are more interesting. In particular, equation length was a reliable predictor of fit; $r = -.19$, $F(1, 261) = 4.87$, $p < .03$, whereas the number of free parameters was not, $r = -.09$, $F(1, 261) = 1.14$, $p > .28$. To be sure, the range of difference in parameters is highly constrained, but the difference values in length of equations is functionally constrained as well. If one dummy codes equation length (1 = long, for FLMP; 0 = short, for other models) and reruns the regression analysis, the partial correlation is essentially the same; $r = -.21$, $F(1, 261) = 6.36$, $p < .01$. In neither case was the difference between partial correlations for length and parameter count reliable.

If conditional statements are included in measurements of equation length, the multiple correlation is again reliable; $R = .22$, $F(2, 261) = 6.83$, $p < .001$. However, equation length is this time not reliable ($r = -.15$, $p > .80$), and parameter count is reliable ($r = -.24$, $p < .01$). Equation length did poorly here because the weighted-averaging model is quite long with its conditionals but fitted the data relatively poorly.

A Tentative Conclusion

Our analysis of equation length and number of free parameters indicates that, at least in some circumstances, a re-

[8] In perceptual psychology equation length has played a role in formal models of perception and in our understanding of the concept of simplicity. This tradition arose in Gestalt psychology and is best represented by the minimum principle (Hochberg, 1957, 1988) where good gestalts were regarded as simpler than other configurations. Over the last 20 years, this idea has been promoted in the structural information theory of Leeuwenberg (1971, 1982; see also Cutting, 1981; Cutting & Proffitt, 1982; Restle, 1979). According to this theory, we perceive the simplest description of the possibilities in the physical stimulus, where simplicity can be measured by equation length: Simpler equations are shorter. It is as if the perceiver constructs many possible representations or models of the object or event and perceives only the object or event with the simplest equation. There are problems with the minimum principle (Hatfield & Epstein, 1985; Hochberg, 1988), but it remains an attractive idea.

Table 6
Equation Lengths and Parameter Counts for the Four Models

Model	Equation	Equation length (no. characters) Without conditionals	Equation length (no. characters) With conditionals	No. of free parameters
Additive	1	9	37	5
FLMP	2	51	121	5
Partial-cue	3	7	34	4
Weighted-averaging	4	9	69	4

Note. FLMP = fuzzy-logical model of perception.

searcher might profitably pay as much attention to the form of a model's equation (without conditionals) as to the number of parameters it contains. On the basis of our analysis, we think it is possible that FLMP fits data because it is longer and hence more complex than other models.

Controlling for Scope to Measure Selectivity: A New Way to Compare Models

Models can be thought to have two data-fitting properties. The first is scope, which we define as the measure of a model's ability to fit all possible data functions. These are represented by a broad sample of functions filled with random numbers. The second is selectivity, or the ability of a model to capture particular patterns of interest in the data to the exclusion of all others. Because a large collection of random functions will contain a few data patterns of interest to the researcher and many patterns of no interest, we are interested in a model's ability to select all and only those patterns of interest.

Thus, we suggest that in any situation where psychological models are to be fit to data they should be compared simultaneously in two ways: The relative fits of the models to individual data sets of interest ought to be compared against their relative fits to random data. In this manner, any advantage of a model in fitting random data (its scope) can be neutralized, and the researcher can focus on the residual and differential advantage of the model's fits to the patterned data of interest (its selectivity).

What we propose is a binomial test (e.g., Siegel, 1956) comparing selectivity with scope, where the probabilities for the obtained fits to psychological data are compared against those predicted by the fits to random data. The standard formula is:

$$z = [(x \pm 0.5) - NP] \,/\, \text{sqrt}[NP(1 - P)], \qquad (8)$$

where x is the number of observations favoring a particular model in comparisons across subject data sets (corrected for continuity), N is the number of subjects, and P is the probability that model has in besting its competitor, computed from many runs on random data sets. The outcome of six such tests is shown in Table 7.

Notice three new results: First, the fits of the additive model and FLMP to the data of 44 subjects from Bruno and Cutting (1988) and from Experiments 1 and 2 here do not statistically differ from their fits to random data. Thus, although the scopes of the two models differ somewhat, that difference (as represented by relative fits to random data) is not sufficiently great to prefer one model over the other. Second, the additive model's and FLMP's overwhelming advantages over the partial-cue and weighted-averaging models in fitting the human data do not differ from their advantages in fitting random data. Thus, on the basis of these comparisons, we conclude that nothing of psychological value can be said about any of those comparisons involving FLMP and the additive model. Third, although the weighted-averaging model fits random data poorly compared with the partial-cue model, it fits our human data remarkably well. Thus, by our account, the weighted-averaging model has narrow scope but high selectivity compared with the partial-cue model.

General Discussion

Massaro's (1987b) paradigm for experimental psychology consists, for our purposes, of three parts: (a) a focus on the analysis of multiple sources of information and (b) a comparison of alternative models in (c) the analysis of individual data. On the basis of more than 10 years of research, Massaro has used his paradigm and found broad support for his model, FLMP. Like Massaro, we embrace the first idea, but we worry about the second and often the third in the current context of his paradigm.

Models and Their Hidden Properties

FLMP has broad scope; it is a powerful model. It has shown noted success in fitting data in many domains, including attention, reading, letter recognition, and speech perception.

Table 7
Binomial Tests for the Fits of Models to Human Data (N = 44) Against Their Relative Fits to Random Data

Model comparison		Fit: Observed	Fit: Predicted	z score
Additive	< FLMP	23	17.2	1.63*
Additive	< Partial Cue	44	44	—
Additive	< Weighted Averaging	44	42.1	.29
FLMP	< Partial Cue	42	38.0	1.53
FLMP	< Weighted Averaging	43	42.1	.29
Partial Cue	< Weighted Averaging	35	41.4	−3.74**

Note. Because better fits are indicated by smaller residuals, additive < FLMP, (fuzzy-logical model of perception) indicates that the additive model fits the data better. Predictions are based on results from Simulations 2 and 4.
* $p < .11$. ** $p < .001$.

However, on the basis of our simulations, we think its success could be based, at least in part, on a property models should not have: It can fit random error. To be sure, variability is part of all human data, but the intent of any model should be to capture systematic trends in data sets, not surreptitiously to capture random error within them. We also think our worry is ironic: Massaro (1988b) warned psychologists against models with too much power, or (in our terms) scope.

The FLMP's ability to fit randomness is, we think, a heretofore hidden property. Hidden properties tend, erroneously, to mold our views of what is being modeled. Uttal (1990) captured this idea best:

> Sometimes models may superimpose their own properties on our concept of the object being modeled. Hence, the modeler's conception of the object may reflect a property that the object does not actually possess. In this sense, therefore, the model may be more (rather than less) than the object it represents. (pp. 195–196)

Thus, FLMP may not capture information integration better than competing models; instead, it may simply capture all patterns of data better, and that property may have masqueraded as a representation of psychological process. At minimum, at least within Massaro's paradigm, we think that comparing numbers of parameters in two models is no longer an adequate method to begin model comparisons. We think running the models on random data is a necessary test of their relative scope.

Consequences of Comparing Models With Unequal Scope

Without baseline comparisons of models run on random data, we think any approach that proceeds by comparing models with matched numbers of parameters may be in jeopardy. Two general effects seem possible. First, when one model may have a moderate advantage over another (such as FLMP compared with the additive model), it may use that advantage in its fits to data of individuals but not of groups, as demonstrated in the results of Experiments 1 and 2 and in the results of Bruno and Cutting (1988). In general, group data are smooth; individual data are more noisy. Because FLMP fits random noise better than the additive model, it will be at an advantage in the noisier, individual comparisons. Second, when one model has a large advantage (such as FLMP compared with the partial-cue and weighted-averaging models), it will win in virtually all comparisons when run on human data because it generally fits all possible functions better. Thus, the model's performance on human data cannot be attributed to the idea that it reflects psychological process; it reflects only the general scope of a much more powerful model.

The problem of the relative scope in modeling, at least as used in Massaro's paradigm, is inherited from the straightforward application of Platt's (1964) idea of *strong inference*. Strong inference is like a horse race, or more particularly a match race between two horses. One horse (one model) is pitted against another, and the null hypothesis is scratched from entry. Winner takes all. Because there is no way to measure statistical error in the margin of a single win, the horses (models) are run many times (run on sets of data from different individuals). It now appears however, that some models (like horses) are differentially handicapped, not so much by their apparent abilities to fit patterned data (to run fast) but by heretofore hidden differences in scope (stewardship at the track).

Conclusions

On the basis of this collection of investigations we have three conclusions. First, given the now reasonably extensive data base of viewers' judgments about the layout of objects in depth beyond that offered by Bruno and Cutting (1988) and Massaro (1988a), there is still no empirical reason to choose between two contenders—an additive model and FLMP. To be sure, finding interactions and other indications of subadditivity impugns the additive model, but as shown by the model fits to individual data, these results do not inherently support FLMP. In fact, given the superior scope of FLMP the additive model may be marginally favored, as shown in Table 7.

Second, if a researcher wishes to compare two or more models and how they fit the data of human participants, we suggest he or she ought to first consider how the models fit random data. This consideration will partial out a model's scope, its ability to fit all patterns of data that are of no interest to a researcher, from its selectivity to particular patterns of data that are of interest. The binomial test we propose is one such partialing method.

Third, we suggest that parameter counting is not the only way to evaluate the fairness of comparisons between models. Considerations of equation length might be given equal weight. We believe that FLMP may have garnered some of its advantage over the other three models due to its increased complexity, as measured by the length it takes to specify.

References

Anderson, N. H. (1981). *Foundations of information integration theory.* San Diego, CA: Academic Press.

Anderson, N. H. (1982). *Methods of information integration theory.* San Diego, CA: Academic Press.

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

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

Burton, G., & Turvey, M. T. (1990). Perceiving lengths of rods that are held but not wielded. *Ecological Psychology, 2,* 295–324.

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

Chandler, J. P. (1969). Subroutine STEPIT—Finds local minima of a smooth function of several parameters. *Behavioral Science, 14,* 81–82.

Collier, C. E. (1985). Comparing strong and weak models by fitting them to computer-generated data. *Perception & Psychophysics, 38,* 476–481.

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

Cutting, J. E. (1986). *Perception with an eye for motion.* Cambridge, MA: MIT Press.

Cutting, J. E. (1991a). Four ways to reject directed perception. *Ecological Psychology, 3,* 25–34.

Cutting, J. E. (1991b). Why our stimuli look as they do. In G. Lockhead & J. R. Pomerantz (Eds.), *Perception of structure* (pp. 41–52). Washington, DC: American Psychological Association.

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., & Millard, R. M. (1983). Three gradients and the perception of flat and curved surfaces. *Journal of Experimental Psychology: General, 113,* 198–216.

Cutting, J. E., & Proffitt, D. R. (1982). The minimum principle and the perception of absolute, common, and relative motions. *Cognitive Psychology, 14,* 211–246.

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

Fodor, J. A. (1983). *Modularity of mind.* Cambridge, MA: MIT Press.

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.

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

Hatfield, G., & Epstein, W. (1985). The status of the minimum principle in the theoretical analysis of visual perception. *Psychological Bulletin, 97,* 155–186.

Hochberg, J. (1957). Effects of the Gestalt revolution: The Cornell symposium. *Psychological Review, 64,* 73–84.

Hochberg, J. (1988). Visual perception. In R. C. Atkinson, R. J. Herrnstein, G. Lindzey, & R. D. Luce (Eds.). *Stevens' handbook of experimental psychology* (2nd ed., pp. 195–276). New York: Wiley.

Jeffreys, H. (1957). *Scientific inference* (2nd ed.). Cambridge, England: Cambridge University Press.

Jeffreys, H. (1961). *Theory of probability* (2nd ed.). London, England: Oxford University Press.

Kemeny, J. (1955). Two measures of simplicity. *Journal of Philosophy, 52,* 722–733.

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

Knudsen, E. I., & Konishi, M. (1979). Mechanisms of sound localization in the Barn Owl (*Tyto alba*). *Journal of Comparative Physiology, 133,* 13–21.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago.

Kuhn, T. S. (1974). Second thoughts on paradigms. In F. Suppe (Ed.), *The structure of scientific theories* (pp. 459–482). Urbana: University of Illinois Press.

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.

Künnapas, T. (1968). Distance perception as a function of available visual cues. *Journal of Experimental Psychology, 77,* 523–529.

Landy, M. S., Maloney, L. T., & Young, M. J. (1991). Psychophysical estimation of the human depth combination rule. *Sensor Fusion III: 3-D Perception and recognition, Proceedings of the SPIE, 1383,* 247–254.

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.

Maloney, L. T., & Landy, M. S. (1989). A statistical framework for robust fusion of depth information. In W. A. Perlman (Ed.), *Visual communication & image processing IV, Proceedings of the SPIE, 1199,* 1154–1163.

Marr, D. (1982). *Vision.* New York: Freeman.

Massaro, D. W. (1984). Building and testing models of reading processes. In P. D. Pearson (Ed.), *Handbook of reading research* (pp. 111–146). New York: Longman.

Massaro, D. W. (1985). Attention and perception: An information-integration perspective. *Acta Psychologica, 60,* 211–243.

Massaro, D. W. (1987a). Integrating multiple sources of information in listening and reading. In D. A. Allport, D. G. MacKay, W. Prinz, & E. Scheerer (Eds.), *Language perception and production: Shared mechanisms in listening, speaking, reading, and writing* (pp. 111–129). San Diego, CA: Academic Press.

Massaro, D. W. (1987b). *Speech perception by ear and eye: A paradigm for psychological research.* Hillsdale, NJ: Erlbaum.

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

Massaro, D. W. (1988b). Some criticism of connectionist models of human performance. *Journal of Memory and Language, 27,* 213–234.

Massaro, D. W. (1989). Review of *Speech perception by ear and eye: A paradigm for psychological inquiry. Behavioral and Brain Sciences, 12,* 741–794.

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

Massaro, D. W., & Hary, J. M. (1986). Addressing issues in letter recognition. *Psychological Research, 48,* 123–132.

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, CA: Academic Press.

Platt, J. R. (1964). Strong inference. *Science, 146,* 347–353.

Pomerantz, J. R., & Lockhead, G. R. (1991). Perception of structure: An overview. In G. R. Lockhead & J. R. Pomerantz (Eds.), *The perception of structure: Essays in honor of Wendell R. Garner* (pp. 1–20). Washington, DC: American Psychological Association.

Popper, K. R. (1959). *The logic of scientific discovery.* New York: Basic Books.

Quine, W. V. (1976). On simple theories of a complex world. In *The ways of paradox* (pp. 255–258). Cambridge, MA: Harvard University.

Reichenbach, H. (1949). *Theory of probability.* Berkeley, CA: University of California.

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

Schilpp, P. A. (Ed.). (1974). *The philosophy of Karl Popper.* Peru, IL: Open Court.

Siegel, S. (1956). *Nonparametric statistics.* New York: McGraw-Hill.

Suppe, F. (Ed.). (1977). *The structure of scientific theories.* (2nd ed.). Urbana: University of Illinois Press.

Uttal, W. R. (1990). On some two-way barriers between models and mechanisms. *Perception & Psychophysics, 48,* 188–203.

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

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology Monograph Supplement, 9*(2, Pt. 2), 1–27.

Received February 25, 1991
Revision received March 16, 1992
Accepted March 30, 1992 ■