Scale Convergence as a Principle for the Study of Perception

MICHAEL H. BIRNBAUM

At one time, scientists accepted the view that celestial and earthly events were governed by different laws. It was believed that the elements of earth, fire, water, and air and their chemical reactivity did not "generalize" to the heavens. Instead, the heavens were supposed to be composed of a different element and to be inert. It can be argued that the dismissal of this belief in favor of a simpler hypothesis in the time of Galileo made possible great new developments in the science of astronomy.

If different laws characterize earthly and heavenly events, then how can earthlings learn about the stars? The simpler view that the heavenly objects are composed of the same elements and obey the same laws as earthly objects has led to modern astronomy. Astronomers now assume that the physical laws that can be tested in simple, tiny laboratory experiments on earth apply to the objects we see in the sky. This premise has not really been tested, it has been assumed. It gives us powerful leverage for studying objects that cannot yet be manipulated or directly observed, whose existence and behavior are known only through a tiny sample of electromagnetic waves detected here on earth. There is a lesson in this story for psychologists, and I will try in this chapter to spell it out.

The principle of scale convergence in psychology may be an analogous assumption of coherence that may prove useful to the study of perception and judgment. The principle of scale convergence asserts that measurements interlock laws of different empirical relationships. To introduce the application of scale convergence in perception, the next section discusses algebraic models with emphasis on size constancy as an example. It will be assumed that the same subjective scale of distance ties together several phenomena of size perception.

Algebraic models of perception and judgment

In many situations, simple algebraic laws have been proposed to explain psychological phenomena. For example, psychologists explain the moon illusion, in which the moon seems larger on the horizon than it does overhead, with the following premises:

$P_1: S = RD$
$P_2: R_H = R_Z$
$P_3: D_H > D_Z$,

where S is perceived size; R is retinal image size; and D is subjective distance. Premise 1 assumes subjective size is the product of retinal size and subjective distance. Premise 2 asserts that the retinal image sizes are equal for the moon on the horizon (H) and zenith (Z). Premise 3, attributed to Ptolemy, is that the subjective distance to the horizon exceeds the subjective distance to the zenith. The three premises imply $R_H D_H > R_Z D_Z$;

therefore $S_H > S_Z$. Thus, subjective size on the horizon exceeds subjective size of the zenith.

Coherent theory

The first premise, $S = RD$, can be regarded as a psychological "law" with potential for great generality. Some of the phenomena that can be explained using this premise are illustrated in Fig. 1. It can be used in explanations of geometric illusions like the Müller-Lyer, and Ponzo illusion, the Ames room, apparent size in stereoscopic views, size constancy, and the subjective sizes of afterimages.

A theoretical system is described as coherent if the same premise can be used in the explanations of a variety of phenomena. The premise, $S = RD$, can be used in explanations of the phenomena illustrated in Fig. 1, given suitable assumptions concerning R and D. To account for the Müller-Lyer, Ponzo, and Baldwin illusions, it is assumed that $R_A = R_B$, but $D_A > D_B$. To account for the sizes of stereo images it is assumed that depth is a function of retinal disparity. If the red filter is worn on the right eye and the green on the left, then the two squares will "fuse" to a larger size when the red square is to the left of the green rather than vice versa. In the Ames room, it is assumed that $D_A = D_B$ but $R_B < R_A$ hence $S_A > S_B$. However, in the normal room it is assumed that retinal size is a function of visual angle, $R = b\Phi_S/\Phi_D$. Thus, subjective distance is proportional to objective distance, $D = a\Phi_D$. Thus, subjective sizes are equal ($S_A = S_B$) when physical sizes are equal, ($\Phi_{SA} = \Phi_{SB}$), i.e., size constancy.

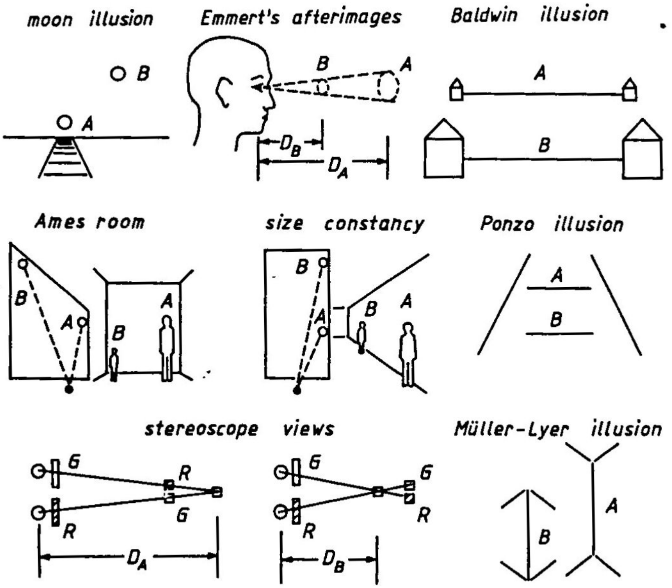

Fig. 1: Some phenomena that can be explained using the premise, $S = RD$. In the moon illusion, the moon on the horizont looks larger than the moon on the zenith. Emmert's after images vary with retinal size and "projection" distance. In the Baldwin illusion, Ponzo illusion, and Müller-Lyer illusion, line A seems larger than line B, even though actual lengths are equal. In the trapezoidal Ames room, a person seems to change size while moving about, although size constancy is maintained (approximately) in rectangular rooms. In stereo views through Red (R) and Green (G) filters, Red and Green squares fuse to different sizes (and distances) as the horizontal distance between them is varied.

Emmert's law extended

Consider the following explanation of the apparent sizes of after-images:

$P_1: S = RD$
P_2: R depends only on flash size
P_3: $D = H(\Phi_D)$
P_4: "S" $= J[S] + \varepsilon$,

where H is the function relating subjective distance (D) to objective distance (Φ_D), J is the function relating subjective size to judged size, and ε is a random error component with a mean of zero. In the experiment, the subject is exposed to a flash presented to one eye, resulting in a circular after-image, The subject then "projects" the after-image onto surfaces of varying actual distance (Φ_D), using both eyes. The visual angle of the inducing flash is assumed to affect retinal size of the flash, R. The subjective distance (D) is assumed to depend on actual distance, though the function need not be linear. The judged size is denoted "S", and is assumed to be a monotonic function of subjective size plus a random error component.

Hypothetical errorless data for this experiment are presented in Tab. 1. Each entry in the table represents judged apparent size in centimeters. The rows of the table represent after-images produced by flashes with differing visual angle (different R). The columns represent different actual distances. The data are perfectly consistent with the model, with H negatively accelerated and J a similarity function. Note that subjective size is directly proportional to physical distance up to 160 cm, but thereafter increases as a negatively accelerated function.

Tab. 1: *Hypothetical data: Perceived sizes of after-images*

	Actual distance (cm)						
After-image size	10	20	40	80	160	320	640
1	.5	1	2	4	7	13	24
2	1	2	4	8	14	26	48
3	1.5	3	6	12	21	39	72

Each entry represents the judged size of an after-image projected to various distances.

Explanations such as the above can be constructed for each of the phenomena of Fig. 1 using the premise $S = RD$ and appropriate premises for R and D.

Comprehensive theory of D is lacking

Although the separate explanations of the phenomena in Fig. 1 satisfy the philosophical criteria for explanations, the system as yet lacks a complete theory of D. In each explanation, the appropriate premise concerning D was plugged in to make the deduction work. In the explanation of Tab. 1, H was estimated from the data. We would like a complete theory from which the appropriate premise for D can be deduced. What we have is a set of mini-theories and a list of variables (cues) that affect D.

Experiments can be done to test mini-theories such as "a proximal rectangle will be perceived as a distal rectangle with equidistant sides" or "converging lines are perceived as parallel lines receding in depth" or "the resultant value of D is the weighted average of the values of D implied by each cue." These mini-theories of distance are testable, but are not yet complete enough to predict the value of D in any visual environment. For this reason, the values of D are derived from the data in an experiment like the one represented in Tab. 1.

Scale convergence for subjective distance

In the absence of a theory of D, it seems reasonable to require that the same transformation $D = H(\Phi_D)$ should appear in several phenomena. Thus, the difference between the explanation of size-constancy and the Ames room illusion is supposed to be due to different H functions in the two rooms. The principle of scale convergence demands that the H functions be taken seriously and forces an additional constraint that requires coherence in a theoretical system.

To illustrate the idea, consider two experiments that can be conducted in the same visual environments (e.g., the normal and Ames rooms). One can "project" after images as in Tab. 1, yielding an H function for each visual environment. One can also present actual objects in the same environment and ask the same subjects to judge their sizes. Consider the following theory for size judgments:

$$S = RD$$
$$R = b\Phi_S/\Phi_D$$
$$D = H(\Phi_D)$$
$$\text{“}S\text{''} = kS + \varepsilon\,.$$

Tab. 2 shows some hypothetical judgments of sizes of actual objects as a function of size and distance that are perfectly consistent with the above premises. Notice that size constancy is maintained for distances less than 160 cm, but for greater distances, objects grow smaller with distance. However, the pattern in Tab. 2 is implied by the H function obtained in Tab. 1. Thus, Tabs. 1 and 2 can be represented with the same H function, using the premise $S = RD$.

Tab. 2: *Hypothetical data: Test of size constancy*

Actual size (mm)	Actual distance (cm)						
	10	20	40	80	160	320	640
50	51	51	51	51	44	41	38
100	102	102	102	102	89	83	76
200	153	153	153	153	134	124	115

Each entry represents the judged size of an actual object. Viewing conditions the same as for Tab. 1.

Despite the lack of a specific theory from which the H function can be deduced, the system of theories is enhanced by being interlocked by the same scale of subjective distance. It is also important to emphasize that $D = H(\Phi_D)$ does not represent an invariant "psychophysical law" for subjective distance, because it only applies to the particular viewing conditions in the "normal" room. In the Ames room, a different

H function would be obtained, but scale convergence would require that this new *H* function should apply to judgments of the size of afterimages projected into the room as well as judgments of actual objects. The *H* function would be further enhanced by the number of phenomena it can be used to explain. For example, subjects could also be asked to judge the "differences in distance" between each pair of positions in the room; hopefully, the same *H* function would reproduce this matrix of "difference" judgments.

The size-distance issue has been simplified to use as an illustration of how coherent laws can be interlocked by the principle of scale convergence. Discussions of this issue, including potential difficulties for the depth explanation of geometric illusions can be found in Gregory (1978), Gogel (1968), Kaufman (1974), and Rock (1975).

Scale convergence as a constraint

Scale convergence can be considered a theoretical constraint that can cause one theoretical system to be preferred to another. Kepler's solar system was preferred to Ptolemy's because Kepler's could be deduced from principles of Newtonian physics that could be tested in the laboratory. Similarly, in psychology a coherent system of laws interlocked by common scales should be preferred to a system that requires new scales for every new situation. Scale convergence can be used in this way to resolve otherwise unsolvable problems that can arise when mathematical models are tested.

Outcomes of model tests

When mathematical models of perception or judgment are tested, three outcomes can occur: (a) The data can show ordinal violations, in which case the model can be rejected. (b) The data may be numerically consistent with the model. (c) The data may be numerically inconsistent with the model, but ordinally consistent. In this case, the experimenter must decide whether to transform the data to fit the model or not. But if transformation is permitted in case c, then the experimenter also needs to consider the possibility that data that fit the model (case b) should also be transformed.

Suppose an investigator hypothesized that the row and column factors of Tab. 3 combine additively. In Tab. 3, the data show ordinal violations that contradict the additive model. The additive model implies that it should be possible to find values of a_i and b_j such that $R_{ij} > R_{kl}$ whenever $a_i + b_j > a_k + b_l$. This implies that if $R_{ij} > R_{il}$ then $a_i + b_j > a_i + b_l$; therefore $b_j > b_l$. Adding a_k to both sides, it follows that $a_k + b_j > a_k + b_l$; therefore $R_{kj} > R_{kl}$. Similarly, $R_{ij} > R_{kj}$ implies $R_{il} > R_{kl}$. However, this property of *independence* is violated repeatedly in Tab. 3. For example $R(1, 1) < R(2, 1)$ whereas $R(1, 5) > R(2, 5)$. If the rank order in Tab. 3 were well-established (based on enough consistent data), there would be no doubt that the additive model should be rejected.

Tab. 3: *Hypothetical data: Ordinal violations*

Level of	Level of B				
A	1	2	3	4	5
1	0	4	8	12	16
2	9	10	11	12	13
3	2	4	6	8	10

In Tab. 4, the data are numerically consistent with the additive model. In other words, it is possible to find values of a_i and b_j such that $R_{ij} = a_i + b_j$. For Tab. 4, let $a_1 = 1$, $a_2 = 2$, $a_3 = 3$, $b_1 = 1$, $b_2 = 2$, $b_3 = 3$, $b_4 = 4$, and $b_5 = 5$. These values perfectly reproduce the entries in the table when added. Thus, the additive model remains consistent with the data in Tab. 4.

Tab. 4: *Hypothetical data: Perfect fit*

Level of	Level of B				
A	1	2	3	4	5
1	2	3	4	5	6
2	3	4	5	6	7
3	4	5	6	7	8

The hypothetical data in Tab. 5, however, pose a problem to an investigator who has hypothesized an additive model. The data are ordinally consistent with an additive model, but numerically inconsistent. That is, it is possible to solve for values of a_i and b_j such that $R_{ij} > R_{kl}$ if $a_i + b_j > a_k + a_l$. However, it is not possible to find a_i and b_j such that $R_{ij} = a_i + b_j$. Put another way, there exists a nonlinear monotonic transformation, T, such that $T(R_{ij}) = a_i + b_j$. For Tab. 5, T is the logarithmic transformation; the logs of the numbers in Tab. 5 are additive. On the other hand, the raw data are perfectly numerically consistent with the multiplicative model, $R_{ij} = a_i^* b_j^*$, where $a_1^* = 1$, $a_2^* = 2$, $a_3^* = 3$, $b_1^* = 1$, $b_2^* = 2$, $b_3^* = 3$, $b_4^* = 4$, and $b_5^* = 5$. In sum, the data in Tab. 5 are ordinally consistent with either an additive or multiplicative model, but they are numerically inconsistent with the additive model.

Tab. 5: *Hypothetical data: Metric violations*

Level of	Level of B				
A	1	2	3	4	5
1	1	2	3	4	5
2	2	4	6	8	10
3	3	6	9	12	15

The problem for the investigator is as follows: given the data of Tab. 5 is there any reason to prefer the additive over the multiplicative model? The criterion of scale convergence allows an additional constraint. The basic idea is as follows: Suppose there are two empirical phenomena to be explained. Suppose there are two rival theories of these phenomena. The principle of scale convergence that one prefers a theory in which the measurements of the stimuli are the same for both phenomena. If the values of b for Tab. 5 were known to be 1, 2, 3, 4, and 5, the multiplicative model would be preferred to the additive for Tab. 5. Similarly, the explanation of the sizes of afterimages and size constancy (or inconstancy) should postulate the same scale of subjective distance, D, for the same actual distances under the same viewing conditions. Tabs. 1 and 2 conform to the criterion of scale convergence given the theories postulated.

Brief review of studies of scale convergence

The principle of stimulus scale convergence asserts that the scale values (measurements) of the stimuli are independent of the task and model interrelating the measurements. By requiring this additional constraint, it becomes possible to differentiate theories that would otherwise be equivalent. Tab. 6 lists studies in which the principle of scale convergence was used to reduce the number of hypotheses that are plausible for a given situation involving two or more empirical relationships. Birnbaum (1974a, 1982) and Birnbaum and Veit (1974) discuss the principle of scale convergence further and relate it to previous conceptions of convergent operationism.

Tab. 6: *Selected studies of scale convergence*

Reference	Dimension	Task
Birnbaum (1974a, Exp. 3)	likableness	D, C, DC
Birnbaum & Veit (1974)	heaviness of lifted weights	R, D, A
Rose & Birnbaum (1975) Birnbaum (1974b)	magnitude of numbers	R, D M (context)
Birnbaum & Mellers (1978)	positions of U.S. Cities	"ratios of easterliness" "differences of easterliness" "ratios of westerliness" "differences of westerliness"
Birnbaum (1980)	(Review)	
Birnbaum & Elmasian (1977)	loudness	R, D
Elmasiam & Birnbaum (1979)	pitch	R, D
Birnbaum (1978)	darkness of dot patterns	R, D
Veit (1978)	darkness of grays	R, D, RD
Hagerty & Birnbaum (1978)	likableness	R, D, RR, RD, DR, DD
Birnbaum (1982)	darkness of dots	R, D, RR, RD, DR, DD
Mellers & Birnbaum (1982a)	darkness	R, D, (context)
Mellers & Birnbaum (1982b)	class performance	C, M, (context)

D = "difference" task, C = "combination" task,
DC = "difference between combinations",
A = "averaging" task, RE = "ratios of easterliness",
DW = "differences in westerliness", M = "magnitude" rating task,

Impression formation

Birnbaum developed scale convergence as a criterion for rescaling in 1970 in order to assess whether derivations from a simple model of impression formation were "real" or due to "nonlinear judgment bias". Judgments of the likableness of hypothetical persons described by adjectives were inconsistent with the constant-weight averaging model that was believed at the time to be acceptable (Anderson, 1979). However, it was unclear whether the major deviations of fit should be attributed to an interactive integration process or to a nonlinear judgment function. To resolve this problem, Birnbaum (1974a, Experiment 3) asked subjects to rate the differences in likableness of combinations of the adjectives. It was possible to reject the following model:

$$P_1: C_{ij} = J_C[(w_0 s_0 + w_1 s_i + w_2 s_j)/(w_0 + w_1 + w_2)]$$
$$P_2: D_{ij} = J_D[s_j^* - s_i^*]$$
$$P_3: s_j = s_j^* ,$$

where C_{ij} = rating of likableness of a person described by the combination of adjectives i and j, which have scale values of s_i and s_j, s_0 and w_0 are the scale value and weight of the initial impression, and w_1 and w_2 are weights. The functions J_C and J_D are assumed to be strictly monotonic judgment functions, and D_{ij} is the judged "difference" in likableness between adjectives i and j, which are assumed to have scale values of s_i^* and s_j^*. The third premise is the assumption of scale convergence, $s = s^*$.

Tabs. 7 and 8 show matrices of C_{ij} and D_{ij} obtained by Birnbaum (1974a) together with the theoretical interpretations of the subtractive model of "differences" and the constant-weight averaging model of "combinations." The rank order in Tab. 7 implies that $s_2 - s_1 > s_5 - s_3$ whereas Tab. 8 implies that $s_2^* - s_1^* < s_5^* - s_3^*$. These ordinal contradictions (and others) require rejection of the theory consisting of P_1, P_2 and P_3.

Instead, the data were consistent with a configural-weight model for combinations in conjunction with the subtractive model for comparisons. This interpretation allows preservation of scale convergence. Birnbaum (1982, Section F) gives a more detailed presentation of the ordinal analysis of scale convergence for this issue.

Tab. 7: *Mean ratings of likableness*

	Level of A				
Level of B	1	2	3	4	5
1	1.54 (0)	2.10 (a)	2.50 ($a + b$)	2.76 ($a + b + c$)	3.45 ($a + b + c + d$)
2	2.10	2.92 ($2a$)	3.82 ($2a + b$)	4.44 ($2a + b + c$)	5.08 ($2a + b + c + d$)
3	2.50	3.82	5.15 ($2a + 2b$)	5.90 ($2a + 2b + c$)	6.72 ($2a + 2b + c + d$)
4	2.76	4.44	5.90	6.53 ($2a + 2b + 2c$)	7.25 ($2a + 2b + 2c + d$)
5	3.45	5.08	6.72	7.25	7.90 ($2a + 2b + 2c + 2d$)

Each entry is the mean judgment of likableness of a person described by both A and B. Each off-diagonal cell is averaged over six pairs of adjectives; 600 judgments from 300 subjects (data from Birnbaum, 1974a, Experiment 1). Algebraic symbols give additive representation, $C_{ij} = J_c[s_i + s_j]$, with $s_1 = 0$, $a = s_2 - s_1$, $b = s_3 - s_2$, $c = s_4 - s_3$, $d = s_5 - s_4$. Arrows represent inequalities showing that $a > b + c$ and $a > c + d$.

Tab. 8: *Mean ratings of differences*

	Level of A				
Level of B	1	2	3	4	5
1	0	1.18 (a^*)	1.86 ($a^* + b^*$)	2.49 ($a^* + b^* + c^*$)	3.20 ($a^* + b^* + c^* + d^*$)
2	−1.18	0	.92 (b^*)	1.64 ($b^* + c^*$)	2.43 ($b^* + c^* + d^*$)
3	−1.86	− .92	0	.53 (c^*)	1.54 ($c^* + d^*$)
4	−2.49	−1.64	− .53	0	.85 (d^*)
5	−3.20	−2.43	−1.54	− .85	0

Each number is the mean judgment of difference in likableness, A—B. Each cell is averaged over six pairs of adjectives, 180 judgments from 90 subjects (from Birnbaum, 1974a, Experiment 3). Algebraic symbols give subtractive representation, $D_{ij} = J_D[s_j^* - s_i^*]$, with $s_1^* = 0$, $a^* = s_2^* - s_1^*$, $b^* = s_3^* - s_2^*$, $c^* = s_4^* - s_3^*$, $d^* = s_5^* - s_4^*$. Arrows represent inequalities showing that $a^* < b^* + c^*$ and $a^* < c^* + d^*$.

A scale-free test verified the interaction (Birnbaum, 1974a, Experiment 4). In this experiment, subjects judged "differences between combinations". The three matrices of data can be represented by the following model, which contains two types of scale convergence. Both scale values and subjective impressions are assumed to be independent of the task.

$$D_{ij} = J_D[s_j^* - s_i^*]$$
$$C_{ij} = J_C[\Psi_{ij}]$$
$$DC_{ijkl} = J_{DC}[\Psi^*_{ij} - \Psi^*_{kl}]$$
$$\Psi_{ij} = \Psi^*_{ij} = I[s_i, s_j]$$
$$s_i = s_i^* ,$$

where J_D, J_C, and J_{DC} are approximately linear, and I is a configural-weight model (Birnbaum, 1982). In the configural-weight model, the weight of an item depends on its rank order within the set of items to be integrated (Birnbaum & Stegner, 1979). The worst trait receives extra weight in impression formation and moral evaluations (Birnbaum, 1972, 1974a, 1982).

Contextual effects in ratings

Birnbaum (1974c) presented subjects with sets of integers from 108 to 992 and asked them to rate the magnitude of the numbers on a ninepoint scale. Nine different groups of subjects received different distributions of the stimuli. For example, in one distribution there were eight stimuli between 100 and 200 in another distribution, only one stimulus fell in this interval. Plotting judgments against stimulus magnitude led to nine different curves that differed drastically from one another. Birnbaum (1974c) fit a version of Parducci's (1965, 1974) range-frequency theory. For the conditions of Birnbaum's experiment the theory can be written:

$$G_{ik} = aF_k(\Phi_i) + s_i \, ,$$

where G_{ik} is the category rating of stimulus Φ_i in distribution k; $F_k(\Phi_i)$ is the cumulative density of stimuli less than or equal Φ_i in context k; a is estimated from the data, and s_i are the scale values of the stimuli, also estimated from the data. When the psychophysical function is assumed to be strictly monotonic and error free, the F_k function is known. Therefore, if this equation can be fit to the data, it provides a basis for deriving scale values from contextual effects. It was found that the data were well-described by this model in terms of a single scale for number.

"Ratios" and "differences" of numbers

The scale for number derived by Birnbaum (1974c) was negatively accelerated, consistent with findings by investigators using other methods (e.g., Rule & Curtis, 1973). Rose and Birnbaum (1975) asked undergraduates to divide a line segment so that either the "ratio of the two lines would equal the ratio of the two numbers" or so that the "difference in the two lines would be proportional to the difference between the two numbers." They found that subjects gave virtually the same responses for the "ratio" and "difference" tasks irrespective of the instructions, despite a careful training procedure that explained mathematical properties of actual ratios and differences and a test to check understanding of these concepts (Rose & Birnbaum, 1975, Experiment 2).[1]

The data were consistent with the theory that subjects used the same operation for both tasks, which could have been either a ratio or a subtractive operation. In order to decide between these interpretations, Rose and Birnbaum (1975) applied the scale convergence criterion to state the following two theories of three data sets:

ratio theory:

$$\begin{aligned} R_{ij} &= J_R[s_j^*/s_i^*] \\ D_{ij} &= J_D[s_j^*/s_i^*] \\ G_{ik} &= aF_k(\Phi_i) + s_i \\ s^* &= s \, ; \end{aligned}$$

subtractive theory:

$$\begin{aligned} R_{ij} &= J_R[s_j^* - s_i^*] \\ D_{ij} &= J_D[s_j^* - s_k^*] \\ G_{ik} &= aF_k(\Phi_i) + s_i \\ s^* &= s \, . \end{aligned}$$

Rose and Birnbaum (1975) found that the ratio model led to a scale of numbers s^*, that was positively accelerated relative to physical value, and positively accelerated relative to s. The subtractive theory led to a single scale, $s = s^*$, that was negatively accelerated relative to physical number. Thus, the ratio theory was rejected in favor of the subtractive theory.

This example illustrates how scale convergence permits an interesting contrast between two theories that would otherwise be equivalent. It also reveals that if one wished to

[1] Quotation marks are used to distingush tasks given to the subject from theories. Quotations are used for "ratio" judgments but not actual ratios. Judged "ratios" may or may not fit the ratio model.

retain the ratio interpretation it would be necessary either to revise range-frequency theory or to give up the premise of scale convergence. The same difficulties confront the relative ratio model, $(s_j/(s_i + s_j))$, which is ordinally equivalent to the ratio theory.

Psychophysical "averaging"

When subjects are asked to judge the "average" value of several psychophysical stimuli, what model describes the combination process? Several experiments indicated that ratings of "averages" violate the constant-weight averaging model (Parducci, Thaler, & Anderson, 1969; Birnbaum, Parducci, & Gifford, 1971). However, a few other studies obtained data that were interpreted as consistent with this model (Andersoń, 1972; 1979; Weiss, 1972).

To investigate the issue, Birnbaum and Veit (1974) applied the scale convergence criterion to the comparison of "difference" judgments and "average" judgments. The "average" heaviness shows a convergent interaction similar to that previously found for loudness and length. The data were *not* consistent with the following:

$$\begin{aligned} D_{ij} &= J_D[s_j^* - s_i^*] \\ A_{ij} &= J_A[w_0 s_0 + w_1 s_i + w_2 s_j)/(w_0 + w_1 + w_2)] \\ s_j^* &= s_j , \end{aligned}$$

where A_{ij} is the rated "average" and J_A is the strictly monotonic judgment function. Instead, the data were consistent with the interpretation that the interaction was "real":

$$\begin{aligned} D_{ij} &= J_D[s_j^* - s_i^*] \\ A_{ij} &= J_A[I(s_i, s_j)] \\ s_j^* &= s_j , \end{aligned}$$

where I is the configural-weight model, and both J functions are approximately linear. As a further check on the judgment functions, A_{ij} was plotted against s_i^*. Most models for "averaging" imply that the "average" of two equal stimuli should be a linear function of the scale value of the stimulus. Birnbaum and Veit (1974) found that A_{ii} was very nearly a linear function of s_i^*, consistent with the theory that J_A was linear, and therefore that the interaction was "real". In sum, the principle of scale convergence in this case provided an indication that the deviations from the constant-weight model of "averaging" should not be attributed to the response scale, but rather to the combination process itself.

"Ratios" and "differences"

Birnbaum and Veit (1974) asked subjects to judge "ratios" and "differences" of heaviness, in addition to "averages". It was initially expected that the two judgments could be transformed to fit the following two-operation model:

$$\begin{aligned} D_{ij} &= J_D[s_j^* - s_i^*] \\ R_{ij} &= J_R[s_j/s_i] \\ s_j^* &= s_j , \end{aligned}$$

where D_{ij} and R_{ij} are judgments of "differences" and "ratios", J_D and J_R are strictly monotonic judgment functions, and s^* and s are the two scales. In principle, if a single scale accommodates both difference and ratio operations, the scale attains ratio scale uniqueness, i.e., only a similarity transformation would allow the scale to successfully reproduce both rank orders using the two corresponding operations (Krantz, Luce, Suppes, & Tversky, 1971).

Tab. 9: *Four theories of "ratio" and "differences"*

Task	Theories			
	two operations		one operation	
	simple	biased	ratio	subtractive
R	A/B	$(A/B)^m$	A/B	exp (A—B)
D	A − B	A − B	log(A/B)	A − B

In each theory, the response is assumed to be a linear function of the table entry. All of the above theories assume scale convergence and imply bilinearity for "ratios" and parallelism for "differences".

However, Birnbaum and Veit (1974) found that the data were consistent with a simpler model:

$$R_{ij} = J_R[s_j - s_i]$$
$$D_{ij} = J_D[s_j^* - s_i^*]$$
$$s_j^* = s_j ,$$

where only one operation is assumed to characterize both comparison tasks. Thus, the data provided nontrivial support for a prior conjecture of Torgerson (1961) that judges compare the two stimuli in the same way, irrespective of instructions.

Four special cases of these models with the *J* functions specified are listed in Tab. 9. The two-operation model assuming *J* linear implies that marginal means for corresponding stimuli should be linearly related, contrary to the data (Birnbaum & Veit, 1974). The "biased" two operation theory with a power function for magnitude estimation implies that marginal mean log "ratios" should be a logarithmic function of marginal mean "differences", contrary of data of nine experiments (Birnbaum, 1980). The one operation theories in Tab. 9 imply this relationship should be linear.

Birnbaum (1980) reviewed nine studies that have investigated "ratios" and "differences". Dimensions studies included loudness and pitch (Birnbaum & Elmasian, 1977; Elmasian & Birnbaum, 1979), darkness of grays or dot patterns (Veit, 1978; Birnbaum, 1978), easterliness and westerliness of U.S. cities (Birnbaum & Mellers, 1978), and likableness of adjectives (Hagerty & Birnbaum, 1978). These studies yielded results consistent with the hypothesis that judges use the same operation to compare stimuli whether instructed to judge "ratios" or "differences". The data from these studies were closely approximated by the following model:

$$R_{ij} = a \exp [s_j - s_i]$$
$$D_{ij} = b[s_j^* - s_i^*] + c$$
$$s^* = s ,$$

where *a*, *b*, and *c* are empirical constants, and the judgment function for magnitude estimations of "ratios" is approximated by the exponential function. An alternative one-operation ratio theory would be consistent with "ratio" and "difference" data but is tested by a further extension.

Extension of ratio-difference problem

The principle of scale convergence has been added to other constraints in order to investigate alternative theories of the ratio-difference problem. Of great concern is the question, can the ratio theory be modified to explain judgments of "ratios"?

Veit (1978), Hagerty and Birnbaum (1978) and Birnbaum (1982) have investigated "ratio" and "difference" tasks along with tasks involving the comparison of two stimulus relations. For example, the subject can be shown four stimuli (A, B, C, and D) and asked to judge the "ratio of the difference" between the first two relative to the difference between the second two ((A — B)/(C — D)). These four-stimulus tasks allow one to compare a larger number of district theories, four of which are listed in Tab. 10.

Tab. 10: *Selected theories of stimulus comparison*

Theory				
Task	model = task	subtractive	ratio	transformation
R	A/B	A — B	A/B	A/B
D	A — B	A — B	A/B	A/B
RR	(A/B)/(C/D)	(A — B) — (C — D)	(A/B)/(C/D)	A/B/C/D
RD	(A — B)/(C — D)	(A — B)/(C — D)	(A/B)/(C/D)	(a — b)/(c — d)
DR	(A/B) — (C/D)	(A — B) — (C — D)	A/B — C/D	A/B/C/D
DD	(A — B) — (C — D)	(A — B) — (C — D)	(A/B)/(C/D)	(a — b)/(c — d)

For the transformation theory, a = logA, b = logB, etc.

Several theories of stimulus comparison that account for such tasks were proposed and discussed by Birnbaum (1978, 1979, 1982). Eisler's (1978) transformation theory is discussed by Birnbaum (1979, 1982). The subtractive theory gave the best account of the data of these experiments. According to this theory, subjects compare two stimuli by subtraction whether instructed to judge "ratios" or "differences". Two differences are also compared by subtraction whether the subject is instructed to judge "differences of differences", "differences of ratios", or "ratios of ratios". However, when instructed to compute "ratios of differences", the subjects use this model. The theory can be written as follows:

$$\begin{aligned}
R_{ij} &= J_R[s_j - s_i] \\
D_{ij} &= J_D[s_j - s_i] \\
RR_{ijkl} &= J_{RR}[(s_j - s_i) - (s_k - s_l)] \\
DR_{ijkl} &= J_{DR}[(s_j - s_i) - (s_k - s_l)] \\
DD_{ijkl} &= J_{DD}[(s_j - s_i) - (s_k - s_l)] \\
RD_{ijkl} &= J_{RD}[(s_j - s_i)/(s_k - s_l)].
\end{aligned}$$

Scale convergence is assumed across all six tasks. Veit (1978) and Hagerty and Birnbaum (1978) derived separate scales from each model for each task and showed that they were linearly related. They found that other theories led to violations of scale convergence. Birnbaum (1982) fit the models to all of the data simultaneously using one set of scale values, and found that the subtractive theory (above) gave the best account of all of the data.

Reverse or inverse attributes

The relationship between loudness and softness of tones, lightness and darkness of grays, etc. provides another application of scale convergence. It seems appealing to suppose that the scale values of tones are independent of the task to judge loudness or softness and that they are only mapped differently into responses.

Birnbaum and Mellers (1978) asked subjects to judge "ratios" and "differences" of easterliness and westerliness of U.S. cities. They found that the data were inconsistent with a ratio model in which distances from zero points were compared. Instead the data were consistent with the assumption that there is only one mental map (one scale) with different judgment functions. Data were well-fit by the model:

$$\begin{aligned} DE_{ij} &= a(s_j - s_i) \\ RE_{ij} &= \exp(s_j - s_i) \\ DW_{ij} &= a(s_i - s_j) \\ RW_{ij} &= \exp(s_i - s_j), \end{aligned}$$

where DE, RE, DW, and RW are "differences" and "ratios" of easterliness and westerliness, respectively. The model predicts that "ratios" of easterliness and westerliness are reciprocally related. All four matrices can be reproduced using the same mental map. Reciprocal relationship between loudness and softness has not been rejected *a priori* by psychophysical theorists, but it seems a very unattractive theory for the mental map. By analogy, the subtractive theory for "ratios" and "differences" of easterliness and westerliness seems attractive for other inverse attributes such as lightness and darkness, etc.

Contextual effects in comparison

In the previous applications, scale convergence was regarded as a necessary condition and the attitude was to reject a model rather than give up the premise of scale convergence. The principle can also be regarded as a testable proposition that may be rejected. When the model is well-established and a plausible theory implies the scales should change, then scale convergence seems more an empirical issue than a principle to be assumed. Experiments by Mellers and Birnbaum (1982a, b) illustrate this use of scale convergence.

Mellers and Birnbaum (1982b) asked subjects to rate the darkness of dot patterns. Six dot patterns (12, 18, 27, 40, 60, 90 dots) were common to two different contexts of spacing. In the positively skewed context there were five extra patterns with between 14 and 25 dots. In the negatively skewed context, the five extra patterns had between 45 and 85 dots. The usual contextual effects occurred in the ratings, consistent with Parducci's (1974) range-frequency theory.

To decide whether such contextual effects can be attributed to changes in the values compared or the judgment function (H or J), Mellers and Birnbaum (1982b) also asked judges to rate the "differences" between pairs of stimuli presented in the same two distributions. Two theories were considered. One theory assumes that ratings of single stimuli are like scale values. Therefore, differences in single ratings should predict ratings of "differences". The other theory assumes that differences in scale values estimated from range-frequency theory would predict the rank order of judged "differences" despite the context in which the "difference" ratings were obtained.

Mellers and Birnbaum also obtained judgments of "ratios" of stimuli in the same two contexts, yielding four matrices of data. The following model gave a good account of the data:

$$R_{ijk} = a_k \exp(s_j - s_i) + b_k$$
$$D_{ijk} = c_k(s_j - s_i) + d_k\,,$$

where R_{ijk} and D_{ijk} are "ratio" and "difference" judgments in context k; a_k, b_k, c_k, and d_k are constants fit to the data; and s_j and s_i are scale values, which are independent of task and independent of stimulus spacing. Thus, the same scale values could be used to reproduce the data in all four matrices. When scale values were estimated separately for aech context, they were found to be virtually identical.

The alternative theory that scale values depend on stimulus spacing was not required by the data. The rank order of differences in rating (from the single stimulus judgments) did not predict the rank order of "differences." That is

$$D_{ijk} \neq J[G_{jk} - G_{ik}]$$

for some monotonic J function.

Mellers and Birnbaum (1982b) found evidence that scale values inferred from additive and subtractive models of cross-modality combination and comparison tasks do appear to vary as a function of the context. It may be that cross-modality comparisons require judgment prior to combination or comparison, whereas within-modality comparison does not require a preliminary relative judgment.

Concluding comments

The scale convergence principle led to confidence that the constant-weight averaging model should be rejected as a representation of psychophysical "averaging" and as a theory of impression formation. Thus, it provided an argument against rescaling the data to fit a model that did not fit raw ratings.

The principle also led to rejection of a model that gave a good fit to raw data. The ratio model gives a reasonable fit to "ratio" judgments when certain experimental procedures are employed. However, the ratio theory fails to give a coherent account of both "ratio" and "differences" judgments, contextual effects, the four-stimulus tasks, and "inverse/reverse" judgments. The subtractive model does give a coherent account of the results.

The social sciences have long envied the coherence of the physical sciences. The dream of Fechner that psychology would develop a coherent system of laws interlocked by scales has not yet been achieved. My suggestion is that psychologists take their measurements seriously enough to assume that they will be reflected in several phenomena. By building scale convergence into our investigations we can find new knowledge that would not be forthcoming in separate experimental studies of single phenomena.

Acknowledgment

Thanks are due to Barbara A. Mellers for a critical reading of the manuscript.

References

ANDERSON, N. H.: Cross-task validation of functional measurement. Perception & Psychophysics, 1972, 12, 389—395

ANDERSON, N. H.: Information integration theory: A brief survey. In: D. H. KRANTZ, R. C. ATKINSON, R. D. LUCE, & P. SUPPES (Eds.), Contemporary developments in mathematical psychology (Vol. 2). San Francisco: Freeman, 1974

ANDERSON, N. H.: Algebraic rules and psychological measurement. American Scientist, 1979, 67, 555 bis 563

BIRNBAUM, M. H.: Morality judgments: Tests of an averaging model. Journal of Experimental Psycology, 1972, 93, 35—42

BIRNBAUM, M. H.: The devil rides again: Correlation as an index of fit. Psychological Bulletin, 1973, 79, 239—242

BIRNBAUM, M. H.: The nonadditivity of personality impressions. Journal of Experimental Psychology, 1974, 102, 543—561, a

BIRNBAUM, M. H.: Reply to the devil's advocates: Don't confound model testing and measurement. Psychological Bulletin, 1974, 81, 854—859, b

BIRNBAUM, M. H.: Using contextual effects to derive psychophysical scales. Perception & Psychophysics, 1974, 15, 89—96, c

BIRNBAUM, M. H.: Expectancy and judgment. In: F. RESTLE, R. SHIFFRIN, N. J. CASTELLAN, H. LINDMAN, & D. Pisoni (Eds.), Cognitive theory (Vol. 3). Hillsdale, N.J.: Lawrence Erlbaum Associates, 1978

BIRNBAUM, M. H.: Reply to Eisler: On the subtractive theory of stimulus comparison. Perception & Psychophysics, 1979, 25, 150—156

BIRNBAUM, M. H.: A comparison of two theories of "ratio" and "difference" judgments. Journal of Experimental Psychology: General, 1980, 109, 304—319

BIRNBAUM, M. H.: Controversies in psychological measurement. In B. WEGENER (Ed.), Social attitudes and psychophysics. Hillsdale, N.J.: Lawrence Erlbaum, 1982.

BIRNBAUM, M. H., & ELMASIAN, R.: Loudness ratios and differences involve the same psychophysical operation. Perception & Psychophysics, 1977, 22, 383—391

BIRNBAUM, M. H., & MELLERS, B. A.: Measurement and the mental map. Perception & Psychophysics, 1978, 23, 403—408

BIRNBAUM, M. H., PARDUCCI, A., & GIFFORD, R. K.: Contextual effects in information integration. Journal of Experimental Psychology, 1971, 88, 158—170

BIRNBAUM, M. H., & VEIT, C. T.: Scale convergence as a criterion for rescaling: Information integration with difference, ratio, and averaging tasks. Perception & Psychophysics, 1974, 15, 7—15, a

BIRNBAUM, M. H., & VEIT, C. T.: Scale-free tests of an additive model for the size-weight illusion. Perception & Psychophysics, 1974, 16, 276—282, b

EISLER, H.: On the ability to estimate differences: A note on BIRNBAUM's subtractive model. Perception & Psychophysics, 1978, 24, 185—189

ELMASIAN, R., & BIRNBAUM, M. H.: A harmonious note on pitch. Unpublished manuscript, 1979

GOGEL, W. C.: The measurement of perceived size and distance. Contribution to Sensory Physiology, 1968, 3, 125—148 (New York: Academic Press)

GREGORY, R. L.: Eye and Brain. New York: McGraw-Hill, 1978^3

HAGERTY, M., & BIRNBAUM, M. H.: Nonmetric test of ratio vs. substractive theories of stimulus comparison. Perception & Psychophysics, 1978, 24, 121—129

KAUFMAN, L.: Sight and mind, New York: Oxford University Press, 1974

KRANTZ, D. H.: Magnitude estimations and cross-modality matching. Journal of Mathematical Psychology, 1972, 9, 168—199.

KRANTZ, D. H., LUCE, R. D., SUPPES, P., & TVERSKY, A.: Foundations of measurement. New York: Academic Press, 1971

MELLERS, B. A., & BIRNBAUM, M. H.: Contextual effects in social judgment. Journal of Experimental Social Psychology, 1982, a (in press)

MELLERS, B. A., & BIRNBAUM, M. H.: Loci of contextual effects in judgment. Journal of Experimental Psychology: Human Perception and Performance, 1982, 8, b

NIHM, S. D.: Polynomial law of sensation. American Psychologist, 1976, 31, 808—809

PARDUCCI, A.: Range-frequency compromise in judgment. Psychological Review, 1965, 72, 407—418

PARDUCCI, A.: The relativism of absolute judgment. Scientific American, 1968, 219, 84—90

PARDUCCI, A.: Contextual effects: A range-frequency analysis. In: E. C. CARTERETTE & M. P. FRIEDMAN (Eds.), Handbook of perception. Vol. 2. New York: Academic Press, 1974

PARDUCCI, A., & PERRETT, L.: Category rating scales: Effects of relative spacing and frequency of stimulus values. Journal of Experimental Psychology, 1971, 89, 427—452

PARKER, S., SCHNEIDER, R., & KANOW, G.: Ratio scale measurement of the perceived lengths of lines. Journal of Experimental Psychology: Human Perception and Performance, 1975, 104, 195—204

POULTON, E. C.: Unwanted range effects from using within-subject experimental design. Psychological Bulletin, 1973, 80, 113—121

POULTON, E. C.: Models for biases in judging sensory magnitude. Psychological Bulletin, 1979, 86, 777 to 803

ROCK, I.: An introduction to perception. New York: MacMillan, 1975

ROSE, B. J., & BIRNBAUM, M. H.: Judgments of differences and ratios of numerals. Perception & Psychophysics, 1975, 18, 194—200

RULE, S. J., & CURTIS, D. W.: Conjoint scaling of subjective number and weight. Journal of Experimental Psychology, 1973, 97, 305—309

SARRIS, V., & HEINEKEN, E.: An experimental test of two mathematical models applied to the size-weight illusion. Journal of Experimental Psychology: Perception and Performance, 1976, 2, 295—298

SCHNEIDER, B., PARKER, S., KANOW, G., & FARRELL, G.: The perceptual basis of loudness ratio judgments. Perception & Psychophysics, 1976, 19, 309—320

SJÖBERG, L.: Sensation scales in the size-weight illusion. Scandinavian Journal of Psychology, 1969, 10, 109—112

STEVENS, S. S., & GALANTER, E. H.: Ratio scales and category scales for a dozen perceptual continua. Journal of Experimental Psychology, 1957, 54, 337—411

TORGERSON, W. S.: Quantitative judgment scales. In: H. GULLIKSEN & S. MESSICK (Eds.), Psychological scaling: Theory and applications. New York: Wiley, 1960

TORGERSON, W. S.: Distances and ratios in psychological scaling. Acta Psychologica, 1961, 19, 201—205

VEIT, C. T.: Ratio and subtractive processes in psychophysical judgment. Journal of Experimental Psychology: General, 1978, 107, 81—107

WEISS, D. J.: Averaging: An empirical validity criterion for magnitude estimation. Perception & Psychophysics, 1972, 12, 385—388