Test of a response bias model of bisection[1]

ROBERT F. FAGOT[2] AND MANARD R. STEWART
UNIVERSITY OF OREGON

An algebraic model of bisection—a special case of Pfanzagl's general measurement system—was tested for brightness. Nonparametric scalability, a condition derived from reflexivity, commutativity, and bisymmetry, was disconfirmed, leading to a rejection of the commutativity axiom, and necessitating the incorporation of a response bias parameter. The systematic bias in the data was substantially reduced by the introduction of the response bias parameter—interpreted as "position" bias, not hysteresis. The data are generally supportive of Pfanzagl's bisection system, although the failure of commutativity requires the incorporation of a response bias parameter.

One disadvantage with scaling methods that rely upon confusion among stimuli, especially in psychophysics, is that scales are generated only for narrow ranges. Piecing together such "local" scales to obtain a scale with a wider range then becomes a special problem. The theoretical structure of such scales has, however, received considerable attention (see, e.g., Luce & Galanter, 1963).

The "direct" ratio scaling methods—ratio production (fractionation and multiplication), ratio estimation, and magnitude estimation—do not depend upon confusion among stimuli and can produce a scale over a wide range without the problem of piecing together local scales. One defect with these methods is that their theoretical structure remains largely unanalyzed, unlike the major "confusion" scales. Some steps have been taken to make explicit and test some necessary conditions for a model of ratio or magnitude estimation scaling (Mashour, 1961; Sjöberg, 1965; Svenson & Åkesson, 1966, 1967; Fagot & Stewart, 1969a, b), but a set of axioms for any of these frequently used methods is conspicuously absent from the literature.

The method of *bisection* is similar to the above ratio methods relative to the lack of reliance on confusion among stimuli, but, unlike the ratio methods, bisection has received little attention in recent years. This is all the more notable in view of the fact that Pfanzagl (1959, 1968) has formulated a general measurement model, one specialization of which provides an algebraic (deterministic) model for bisection. Hence, bisection is in the curious position of being the least studied of those methods that do not rely on confusion among stimuli, yet the only one of these methods for which a formal theoretical structure has been provided. The purpose of this paper is to test an important consequence of Pfanzagl's axioms for bisection.

A bisection specialization of Pfanzagl's general measurement system is presented below in a relatively informal way, omitting some technical points.

Pfanzagl bisection system. Let $\mathscr{S}$ denote the set of stimuli, weakly ordered by $\succcurlyeq$ on the physical scale.

Axiom 1 (Existence) For all a, b, in $\mathscr{S}$ there exists a unique element aob in $\mathscr{S}$, which is interpreted as the *bisection point* of a and b.

Axiom 2 (Reflexivity) For all a in $\mathscr{S}$, aoa = a.

Axiom 3 (Bisymmetry) For all a, b, c, d in $\mathscr{S}$, (aob) o (cod) = (aoc) o (bod).

Axiom 4 (Monotonicity) If $a \precsim b$, then for all c in $\mathscr{S}$, $aoc \precsim boc$.

Axiom 5 (Continuity) aob is a continuous function of both a and b.

Axiom 1 introduces the binary operation "o," which assigns to each pair a, b, in $\mathscr{S}$ an element aob (bisection point) in $\mathscr{S}$. This points up a major way in which bisection differs from the numerical response methods (ratio and magnitude estimation): In the latter, S produces a *number*, whereas in bisection (as well as in ratio production), S produces a stimulus, i.e., performs an operation on presented stimuli (a, b) to produce a third stimulus (aob).

Axiom 2 is frequently taken for granted without testing in many applications, but could fail due to the presence of response biases. Axiom 5 is a technical axiom that has no testable consequences. Axiom 3 contains the principal power of the system.

Pfanzagl (1968) showed (Representation Theorem) that if Axioms 1-5 hold, then there exists a real-valued function Ψ on $\mathscr{S}$ that is unique up to a linear transformation (interval scale), and a real number δ ($o < \delta < 1$) such that

$$\Psi(aob) = \delta\,\Psi(a) + (1 - \delta)\,\Psi(b). \quad (1)$$

Another condition of considerable importance, formulated here as an axiom, is commutativity:

Axiom 6 (Commutativity) For all a, b, in $\mathscr{S}$, aob = boa.

It is easy to see from Axiom 6 and Eq. 1 that if, in addition to Axioms 1-5, commutativity holds, then $\delta = 1/2$.

The parameter δ is a *response bias* parameter and could be used to explain such effects as hysteresis or other effects produced by experimental conditions that result in different orderings of the stimuli.

An important condition, which we name *Non-Parametric Scalability*, is the following:

Non-Parametric Scalability (NPS). For all a, b, in $\mathscr{S}$, [a o (aob)] o [(aob) o b] = aob.

This condition follows directly from bisymmetry, reflexivity, and commutativity. If the condition holds, then $\delta = 1/2$, and an interval scale can be constructed without estimating the parameter δ.

The purpose of the present study was to test NPS for brightness. This condition was selected rather than bisymmetry as a first step partly because confirmation of NPS provides considerably more information about the bisection system, inasmuch as this theorem depends on commutativity and reflexivity as well as bisymmetry. Confirmation of NPS not only gives indirect evidence of the validity of bisymmetry—the key axiom in the system—but provides a method of *constructing* a scale without using the psychophysical function. If NPS is rejected, δ can be estimated to determine if the data can be accounted for by a response bias parameter. However, since the representation theorem (Eq. 1) is stated in terms of nonobservable Ψ magnitudes, rejection of NPS necessitates the use of the psychophysical function in order to construct a scale.

There appear to be no published studies testing the bisymmetry axiom directly, but

two studies of loudness have reported tests of NPS. Gage (1934), using low loudness levels and ascending order only, found that the final bisection—[ao(aob)] o [(aob)ob]—was consistently louder than the initial bisection—aob. Newman, Volkman, and Stevens (1937), using higher loudness levels and both ascending and descending orders, found a much smaller discrepancy in the same direction. Neither of these studies refutes bisymmetry, since commutativity or reflexivity could be at fault.

There are three problems with these reported tests of NPS.

(1) *The averaging problem.* Bisection points were determined by averaging several replications. Thus, NPS is interpreted to hold for such averages, and no theory was presented that dictated what average to use. To illustrate (denoting here and henceforth by Φ_i both the physical magnitude and the name of the stimulus), let Φ_1 and Φ_5 be two stimuli presented to S with instructions to bisect. Then let Φ_1 o$_i$ $\Phi_5 = \Phi_{3_i}$ be the i^{th} observation of the bisection point ($i = 1, \ldots . N$). Previous studies have averaged the N observations (using the geometric or arithmetic mean) to obtain Φ_3, and similarly obtained Φ_2 as the bisection point for Φ_1 and Φ_3, and Φ_4 as the bisection point for Φ_3 and Φ_5. The final step was to obtain $\Phi_{3'}$, as the bisection point for Φ_2 and Φ_4, in which case NPS is equivalent to $\Phi_3 = \Phi_{3'}$.

The procedure followed in our experiment was to base bisection points on single observations. Thus, the initial pair (Φ_1, Φ_5) is presented and S's single response Φ_3 is used to generate the second pair (Φ_1, Φ_3) and Φ_1 and Φ_3 are in turn bisected to produce Φ_2, etc. In this way, bisection points are not dependent on the particular measure of central tendency arbitrarily selected.

(2) *Single comparisons.* The studies by Gage (1934) and Newman, Volkman, and Stevens (1937) used a *single pair* of bisection points to test NPS. Considerably more information, possibly concerning the nature of systematic biases, can be obtained by using several pairs of bisection points covering a wide range of stimulus intensity, and this procedure was followed in the present experiment.

(3) Finally, the above studies did not attempt to account for the data by introduction of a response bias parameter, as will be done in this study.

METHOD

Subjects

The Ss were four undergraduate psychology majors, one male and three female. The male S had participated in a fractionation experiment prior to the present experiment.

Apparatus

The apparatus, described in detail by Eskildsen (1963), provided three luminous white circular targets, 13 mm in diam, placed horizontally and separated by 90 mm center to center. The middle target was continuously variable and under the control of the S, and provided a rate of change matching a .3 exponent power function.

Procedure

Each of nine basic pairs of stimuli were used to generate an elementary bisection set (EBS) as follows: Each block of four trials began with the presentation of one of the basic pairs, which we shall denote by (Φ_1, Φ_5). The S then bisected this interval, and the response was designated Φ_3. On the following two trials, S was presented the pairs (Φ_1, Φ_3) and (Φ_3, Φ_5), producing responses designated Φ_2 and Φ_4, respectively. The values Φ_2 and Φ_4 were then used to obtain the final observation, $\Phi_{3'}$ of the EBS. The basic pairs used were (8, 600), (3, 600), (1, 600), (.3, 600), (.1, 600), (.1, 200), (.1, 60), (.1, 20), (.1, 7.5) ft-L, each pair generating an EBS. Each EBS was replicated 10 times.

Prior to the start of the experiment, the following set of instructions was read to the Ss.

"You will be presented with three lights. The light on the left will be somewhat dimmer than the light on the right. The light in the middle will be of a brightness somewhere in between the right and left hand lights. Your task will be to adjust the brightness of the middle light such that its brightness is halfway in distance between the bright and dim lights. In other words, the interval of brightness between the dim and middle light should be equal to the interval of brightness between the middle and bright lights. You will repeat this judgment many times each session and the three lights will vary from trial to trial.

"There are certain things you should be cautioned against doing. Do not make a judgment on the basis of the size of the lights (they tend to appear larger or smaller to some people as the intensity changes), nor on the distance of the lights (they tend to appear further away or closer to some people as the intensity changes), nor on the basis of time. What I mean by that is, it will take the light a certain amount of time to go from the starting point to the point where you want it to be. Since we will start the light at a different place for each judgment, you will just increase your error if you take into consideration how long it is taking. If by chance the light is started on a point which you think is just right you *must* move the light away from that point even if you move it right back to the same place. You will not be limited in the amount of time you take to make an adjustment or the number of times you increase or decrease the intensity of the light.

"Are there any questions? Now, let's run through the procedure a few times."

Each S participated in 10 sessions, each lasting about 1 h, and before each session S was dark-adapted for a period of 15 min. All nine basic pairs were presented in random order in each session so that a complete replication of the nine EBSs was obtained in 36 trials. *The brighter stimulus was always on the right as viewed by S.* The order of presentation varied over days but was constant over Ss. On half the days, the bisection points were generated in the order $\Phi_3, \Phi_2, \Phi_4, \Phi_{3'}$ and on half the days in the order $\Phi_3, \Phi_4, \Phi_2, \Phi_{3'}$. Sessions using the two different orderings were intermixed randomly.

A basic set of 25 logarithmically spaced starting points ranging from .1 to 600 ft-L was used. The starting points for each of the nine basic pairs were taken from blocks of nine consecutive luminances from the basic set. The blocks were chosen so that the middle element of the block was close to where E estimated the average S would judge the midpoint of the interval to be. For a given basic pair, the starting points were chosen randomly without replacement until Day 10 when the starting points of Day 1 were re-used. A similar procedure, with two modifications was used to generate starting points for the remaining three trials of each EBS. Since the intervals to be bisected were smaller than for the first trial of each EBS, blocks of only seven consecutive starting points were used, with Days 8, 9, and 10 repeating the starting points of Days 1, 2, and 3, respectively. Secondly, since E did not know in advance one or both terminal stimuli for these trials, occasionally a starting point would turn out to be outside the interval presented to S. In these cases, the upper and lower bounds of the intervals were used as starting points, depending on whether the original starting point had been too high or too low.

RESULTS AND DISCUSSION

Two basic concepts in our analysis are *construction set* and *test set.* By construction set is meant the minimum set of bisection points (BP) sufficient to construct an interval scale. The number of BPs in the construction set is $n - 2$, where n is the number of stimuli scaled (since there are two arbitrary assignments, unit and zero point). For example, in our experiment there are five stimuli $(\Phi_1, \Phi_2, \Phi_3, \Phi_4, \Phi_5)$ which include three bisection

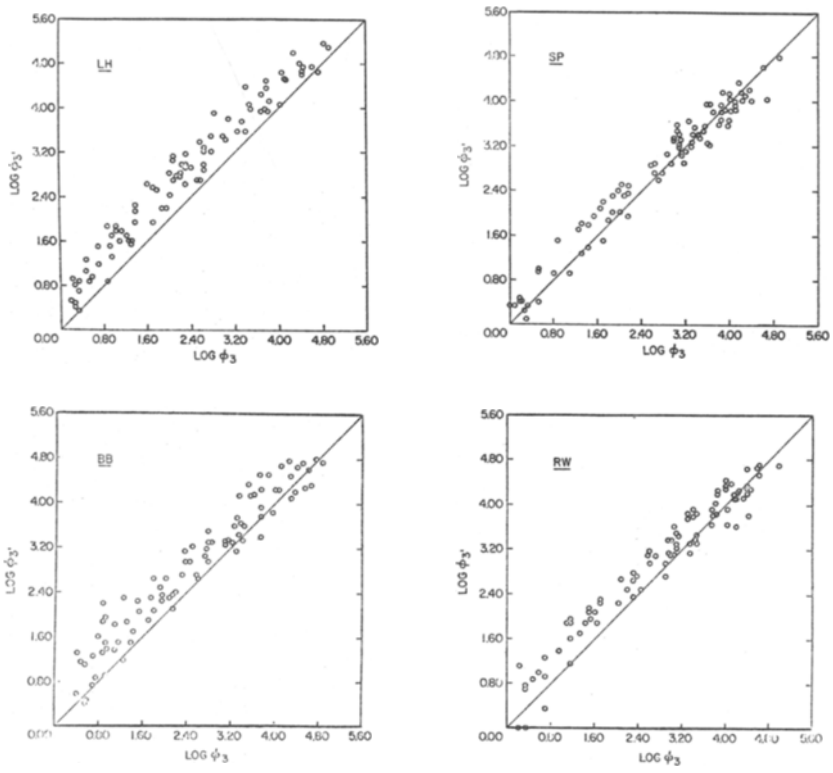

Fig. 1. Test of NPS: Log Φ_3' plotted as a function of Log Φ_3.

points (Φ_2, Φ_3, Φ_4) in the construction set.

The *test set* consists of those observations implied by the construction set via the axioms. Thus, relative to NPS, the *test set* consists of one testable consequence, $\Phi_3 = \Phi_3'$, for each EBS.

In many analyses following a "predicted vs observed" paradigm, *all* of the data are used to estimate parameters and "predict" values, *including the so-called "observed" values.* Partitioning the observations into a construction set and a test set (a natural consequence of conditions such as NPS, which involve a "premise" based on a construction set and a "conclusion" based on a test set) provides a more powerful test, since only the construction set data are used to estimate parameters and an independent set of observations (the test set) are predicted using these parameter estimates.

Test of NPS. If $\delta = 1/2$ (no response bias) then a plot of Φ_3' as a function of Φ_3 should result in a straight line through the origin with unit slope. Construction of such a plot was done as follows.

Each of the nine EBS generated 10 replications of the bisection point Φ_3 and 10 replications of the bisection point Φ_3', resulting in a total of 90 data points. Figure 1 shows a plot of $\log \Phi_3'$ as a function of $\log \Phi_3$, separately for each of four Ss. The line is the theoretical line predicted by NPS.

Inspection of Fig. 1 shows a clear systematic trend: The test stimulus Φ_3' tends to be *over-estimated* relative to the construction stimulus Φ_3. This effect is very pronounced for three Ss (BB, RW, LH). In the case of LH, for example, all but one of the 90 points are on or above (mostly above) the theoretical line. The plot for SP, however, is somewhat deviant: The modal tendency to overestimate the test stimuli relative to the construction stimuli is moderately evident for the lower half of the scale but absent for the upper half. In any event, NPS must be rejected, and we shall explore the possibility of accounting for the data by incorporating δ into the model as a response bias parameter.

Response bias. Since commutativity is not one of the axioms of Pfanzagl's bisection system, the model permits response biases of the form aob $\neq$ boa, which implies $\delta \neq 1/2$ from Eq. 1. Luce and Galanter (1963, p. 161) have offered one possible interpretation of δ as a *hysteresis* parameter, in which case, for $a < b$, aob could denote, say, the bisection point using the "up-method" and boa the bisection point using the "down-method."

Although approximately half the starting points were below BPs (this was not under absolute control of E since the BP was not known prior to setting up starting points), the design did not provide a fair test of hysteresis since an S was allowed to "back up" if he went beyond his perception of the BP. This was done because the main interest was in getting a good test of NPS rather than in hysteresis.

In any event, the fact that the data points in Fig. 1 tend to lie mostly on one side of the theoretical line for LH, BB, and RW indicates that, at least for these three Ss, hysteresis could not account for the systematic trend, since both the "up-method" and "down-method" empirical lines lie above the theoretical line. An analysis was done for each method separately, and the two lines were, in fact, indistinguishable. It is clear that hysteresis cannot account for the bias in the data.

A subsequent experiment suggested *position bias* as a possible explanation: In an unpublished pair-comparison experiment, Fagot and Stewart found that (for most Ss) when a given pair of stimuli was presented a large number of times in a horizontal display with left-right position balanced, then a stimulus was judged the brighter of the pair more often when it appeared on the *left* than when it sppeared on the right—which we designate *"left-dominance."* It can be shown that left-dominance implies $\delta < 1/2$ and $\Phi_3 < \Phi_3'$ (Eq. 1 with aob denoting a on the left). Therefore, left-dominance provides at least a qualitative explanation for the systematic deviations observed in Fig. 1, except possibly for SP.

The next step in the analysis was to estimate the parameter δ and determine if such a position bias parameter could account quantitatively for the data in Fig. 1. In order to apply Eq. 1, we need to distinguish between aob and boa. Since the brighter stimulus was always on the right, aob denotes a was on the left, and boa was never observed.

Forms of the power law. Since Eq. 1 is expressed in terms of psychological magnitudes Ψ and observations in terms of Φ, it is necessary to introduce the *psychophysical function,* $\Psi = f(\Phi)$. We consider here three different functions f (forms of the power law) and hence three *bias models.* These three forms of the power law are (see Fagot, 1966):

The general power law:

$$\Psi = c(\Phi - \tau)^k + d; \qquad (2)$$

the *Φ-Law* (translation on the intensity axis)

$$\Psi = c(\Phi - t)^k, \qquad (3)$$

and the *Ψ-Law* (translation on the psychological axis)

$$\Psi = c(\Phi^k - t^k). \quad (4)$$

For both Eqs. 3 and 4, the parameter t is interpreted as a *threshold* parameter, i.e., $\Psi(t) = 0$. The Φ-Law is a special case of the general law in which $\tau = t$ and $d = 0$; and the Ψ-Law is a special case in which $\tau = 0$ (and $d = -ct^k$).

Bias models. The General Bias Model (Model G) is derived from Eqs. 1 and 2 and is given by Eq. 5:

$$(\Phi_{aob} - \tau)^k = \delta(\Phi_a - \tau)^k + (1 - \delta)(\Phi_b - \tau)^k, \quad (5)$$

a three-parameter (δ, τ, k) model.

The Φ-Law Bias Model (Model Φ). This is also a three-parameter model and is given by Eq. 5 with the substitution $\tau = t$. Thus, the Φ-Law model is distinguished from the general law only by the interpretation of τ as a threshold parameter.

The Ψ-Law Bias Model (Model Ψ). This model is obtained by substituting $\tau = 0$ in Eq. 5.

$$\Phi^k_{aob} = \delta\,\Phi^k_a + (1 - \delta)\,\Phi^k_b. \quad (6)$$

Thus Eq. 6 is a two-parameter (δ and k) model, one less parameter than for Models G and Φ. Furthermore, this model alone of the three implies that bisections [and interval judgments in general (Fagot & Stewart, 1969b)] are *independent of the threshold parameter t.*

Estimation of parameters. An "observation" is a physical value (luminance) obtained when S "produces" a setting on a dial that E records in ft-L, the actual value not observable, of course, by the S. An observation is denoted Φ_{ijk} where

$i = 2, 3, 4, 3'$ (bisection points)
$j = 1, 2, \cdots, 9$ (EBS)
$k = 1, 2, \cdots, 10$ (replications)

In addition, the pairs (Φ_{1j}, Φ_{5j}) are independent variables, each of the nine pairs generating an EBS with replications.

In the estimation procedure, $\log \Phi_{ijk}$ was treated as the dependent variable, and an iterative least squares procedure was applied to estimate parameters for Models G and Ψ, separately for each of the four Ss. Estimates of the parameters were based entirely on the construction set observations (Φ_{2jk}, Φ_{3jk}, Φ_{4jk}), a procedure that permits prediction to the test stimulus observations ($\Phi_{3'jk}$) as "independent" data. Hence, there were a total of 270 observations (for each S) on which to base parameter estimates. A summary of parameter estimates is given in Table 1.

Table 1
Parameter Summary, Individual Data

S		$\hat{\kappa}$	$\hat{\delta}$	$\hat{\tau}$
(BB)	Model Ψ	−0.0650	0.3745	–
	Model G	−0.0725	0.4101	−0.1156
(RW)	Model Ψ	0.0684	0.4236	
	Model G	0.0768	0.4744	−0.1813
(LH)	Model Ψ	−0.1356	0.3448	–
	Model G	−0.1512	0.3638	−0.0687
(SP)	Model Ψ	0.1193	0.4777	–
	Model G	0.1411	0.5433	−0.2264

Note that for two Ss (BB and LH) $k < 0$, i.e., the psychophysical function is more negatively accelerated than a log function (Fagot, 1966). This means that, for these two Ss, the general power law cannot be interpreted as the Φ-Law since exponents for the Φ-Law are restricted to the range $k > 0$. Note that for all four Ss $\hat{\tau} < 0$, ruling out the interpretation of τ as a threshold parameter (t) for all Ss. If we wish to retain the interpretation of t as a threshold parameter, we must reject the Φ-Law for these data. This result is consistent with some previous research on brightness (Fagot & Stewart, 1969a, b).

For three Ss (BB, RW, LH) $\hat{\delta} < 1/2$ for both models, consistent with the left-dominance reported in the pair-comparison experiment, and with the very large proportion of data points above the theoretical lines (Fig. 1). Since in the case of SP inspection of Fig. 1 suggests only slight bias, it is not disturbing that the two models estimate different directions for the bias. If we use the number of points above the theoretical lines (Fig. 1) as a rough measure of bias, then for Model Ψ the ranking of the four Ss on this measure is identical to the ranking given by the $\hat{\delta}$, whereas there is one reversal for Model G.

The parameter estimates in Table 1 were used to investigate the adequacy of the response bias parameter δ in eliminating the systematic deviations observed in Fig. 1. First, $\hat{k}$ and $\hat{\delta}$ from Model Ψ were applied to predict the $\log \Phi_{3'jk}$, and these values were compared to the observed $\log \Phi_{3'jk}$, where the observed values were *not* included in the estimation data.

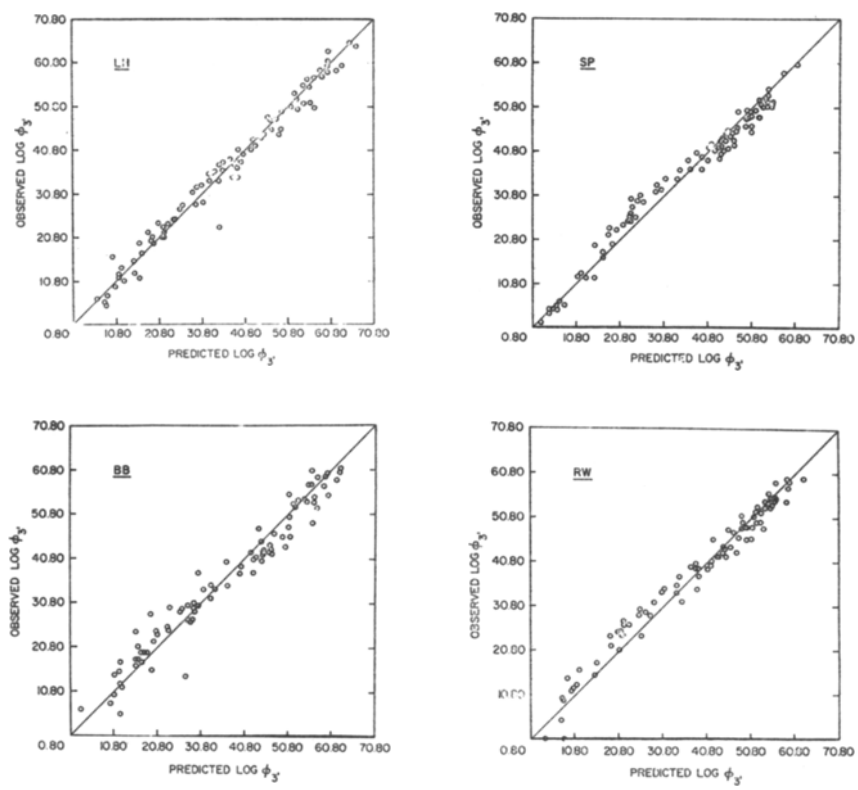

Fig. 2. Test of Model Ψ: Plot of observed Log $\Phi_{3'}$, as a function of Log $\Phi_{3'}$ predicted from construction set data using Model Ψ.

Figure 2 shows a plot of observed log $\Phi_3'_{jk}$ as a function of log $\Phi_3'_{jk}$ predicted from the construction set data using Model Ψ.

In general, the marked systematic deviations exhibited in Fig. 1 are eliminated, which implies that the systematic trend is largely due to response (position) bias. However, there still appear to be slight residual systematic deviations that are unaccounted for by the bias parameter δ. Inspection of Fig. 2 shows the same general pattern for all Ss: a tendency for predicted values to be too low relative to observed values at the low end of the scale and too high at the high end of the scale. The data for all Ss suggest a function with an inflection point, slightly concave down at the low end and slightly concave up at the high end. It is interesting to note that LH exhibited the greatest bias in Fig. 1 (and worse fit to the theoretical line), but after introduction of the bias parameter δ, LH appears to fit the theoretical line best (Fig. 2).

The rejection of NPS does not necessarily imply the rejection of bisymmetry, since NPS depends on commutativity and reflexivity, as well as bisymmetry. Failure of commutativity is implied by $\delta \neq 1/2$ provided Eq. 1 holds, but there is the possibility that reflexivity does not hold, in which case the representation theorem given by Eq. 1 is not correct. Dropping both reflexivity and commutativity as axioms, the numerical representation is given by Eq. 7:

$$\Psi(aob) = \alpha \Psi(a) + \beta \Psi(b) + \gamma. \quad (7)$$

Pfanzagl (1968) has shown that (1) $\alpha + \beta = 1$, $\gamma = 0$ if and only if o is reflexive; and (2) $\alpha = \beta$ if and only if o is commutative.

To test for possible violations of reflexivity, we set $\gamma = 0$ and estimated the parameters α and β from the construction set data, using a least squares iterative procedure with log Φ as the dependent variable. The sums $\hat{\alpha} + \hat{\beta}$ were 1.007 (BB), 1.003 (RW), .9867 (LH), and .9989 (SP). The requisite sums were judged to be sufficiently close to unity to rule out nonreflexivity, although no attempt to devise a statistical test seemed warranted.

Equation 1 would appear to give a satisfactory numerical representation, and therefore $\delta \neq 1/2$ implies a violation of commutativity but does not give evidence of a failure of bisymmetry.

The failure of commutativity does not require the rejection of Pfanzagl's bisection system, nor does it imply the nonexistence of a metrical scale, although the construction of a scale does require estimation of δ (Eq. 1). Figure 2 shows that the introduction of the response bias parameter removed most of the systematic bias exhibited in Fig. 1.

The unfortunate consequence of the necessary introduction of the psychophysical function resulted in a confounding between the function and bisymmetry as sources of error in producing the residual bias in Fig. 2. (Since Fig. 2 is based on $\hat{\delta}$, response bias due to the failure of commutativity is essentially "removed" and any residual bias must be due to failure of bisymmetry or such other factors as a poorly fitting psychophysical function.) Since Fig. 2 involves a prediction from the construction set to the test stimuli, one approach that was taken in an attempt to disentangle the two possible sources of error was to analyze the data from the construction set alone. Parameter estimates $\hat{\kappa}$ and $\hat{\delta}$ for Model Ψ (Table 1) derived from construction set observations were used to "reproduce" all 270 observations of the construction set (Φ_{2jk}, Φ_{3jk}, Φ_{4jk}). A plot of "log observations" as a function of "log reproductions" was constructed (not presented here) for each of the four Ss. These plots showed the same general trend exhibited in Fig. 2, except that the trend was not as pronounced. On this basis, we tentatively conclude that the systematic trends exhibited in Fig. 2 reflect more a general judgmental bias or a failure of the psychophysical functions rather than a special failure of bisymmetry.

The question arises as to whether or not the additional parameter (τ) in Model G provides substantial added predictive power over Model Ψ. A plot of observed log $\Phi_3'_{jk}$ as a function of log $\Phi_3'_{jk}$ predicted from Model G appeared upon inspection to be so similar to the corresponding plot for Model Ψ (Fig. 2) that the figure was not presented. If we compare the predictions for each model with the observed values (the observed and predicted values based on different sets of data), we find that the Model Ψ predictions were actually closer to the observed values in 51.7% of the cases, pooling over Ss, and as close or closer for three of the four Ss. On the other hand, if we take account of the magnitude of the difference between prediction and observed, there is some advantage to Model G: Following Fagot and Stewart (1969b), let d_{jk} (G) denote the absolute value of the difference between the $\Phi_3'_{jk}$ predicted from Model G and the corresponding observed value, and d_{jk} (Ψ) the corresponding difference based on Model Ψ. Assuming that the differences $D_{jk} = d_{jk}(G) - d_{jk}(\Psi)$ are normal, the hypothesis of no difference was tested separately for each of the four Ss.[3] Results showed that for two of the Ss (BB and RW), the differences were significant at the .05 level but not significant for the other two Ss. Hence, if the normality assumption can be supported, Model G (with an additional parameter) does somewhat better for half the Ss. It is perhaps remarkable that Model Ψ, with one less parameter, compares so favorably with Model G.

Group data. Since for both bias models the exponents for the four Ss average close to zero, the log law will give a good fit to the group data (Fagot, 1966). Table 2 presents parameter estimates for Model G, Model Ψ, and the log model for the group data. Estimates were obtained by pooling observations for all Ss. We see that, for both Model G and Ψ, the group exponent is slightly negative, i.e., the group power function is slightly more negatively accelerated than a log function. Note that τ is substantially negative, ruling out a threshold interpretation (and presumably the Φ-Law) for the group data. The fact that the exponents are so close to zero means that neither power law can be discriminated from the log law for the group data, although a choice in favor of the power law can be made for each individual S. Considering that the log model contains only one parameter (δ) compared to two and three for Models Ψ and G, respectively, it accounts for the data remarkably well.

Concluding discussion. The systematic bias (for brightness) exhibited in Fig. 1 is in the same direction ($\Phi_3' > \Phi_3$) as the bias observed for loudness in the studies by Gage (1934) and Newman, Volkman, and Stevens (1937). In the Gage (1934) study, no explanation was offered to account for the observed response bias. Newman, Volkman, and Stevens (1937) attempted to eliminate the possible systematic effect of hysteresis by averaging the ascending and descending series of observations, and succeeded in eliminating most of the bias in the former study but provided no means of evaluating the hysteresis effect. It would appear that the major contributing factor to the failure of NPS in all three studies was the violation of commutativity.

What evidence is there for the violation of commutativity in the Gage (1934) study? Our evidence here is indirect. Stevens (1957) reported that his Ss set the bisection point in the ascending order 5-8 dB louder than in the descending order.

Table 2
Parameter Summary, Group Data

	$\hat{\kappa}$	$\hat{\delta}$	$\hat{\tau}$
Model Ψ	−.0068	0.4021	–
Model G	−.0090	0.4431	−0.1360
Log Model	–	0.4069	–

In our terminology, for $b > a$, aob is the BP of the ascending order and boa the BP of the descending order; then aob > boa, a violation of commutativity. From Eq. 1 it is obvious that for $b > a$, aob > boa implies $\delta < 1/2$, and NPS fails in the direction $\Phi_3' > \Phi_3$. Since Gage (1934) used ascending order only then the hysteresis effect accounts for the failure of commutativity and NPS. The bias in the Newman, Volkman, and Stevens (1937) study, although in the same direction, was slight, presumably due to averaging ascending and descending orders and largely "removing" the hystersis effect.

The brightness data reported here did not permit a direct comparison of aob and boa, but the estimations of the bias parameter implied violation of commutativity ($\delta < 1/2$) and, in particular, left-dominance. One additional result of our study was to show that the bias existed generally at all levels. Although response bias appeared to be present in all three studies, different effects were responsible: hysteresis in the Gage (1934) experiment and position bias (left-dominance) in the present brightness experiment. Thus δ is a general response bias parameter, the particular interpretation depending on the experiment.

We conclude that these data (as well as Gage, 1934) show a violation of commutativity and NPS but do not give evidence against bisymmetry, the key axiom of the Pfanzagl system. Furthermore, since commutativity is not a necessary condition in the Pfanzagl system, its violation does not refute the *existence* of an interval scale. As in many measurement systems, not all the Pfanzagl axioms are directly testable. Given the failure of commutativity and NPS, one would usually be satisfied with a test of bisymmetry as sufficient to imply the existence of a metrical scale. However, since the representation theorem (Eq. 1) is expressed in terms of unobservable Ψ magnitudes and a parameter δ that must be estimated if commutativity fails, then rejection of commutativity would appear to require the use of the psychophysical function in order to *construct* a scale. Pfanzagl (1968) has presented a so-called "derived middling operation" that would permit the construction of a scale without assuming the existence of a psychophysical function, but it is not at all clear what set of experimental conditions would satisfy this complicated "operation."

In any event, these data are interpreted as generally supportive of the Pfanzagl bisection axiom system, and as pointing to the necessary introduction of response bias parameters. Follow-up studies should focus on direct tests of bisymmetry and commutativity, and on designs that permit the independent estimation of possible response biases.

REFERENCES

ESKILDSEN, P. An apparatus for studies of brightness which controls variation of intensity with time. American Journal of Psychology, 1963, 76, 321-323.

FAGOT, R. F. Alternative power laws for ratio scaling. Psychometrika, 1966, 31, 201-214.

FAGOT, R. F., & STEWART, M. R. Tests of product and additive scaling axioms. Perception & Psychophysics, 1969a, 5, 117-123.

FAGOT, R. F., & STEWART, M. R. Individual half-judgment brightness functions. Perception & Psychophysics, 1969b, 5, 165-170.

GAGE, F. H. The measurability of auditory sensations. Proceedings of the Royal Society, 1934, 23, 35-40.

LUCE, R. D., & GALANTER, E. Psychophysical scaling. In R. D. Luce, R. R. Bush, and E. Galanter (Eds.), *Handbook of mathematical psychology*. Vol. 1. New York: Wiley, 1963.

MASHOUR, M. On the validity of scales derived by ratio and magnitude estimation methods. Report of the Psychology Laboratory, University of Stockholm, 1961, No. 105.

NEWMAN, E., VOLKMANN, J., & STEVENS, S. S. On the method of bisection and its relation to a loudness scale. American Journal of Psychology, 1937, 49, 134-137.

PFANZAGL, J. A general theory of measurement: Applications to utility. Naval Research Logistics Quarterly, 1959, 6, 283-294.

PFANZAGL, J. In cooperation with V. Baumann and H. Huber, *Theory of measurement*. New York: Wiley, 1968.

SJÖBERG, L. On ratio estimation. Report of the Psychology Laboratory, University of Stockholm, 1965, No. 191.

STEVENS, S. S. On the psychophysical law. Psychological Review, 1957, 64, 153-181.

SVENSON, O., & ÅKESSON, C. A. Fractional and multiple estimates in ratio scaling. Report of the Psychology Laboratory, University of Stockholm, 1966, No. 202.

SVENSON, O., & ÅKESSON, C. A. A further note on fractional and multiple estimates in ratio scaling. Report of the Psychology Laboratory, University of Stockholm, 1967, No. 224.

NOTES

1. This research was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Air Force Office of Scientific Research under Contract No. F44620-67-C-0099. We are indebted to Robyn Dawes for helpful comments.

2. Address: Department of Psychology, University of Oregon, Eugene, Oregon 97403.

3. This test would not make much sense if all the data were used to estimate parameters, since then Model G must do at least as well as Model Ψ. However, since observations and predictions were based on different sets of data (construction set and test set), it was theoretically possible for Model Ψ to do better than Model G.

(Accepted for publication September 30, 1969.)