Journal of Experimental Psychology
1969, Vol. 82, No. 1, 52–63

SUBJECTIVE AVERAGING OF LENGTH WITH SERIAL PRESENTATION [1]

DAVID J. WEISS [2] AND NORMAN H. ANDERSON

University of California, San Diego

The *S*s estimated the average of several lengths presented serially one at a time. In Exp. I, the judgment was made only at the end of the sequence. In Exp. II, *S* estimated a cumulative average as each new length was presented. The main phases of these experiments used sequences of six lengths. For the most part, each *S*'s data could be described by a subjective averaging model as tested in single-*S* analyses. There was a general recency effect, the later lengths in the sequence having greater influence. Recency was fairly uniform across *S*s with the end responding procedure of Exp. I, but large individual differences in the serial position curves appeared with the continuous responding procedure of Exp. II. In Exp. III, two hypotheses about the cause of recency were tested, but received little support. Functional measurement technique showed that subjective length differed from objective length, apparently by a constant error for each *S*. It was noted that the present methods could be applied to psychophysical scaling of other stimulus dimensions.

Suppose that you are shown several lines, one at a time, and asked to estimate their average length. Can your response be described as a mathematical average? If so, is it an arithmetic mean, geometric mean, weighted midpoint, or some other measure of central tendency? The present experiments were designed to consider this question. They test a subjective averaging model for length.

Averaging model.—The basic assumption of the model is that the response (R) at Serial Position N is simply a weighted sum:

$$R_N = \sum_{1}^{N} w_k s_k. \qquad [1]$$

Here w_k is the weight of the kth stimulus, and s_k is its scale value. For the present application, it is assumed that the w_k sum to unity across serial positions, so that Equation 1 is an arithmetic averaging model.

Two further restrictions are also made. First, the weight of any stimulus is assumed to depend only on serial position. Second, the scale value or subjective length of any stimulus is assumed to be constant, independent of the other stimuli in the sequence. Different *S*s, of course, will have different weights and scale values, but this is taken into account in applying the model.

Goodness of fit.—If mathematical correctness were taken as the standard, three sources of inaccuracy might arise. First, subjective length might differ from objective length. Second, the serial positions might receive unequal weighting, e.g., if weight were related to the recency of the memory trace. In either case, the subjective averages will generally be unequal.

Both the aforementioned sources of inaccuracy are allowed by the model. They correspond to values of w_k and s_k that differ from the arithmetically correct values. Neither of these inaccuracies, therefore, poses any immediate problem.

There is a third source of inaccuracy that would reflect adversely on the model as applied here. If the stimuli interact with one another, or with serial position, then the model will not hold in general. Such interactions might arise from psychophysical assimilation or contrast, or from other sorts of context effects.

It is important, therefore, to test the model, even while allowing for subjective values of the weight and value parameters that are specific to each *S*. Fortunately, testing is straightforward. If the stimuli are constructed from a factorial design,

[1] This work was supported by National Science Foundation Grants GB-3913 and GB-6666.

[2] Requests for reprints should be sent to David J. Weiss, Department of Psychology, University of California, San Diego, P. O. Box 109, La Jolla, California 92037.

analysis of variance provides a direct and powerful test of goodness of fit (Anderson, 1962a, 1964a).

Functional measurement.—If the model is validated, it may be used to scale both subjective length and the weights at each serial position. This application of functional measurement is illustrated for the present experimental design.

In the main body of the present data, there were just two lengths, 16 and 24 cm., at each serial position. The design yields estimates of their observed differential effect (D_k) at Serial Position k. The theoretical expression for D_k is:

$$D_k = w_k(s_{24} - s_{16}). \qquad [2]$$

Since the difference in scale values ($s_{24} - s_{16}$) is the same at each serial position, D_k is proportional to w_k. Hence, if each D_k is divided by their sum, these quotients add to unity and provide estimates of the w_k. These estimates define the serial position curve.

Subjective length may also be scaled using the model. At Serial Position N, the model implies that

$$\sum_1^N D_k = \sum_1^N w_k(s_{24} - s_{16}) = s_{24} - s_{16},$$

since the w_k sum to unity. Also, the grand mean response equals $(s_{24} + s_{16})/2$. The data thus provide estimates of both sum and difference of the scale values and hence of each scale value separately.

METHOD

The same basic length-averaging task was employed in all three experiments reported here. In Exp. I, S saw three or six lengths in sequence and estimated their average length at the end of the sequence after the last length had been presented. In Exp. II, the procedure was quite similar except that S estimated a running average, giving a judgment after each successive length was presented. In Exp. III, a procedure of simultaneous-serial presentation was introduced to test two hypotheses about the source of the recency effect observed in Exp. I and II.

Experiment I

Apparatus and procedure.—The line stimuli in this experiment were displayed horizontally on a 43×28 cm. screen using a rear projector. The projected lines were black, .5 cm. wide, and ranged from 14 to 26 cm. in length. The screen was 168 cm. in front of S, and the lines were displayed 28 cm. above the response panel. Each line was projected for 4 sec., with an interstimulus interval of 2 sec. during which the screen was blank. Normal room illumination was used.

Responses were made by the method of reproduction after the last line stimulus disappeared. By throwing a two-way, spring-loaded switch, S controlled a motor that adjusted the length of a looped horizontal white tape displayed in a panel 122 cm. in front of him.

On E's side of the response apparatus, a marker attached to the response tape passed along a meter stick to monitor S's response. Responses were read to the nearest millimeter by E, who sat concealed behind a screen by the side of the response apparatus. After each sequence, E reset the white tape to zero length.

Design.—In the main part of the experiment, there were six lengths in each sequence. At each of these six serial positions, the length could be long (24 cm.) or short (16 cm.). There are then 64 possible sequences, and they form a 2^6 factorial design.

Each regular daily session included a repetition of the instructions, 4 practice sequences, a half replication (32 sequences) of the basic design, plus 6 interspersed filler sequences. The practice and filler sequences included lengths of 14, 18, 22, and 26 cm., as well as 16 and 24 cm., and some of these sequences also served as end anchors. Although end effects did not seem likely with a continuous, unmarked scale, these anchor sequences were an extra precaution against distortion of the effective response scale.

After an initial practice day, each S was run in eight daily sessions, yielding four replications of the main design. This was followed by two sessions in which sequences of three lengths were judged. Analogous to the main design, this yielded a 2^3 design; each S was run through it four times each session.

Subjects and instructions.—The Ss were eight students who received $1.85 for each of the 11 sessions. They were told to estimate the average length of the six (or three) lengths in each sequence, and some preliminary sequences were used to ensure that they understood the task.

Experiment II

General procedure was similar to that of Exp. I, except that S responded after each length of the sequence, giving the running average of the lengths he had seen so far. In addition, the stimulus display and response panel were somewhat different. Experiment II was run before Exp. I, but is presented second for expositional simplicity.

Apparatus and procedure.—The line stimuli in this experiment were sticks, 1 cm. in diameter, painted black, and of the same lengths as used in Exp. I. They were presented singly between guides

on the table 60 cm. in front of *S* and centered about 25 cm. left of the zero region on the response apparatus.

Responses were made with the stick present. Responding was self-paced and generally required 4–5 sec. per length. A complete sequence of six lengths required about 1 min.

The response apparatus consisted of a meter stick surmounted by a brass rod that held a sliding indicator about 5 cm. above the table. On the side facing *S*, the meter stick was covered with black felt, except for the rightmost 20 cm., which was white. The black-white boundary served as the zero reference from which *S* estimated the average length. Within each sequence, the response indicator remained at the last response until *S* made his new estimate. At the end of each sequence, *S* returned the indicator to the zero position.

The *E* sat on the other side of the table from *S*, presented each stick in turn, and recorded the response by reading the meter stick to the nearest millimeter. Except for the one present stimulus, the sticks were concealed from *S* behind a small blind.

Design: Phase 1.—In the main part of the experiment, there were six lengths in each sequence. At each of these serial positions, the lengths could be long (24 cm.) or short (16 cm.). As in Exp. I, there are thus 64 possible sequences and they form a 2^6 factorial design.

Each regular daily session included a repetition of the instructions, four practice sequences, a half replication (32 sequences) of the 2^6 design, plus six interspersed filler sequences. Practice and fillers were the same as in Exp. I. Each *S* served one practice and eight regular sessions, yielding four replications of the basic design.

Design: Phase 2.—In this phase, 10 lengths were used in each sequence in order to obtain serial curves for longer sequences. Procedure was the same as in Phase 1: a 2^6 design was again employed. To compress 10 serial positions into six factors, each of the first four pairs of serial positions was treated as a factor, each with the same two levels. One level consisted of 14 cm. followed by 18 cm.; the other level consisted of 22 cm. followed by 26 cm. The levels of the last two factors, Serial Positions 9 and 10, were 16 cm. and 24 cm., as before. Each *S* went through the complete design twice at the rate of one-quarter replication per day.

Subjects and instructions.—Undergraduate *S*s received $1.85 per session, except for *S*4, a department secretary. An initial squad of four *S*s was run through both phases of the experiment. The data supported the model, but there were large individual differences in the serial position curves. To get more information on the distribution of individual differences, *S*5–*S*10 were then run through the Phase 1 design.

Running average instructions were employed. The *S*s were told to simply estimate the length of the first length, then the average of the first two lengths, etc. This pointer position was then to be changed to incorporate each new length in the cumulative average. A movement of the pointer was required at each stimulus presentation; however, *S* was told that he could move it very slightly, or away and back.

Experiment III

A simultaneous-serial presentation procedure was employed to test two hypotheses about the recency effect. The main procedural change was to present several lengths simultaneously in the first serial position. Thereafter only one length at a time was presented, and *S* responded after each presentation with a running average, as in Exp. II.

Apparatus and procedure.—The response apparatus was the same as in Exp. I. The stimuli were sticks, as in Exp. II, ranging 10–28 cm. in 2-cm. steps. Stimulus presentation was the same as in Exp. II, except that from 1 to 5 lengths might be shown simultaneously on the first serial presentation of a sequence. When several lengths were shown together, they rested on separate guides parallel to the response panel, with the sticks placed haphazardly among themselves. Responding was self-paced and generally required about 10 sec. per response for the naive *S*s of this experiment.

Design.—There were seven basic conditions, as shown in Table 3. The first number in the condition designation is the number of lengths presented together in the first serial position. Thus, in Cond. 4-1-1, four lengths were shown together initially; then the next two serial positions had one length each, for a total of six lengths.

Six different pairs of lengths, each pair differing by 8 cm., were used for the first five conditions listed in Table 3. Length pairs were balanced over serial positions with a 6×6 Latin square, as in Anderson (1964b). As a consequence, the lengths in the initial presentation were all different for the first five conditions.

For the last two conditions listed in Table 3, the four or five lengths presented in the initial position were all equal. It was thought that this would produce greater confidence in the estimated average than when the first four or five lengths were all different, thereby reducing the recency effect.

Each *S* judged a total of 64 sequences, from 4 to 16 in each of the seven conditions. As in the previous experiments, these were constructed from factorial designs in order to allow estimation of the weights associated with each serial position. For the longer sequences, this required use of fractional replication (Cochran & Cox, 1957).

Subjects and instructions.—The 24 undergraduate *S*s were each paid $1.00 plus class credit for the 1.5-hr. session. Instructions were similar to those used in the first two experiments.

RESULTS

The main tests of the model, in Exp. I and II, were made by applying analysis of variance separately to the data of each single *S*. The theoretical basis for this analysis has been noted previously, and

here it need be kept in mind only that the model implies that the interaction terms are zero. The within-replicates variability was used as the error term in these single-S analyses. Certain group analyses were also made for various purposes. In these, of course, the error term is the appropriate S interaction.

There is a difficulty that is partly an embarrassment in reporting these analyses because of the multitude of interactions. With six serial positions, a complete factorial yields 57 interaction terms, and these are difficult to present in simple form. Moreover, with so many tests, just about three interactions might be expected to be significant at the .05 level by chance alone. It is desirable to avoid testing so many interactions if that can reasonably be done (Anderson, 1968). A few remarks on this question, written in light of subsequent results, may conveniently be included at this point.

There seem to be three main ways to reduce the number of interactions tested. One method is to employ confounding procedures (Cochran & Cox, 1957). For instance, a 2^6 design in $\frac{1}{8}$ replicate leaves just 1 df for a test of pooled interactions. It also reduces the number of stimulus sequences from 64 to 8, a feature that is useful in many experimental applications (e.g., Anderson, 1964b).

A second method is to apply partial analysis (Anderson, 1968). With partial analysis, high-order interactions are not tested when there is reasonable ground to expect that they will be negligible. Related to this is the third method, which focuses on particular interactions or other tests in which there is specific reason to expect discrepancies from the model to be located. For example, contrast or other patterning effects might be expected to appear largely in the two-way interactions of successive serial positions.

Since this was an initial investigation of length averaging and since a previous experiment on loudness averaging (Parducci, Thaler, & Anderson, 1968) had raised some doubt about the model, it was decided to employ a complete design and analysis. Fortunately, the general pattern of the results supports the use of the aforementioned methods to reduce the number of interactions tested. In fact, confounding formed an essential part of the design in Exp. III.

Experiment I

In this experiment, the response was made only at the end of the sequence. Two aspects of these data are of interest: first, the test of the averaging model; second, the serial position curves.

Goodness of fit.—A summary of the tests of the model is given in Table 1. The tabulations include all separate interactions that were significant for each S, together with an overall F ratio based on all interactions pooled. The main data, for sequences of six lengths, are in the left half of the table. Since the 2^6 design has 57 interaction terms, an average of 3 separate interactions would be expected to be significant by chance alone. On such a rough assessment, $S1$

TABLE 1

SUMMARY TEST OF FIT, EXP. I

S	Six-length sequences				Three-length sequences			
	Interactions[a]	MS_e	Interaction F ratio[b]	r[c]	Interactions[a]	MS_e	Interaction F ratio[d]	r[c]
1	12, 23, 25, 26, 135, 145, 146, 345, 1256, 1356, 12346	.95	2.11*	.910	123	.70	3.88*	.981
2	12345	1.59	.82	.956	None	.98	.51	.996
3	23456	1.95	.90	.951	None	1.14	.26	.998
4	12	.95	.81	.963	123	.41	11.66*	.974
5	12, 34, 46, 234, 2356	2.48	1.14	.923	None	2.26	.13	.998
6	25, 36, 1246, 12356	2.64	1.25	.902	None	1.76	.44	.995
7	None	1.79	.63	.953	123	.19	3.50*	.995
8	56, 13456	5.75	.68	.944	123	.84	3.87*	.992

[a] Code numbers indicate significant interactions; each digit represents corresponding serial position.
[b] F ratio for pooled interactions, df = 57 and 192.
[c] Correlation (observed, predicted).
[d] F ratio for pooled interactions, df = 4 and 56.

clearly deviates from the model, showing 11 separate interactions, as well as a significant pooled interaction. The other seven *S*s appear to satisfy the test of fit reasonably well.

Inspection of Table 1 fails to show any clear pattern among the significant interactions; e.g., each serial position is represented with nearly equal frequency. Inspection of the data tables for the two-way interactions also failed to show any pattern. However, it was possible to localize the discrepancies for *S*1 in a single sequence. Inspection of the raw data showed an unusually high response to that sequence containing the longer length at all six serial positions. Subtracting a constant, 3.9 cm., from the response to each replication of this sequence reduced the number of significant interactions from 11 to 2, about what would be expected by chance. No other unusual responses were found, and the cause of this one discrepancy is unknown.

Deviations from the model that were consistent across *S*s would tend to show up in the overall test. In this group analysis, however, only two high-order interactions were significant, no more than would be expected by chance. This situation remained unchanged when the pooled within-*S*s error on 1,536 *df* was used for all the tests. These results suggest that what real discrepancies there are from the model are idiosyncratic. They may present difficulties, but they do not reflect any uniform preceptual effect.

The right half of Table 1 shows the summary for the last phase of the experiment, in which each sequence had three lengths. No two-way interactions were significant, but four *S*s show a significant 123 interaction. However, it took a different form in each case and did not approach significance in the group analysis, even when tested against the pooled within-*S* error on 448 *df*.

The seriousness of the interactions depends heavily on their magnitudes. It is not feasible to report these in detail, but two data columns in Table 1 have relevant information. The first is the overall *F* ratio, already noted, which gives the

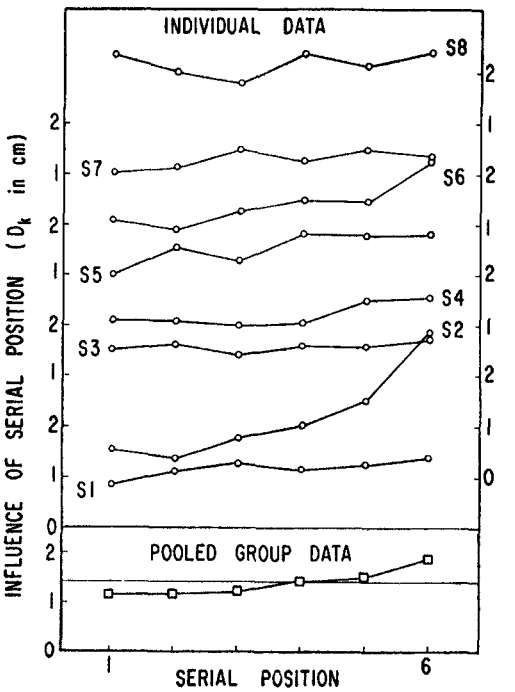

FIG. 1. Serial position curves, Exp. I. (Individual curves displaced upward, as indicated by repeating ordinates.)

magnitude of the total interaction component relative to the error variability. These are all near unity, except for *S*1. The second is the correlation between the observed mean response and that predicted by the model. These correlations are, in effect, the ratios of the *SS* for main effects to the total *SS* for systematic sources. Thus, the degree to which they fall short of unity is an index of the relative size of the pooled interactions. Although these correlations seem satisfactorily high, they must be interpreted with due care. In particular, that smaller correlations are obtained for six lengths than for three lengths is almost entirely artifactual, arising from the more numerous interaction terms in the former designs.

Serial position curves.—These curves, in Fig. 1, show the differential effect of the 24- and 16-cm. lengths as a function of serial position. With one exception, the curves are fairly similar in shape, showing a mild recency effect. For graphical clarity, the

curves are separated vertically, as indicated in the graph.

The calculation and interpretation of these serial curves require a brief comment. The entry for Serial Position k is the difference, D_k, between two marginal means: for all those sequences that had the 24-cm. length at Position k, and for all those sequences that had the 16-cm. length at Position k. If the model is correct, then this observed difference is an estimate of $w_k(s_{24} - s_{16})$, where w_k is the weight for Position k and s_{24} and s_{16} are the subjective lengths of the two stimuli. If the difference in subjective lengths is indeed constant across serial position (see Exp. II), then

TABLE 2

SUMMARY TEST OF FIT: SIX-LENGTH SEQUENCES, EXP. II

Interaction	S									
	1	2	3	4	5	6	7	8	9	10
Serial Position 6										
12			4.49			5.89	6.99			5.05
34			4.19				4.92			
56	4.85									
134					6.48					
345					5.94					
1235					4.40					
1245		4.24								
1346									4.82	
12345				7.18						
12346		10.49	4.79							
13456				4.48						
Main[a]	632.50	185.92	223.77	38.01	158.18	86.37	175.22	219.52	164.72	64.85
MS_e (192 df)	.51	.50	.47	2.57	.85	1.77	.63	.83	.57	1.58
Serial Position 5										
12			6.82			6.67	8.90			
23			4.68							
34							4.70			
45				7.73						
134		3.89			6.12					
234							5.58			
345					6.26		4.85			
1235					6.61					
12345					6.75	4.70	4.44			
Main[a]	666.06	227.04	343.48	62.08	229.26	142.54	236.48	316.85	209.04	99.58
MS_e (224 df)	.57	.60	.41	2.63	.80	1.45	.61	.68	.57	1.45
Serial Position 4										
12			10.68			5.72	8.69			7.96
13			5.53							
23		4.40	5.53							
24					5.31		8.88			
123		4.01								
124						10.21			4.40	
134					4.86				11.23	
Main[a]	1044.13	343.15	578.05	105.13	356.73	214.24	367.90	368.73	365.60	162.68
MS_e (240 df)	.47	.61	.36	2.37	.76	1.43	.59	.74	.49	1.38
Serial Position 3										
12			12.75			5.70	11.30		3.95	
13			10.68							
23			5.26	6.61	4.86					
123									24.27	
Main[a]	1331.20	578.73	1015.33	287.50	596.50	355.50	602.67	627.37	704.40	290.43
MS_e (248 df)	.51	.66	.38	2.03	.76	1.36	.61	.65	.47	1.44
Serial Position 2										
12			23.00	13.67		4.70	14.61			3.93
Main[a]	2153.33	1213.73	2075.86	772.48	1241.37	731.95	1407.93	870.68	1845.90	630.59
MS_e (252 df)	.59	.71	.42	1.53	.76	1.50	.59	.76	.42	1.52
Serial Position 1										
MS_e (254 df)	.77	.85	.49	.98	1.00	1.35	.46	.61	.50	1.78

[a] Mean F ratio for main effects.

the curves of Fig. 1 are proportional to the values of w_k. In other words, these serial curves are the weight curves.

The serial curves for the sequences of three lengths are not shown here, but they also showed a mild recency. Moreover, each *S* showed about the same curve shape for three lengths as for six. This individual consistency in serial position curves was also found in Exp. II.

Experiment II

In this experiment, *S* responded to each successive length in the sequence. The analysis of the terminal response is exactly the same as in Exp. I. In addition, the data at each earlier serial position may be analyzed in the same way. The successive responses in a sequence are not independent, of course, and neither are the analyses. Nevertheless, the additional responses do give useful information.

Goodness of fit: Six lengths.—An overview of the tests of the model is given in Table 2. All interaction *F* ratios that were significant for each *S* for each serial position are listed. For comparative purposes, the mean *F* ratio for the main effects at each serial position is also included.

The upper section of Table 2 summarizes the tests of the response at Serial Position 6. These entries are comparable to the corresponding analysis of Exp. I summarized in Table 1. On a gross count, the two experiments show about the same number of significant interactions. In Exp. II, however, there is a marked tendency toward significant 12 interactions, between Serial Positions 1 and 2. This same tendency is found in the analysis of each of the previous responses. In addition, *S*3, *S*7, and perhaps *S*5 show more significant interactions than would be expected by chance. Indeed, *S*3 and *S*7 show a significant 12 interaction at each response.

All told, the 12 interaction accounts for nearly a third of the significant *F* ratios. The initial response is just a reproduction of the first length and presumably presents no difficulty. Accordingly, the source of the interaction would seem to be at Serial Position 2, the first position at which *S* is required to average. As already noted, successive responses are not independent, and it is straightforward to show that an interaction induced at one position tends to reappear automatically at subsequent positions. Presumably, therefore, the 12 interactions at the later positions reflect a discrepancy perpetuated from Serial Position 2.

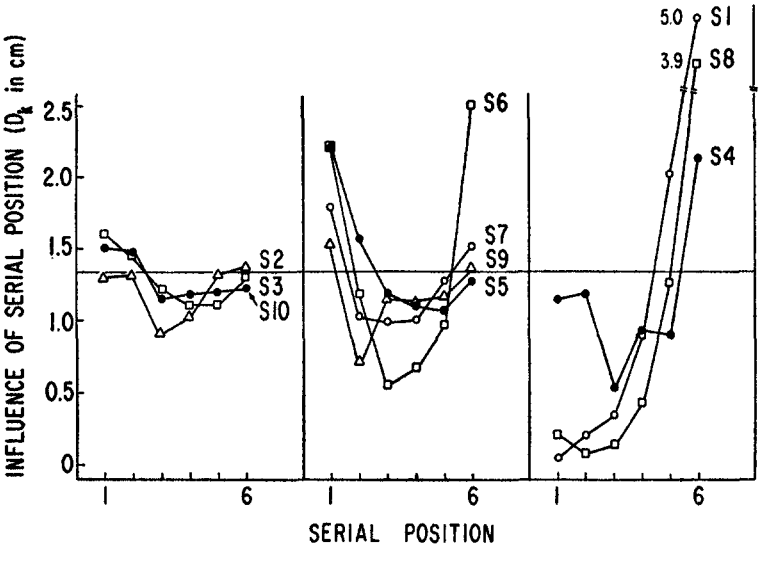

FIG. 2. Serial position curves for each *S*, Exp. II. (Curves based on data from response at Serial Position 6.)

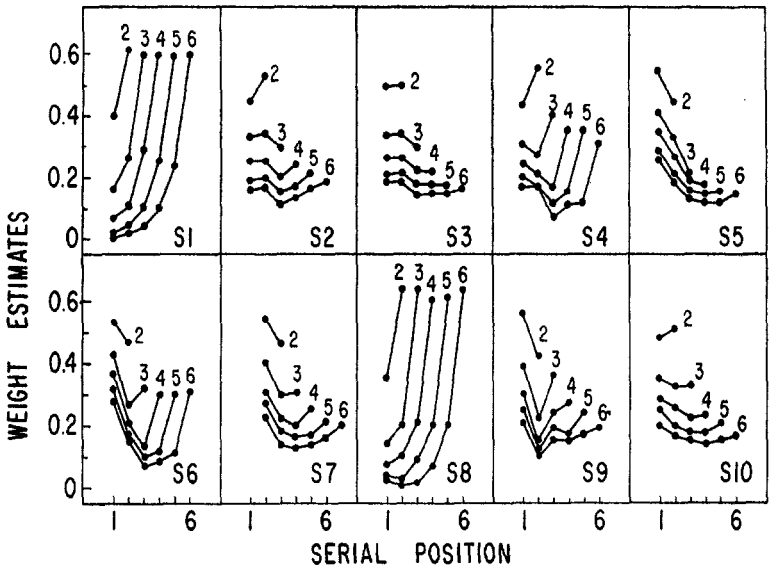

FIG. 3. Serial position curves for each *S*, for each successive response, Exp. II. (Numbers by curves in each panel index successive responses within sequence.)

Other than the 12 interaction, not much pattern is evident in Table 2. The mean total of significant interactions is 5.9 per *S*, only slightly greater than would be expected by chance. Even for the 12 interaction at Serial Position 2, there was no consistent pattern across *S*s. It had a different shape for different *S*s and was not significant in the group analysis.

Response variability was roughly constant across serial position. This can be seen by inspection of the error mean squares of Table 2. Despite considerable individual differences, each *S* shows about the same variability in response at each serial position. The detailed analyses of the successive responses for the sequences with 10 lengths showed a similar pattern.

Serial curves: Six lengths.—The serial curves calculated from the terminal response are shown in Fig. 2. For graphical clarity, *S*s have been grouped according to curve shape, which varies considerably. Some show extreme recency (right panel), some show relatively flat serial curves (left panel), and others show both primacy and recency components (center panel). This variety in curve shape is markedly greater than for Exp. I (Fig. 1), in which only a terminal response was made.

Serial curves may also be obtained from the earlier responses and these are shown in Fig. 3. These plots are in terms of *w* values in order to facilitate comparison of the several curves for each *S*. As can be seen, each *S* tends to have a stable characteristic curve shape.

Scale values: Six lengths.—The scaling procedure described in the introduction provides estimates of the subjective lengths at each serial position. These are shown

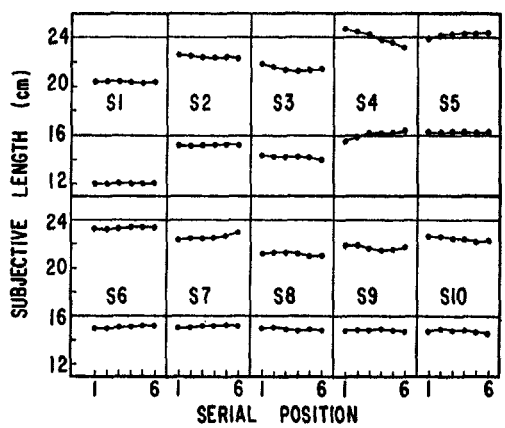

FIG. 4. Subjective length estimated from response at each serial position, Exp. II. (The two curves for each *S* give subjective lengths of long, 24-cm., and short, 16-cm., lengths.)

in Fig. 4. For nearly every *S*, subjective length is constant over serial position. This indicates that there are no within-sequence adaptation effects.

Figure 4 shows marked individual differences in the subjective lengths. These individual differences are real enough since the first point on each curve is nothing but a direct estimate of the first length in the sequence. This fact, it may be emphasized, helps validate the scaling procedure. Curiously enough, the difference between the two subjective lengths is fairly close to its arithmetically correct value of 8 cm. for every *S*. The individual differences in subjective length can, therefore, be considered as constant errors.

Serial curves: Ten lengths.—The main function of these data was to yield serial curves for longer sequences, and only the terminal response is considered here. The tests of fit showed about the same general picture as in the main phase. Of the 228 interactions, only 14 reached significance. However, *S*3 showed an array of discrepancies similar to that noted previously, and *S*4 appeared to have some possibly real discrepancy from the model. As before, these discrepancies tended to be relatively small in magnitude, and inspection of the data failed to show any pattern among them.

The serial curves are shown in Fig. 5. Although there are 10 serial positions, there are only 6 points on the curves since the design used pairwise confounding over the first 8 serial positions. The most important aspect of these data is that *S*s show the same shape curve with 10 lengths as with 6 (compare Fig. 5 with Fig. 2). This individual consistency in shape of serial curve confirms that found in Exp. I.

Averaging accuracy.—In the length-averaging task, there is an objective criterion of accuracy to which the behavior may be compared. With serial presentation, of course, the recency effect guarantees inaccuracy. Nevertheless, the balance in the stimulus design allows a comparison of objective length and subjective length independent of recency. This is obtained by averaging over all sequences with the same frequency distribution of lengths regardless

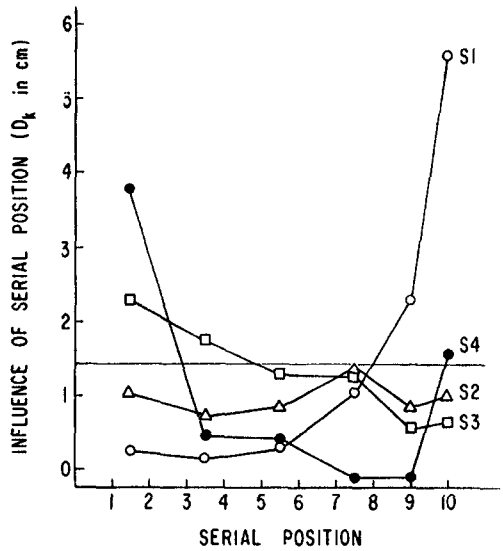

FIG. 5. Serial position curves for the four *S*s in Phase 2 of Exp. II. (Compare with Fig. 2.)

of their position. Figure 6 plots subjective average as a function of objective average for the sequences of six lengths. The second point on the curves is the mean response to all six sequences that had just one short length, etc. With end responding (Exp. I), *S*s are quite accurate. With continuous responding (Exp. II), marked underestimation occurs. This constant error parallels that of Fig. 4. Amount of constant error varies from *S* to *S*, but to the degree that the behavior follows the averaging model, each *S*'s curve would be parallel to the group curve.

Experiment III

A method of simultaneous-serial presentation was adopted in this experiment to test two hypotheses about the recency effect observed in Exp. I and II. The seven types of sequences are shown in the left column of Table 3. In each row, the first number in the sequence type is the number of lengths presented simultaneously in the initial position of the sequence; thereafter only one length was presented at a time.

The entries in Table 3 are the weights estimated from the terminal response. The weight for the first serial position, which represents several stimuli, was reduced to a per stimulus basis as indicated in the table. Two principal features of Table 3 merit comment. First, w_6 is the largest entry in each row, a sizable recency effect. Second, the prorated weight of a member of the initial set is roughly constant for sets of more than one length.

To understand the two hypotheses about the recency effect, *S*'s task should be kept clearly in mind. At any step in the sequence, both the new stimulus and *S*'s previous estimate are in view. All *S* need do is average them, and Exp. I and II indicate that he does just that. Since both relevant stimuli are visibly present, the recency presumably does not arise from any misperceptions or memory changes in the scale values. Accordingly, the problem is why *S* overweights the new stimulus and underweights the previous estimate.

Now the weight associated with the previous estimate is, in fact, directly related to the number of previous stimuli. This weight, therefore, may be interpreted as a subjective *N*. The first hypothesis is that *S* does not keep in mind the exact number of previous stimuli, but that this subjective *N* undergoes a temporal decay. If this is true, then the recency should be less for the shorter sequences, in which several stimuli are given simultaneously at Serial Position 1.

This hypothesis is tested by the first five sequence types of Table 3. It implies that w_6 should decrease as the size of the initial simultaneous set increases. Inspection of the data shows no trend, and the statistical test was nonsignificant, F (4, 92) = 1.99. Similar data from the fifth and fourth responses also showed no effect of size of initial set. In short, there is no support for the first hypothesis.

The second hypothesis states that the overweighting of the present stimulus is caused by lack of confidence in the previous response. The last four rows of Table 3 test this hypothesis. In the last two rows, the four or five lengths of the initial set were all equal; in contrast, they were all different for the next-to-last two rows. The confidence interpretation would imply less recency for the "same" than for the "different" condition.

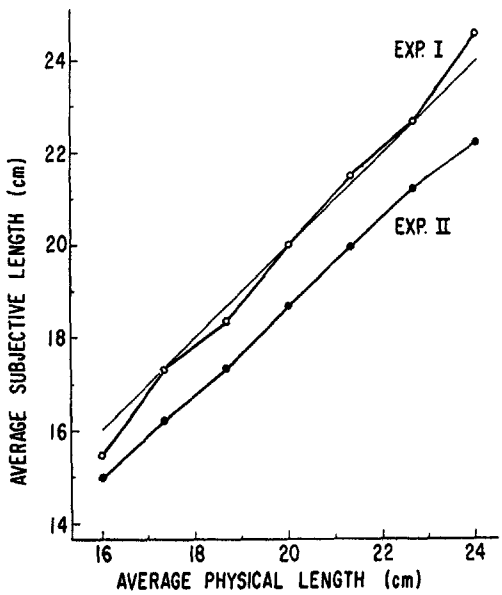

FIG. 6. Subjective average as a function of objective average, sequences of six lengths, with recency effect averaged out.

TABLE 3

EFFECTS OF EACH SERIAL POSITION ON FINAL RESPONSE: WEIGHT ESTIMATES, EXP. III

Sequence type	w_1	w_2	w_3	w_4	w_5	w_6
1-1-1-1-1-1	.21	.12	.13	.13	.16	.26
2-1-1-1-1	(.14)	(.14)	.14	.12	.14	.31
3-1-1-1	(.13)	(.13)	(.13)	.16	.19	.24
4-1-1	(.14)	(.14)	(.14)	(.14)	.13	.33
5-1	(.15)	(.15)	(.15)	(.15)	(.15)	.26
4-1-1 (same)	(.16)	(.16)	(.16)	(.16)	.15	.21
5-1 (same)	(.15)	(.15)	(.15)	(.15)	(.15)	.25

Note.—Entries in parentheses are weights of initial set on a per stimulus basis. Initial simultaneous set in last two rows had all equal lengths. Sum of w values in some rows differs from 1.00 because of rounding error.

Some slight support for the confidence hypothesis was obtained from Sequences 4-1-1 since the difference between the estimates of .33 and .21 for w_6 was significant, F (1, 23) = 8.43. However, the corresponding test on w_5 estimated from the previous response was not significant. Moreover, the data for Sequences 5-1 in Table 3 show no effect.

DISCUSSION

This discussion takes up two main topics. The first is the problem of psychophysical integration, here exemplified by the serial averaging task. On the whole, despite certain discrepancies noted previously, the model seems to have done reasonably well for subjective length averaging. This result is not trivial. It is true that *S*s were instructed to "average," but they were under no constraint about how they averaged. There are many measures of central tendency, and each *S* was free to choose his own. This averaging rule could even be variable, depending on the particular sequence of lengths. It has considerable interest, therefore, that the behavior was well described as a weighted arithmetic mean, with weights dependent only on serial position. Moreover, this result is important in the interpretation of the serial position curves.

In this connection, the predominant recency observed in each of the experiments needs consideration. When a response was required only at the end of the sequence, individual differences were relatively small (Fig. 1) and all *S*s showed moderate recency. But when a response was required at each successive serial position, individual differences were extreme (Fig. 2, 3, and 5). Although these two experiments differed in procedural details and so are not strictly comparable, the difference

between them is sufficiently strong to deserve attention.

Memory storage requirements might be related to the difference in range of recency in the two main experiments. With continuous responding, no memory of the individual previous stimuli is required since their average is represented visibly in the current position of the response indicator. Even so, S still must assign weights in some manner, and this requires a revision at each successive serial position of the relative weight to be assigned to the cumulated average. The large individual differences obtained with continuous responding would then presumably be cognitive, rather than perceptual, reflecting different weighting strategies. This is consistent with the within-S consistency in curve shape. With end responding, on the other hand, S could simply store the several subjective lengths more or less separately and make only one overall integrative response, thereby avoiding the successive weighting revisions. Task variations based on intermittent responding and component recall might help pin down this problem.

In any case, the recency itself remains to be explained. For continuous responding, Exp. III indicates that recency does not arise from decay of subjective N with time or serial position. The second recency hypothesis, that lack of confidence in the previous judgment produced recency, received only weak, equivocal support. In view of the large individual differences with continuous responding, it might be preferable to use an end responding procedure in future tests of this kind.

Whatever the cause of the recency, it appears regularly in serial averaging of simple perceptual stimuli. Besides the present length dimension, recency has been found with loudness (Parducci et al., 1968) and lifted weights (Anderson, 1967; Anderson & Jacobson, 1968), as well as with numbers (Anderson, 1964b). These stimulus classes do not, however, give uniform support to the averaging model. Loudness averaging, in particular, showed a sizable discrepancy. In weight averaging, the model did well with six weights, but a possible discrepancy appeared with three weights. This is consistent with the present indications of interactions at the second and third serial positions.

The second topic for discussion is functional measurement and its application to psychophysical scaling. Functional measurement provides a basis for scaling both stimulus and response dimensions at the same time. The essential ideas were given in Anderson (1962a, 1962b); recent applications are given by Shanteau and Anderson (1969) and Anderson and Jacobson (1968). Psychophysical averaging involves not only the subjective value of the stimulus, but also its subjective weight. The two are not always separable, but the continuous responding procedure has technical interest because it provides a basis for measuring both value and weight at each serial position.

A key feature of functional measurement is the use of an integration task, in which several stimuli are to be combined into a single judgment. The quantities that S combines are, of course, the subjective values of the stimuli. If the integration rule is simple, then the stimulus values may appear fairly directly in the response. This is one reason for attempting to develop an averaging model: in its simplest applications, the marginal response means of the stimulus design are equivalent, up to a linear transformation, to the subjective values of the stimuli. This remains true, of course, when more than two stimulus values are used in each factor of the design, a desirable property in determining the psychophysical law relating subjective and objective stimulus values.

Garner (1954a, 1954b), Garner and Creelman (1967), and Treisman (1964) have pointed out the perplexities in the interpretation of psychophysical scales based on direct, numerical response methods. For the most part, these methods obtain judgments of single stimuli and yield results of the form $R = F(S)$, where R is the observed response, S is the physical value of the stimulus, and F is the presumptive psychophysical law. Ordinarily, S is known. Hence, if R could be taken at face value, F would be completely determined, and the problem would be solved. But if R is simply assumed to be valid, then any observed relationship is arbitrary. As both Garner and Treisman have emphasized, numerical response methods typically take R at face value and provide no means for assessing that assumption. Similarly, Luce and Galanter (1963) have noted that numerical biases may vitiate the use of numerical response measures.

Numerical response measures can often be considered no more than ordinal scales. The proper response scale, R_T, if it exists, will then be some monotone transformation, $R_T = \mathfrak{M}(R)$, of the observed response (Anderson, 1962b). If $R = F(S)$ is observed, then the

proper psychophysical law will be $\mathfrak{M}(F)$ since $R_T = \mathfrak{M}(R) = \mathfrak{M}(F(S))$. In the absence of any validation criteria, all continuous, strictly monotone transformations of F would have equal claim to be the psychophysical law.

The methods proposed here avoid this indeterminacy. The integration task provides a basis for determining a relation of the form $R = f(s_1, s_2, \ldots, s_N)$, where s_k is the subjective value of the kth stimulus. Because R is based on more than a single stimulus, constraints are available that allow application of a monotone rescaling procedure for the response dimension (Anderson, 1962b). In the present case, f was a weighted average, and the observed R satisfied the model so that no response rescaling was needed. Given the validity of the response, the subjective values of the stimuli could be derived fairly simply, in part because the integration rule was simple, in part because the stimulus combinations were constructed according to a factorial design. In this formulation, the psychophysical law is expressed as $s_k = F(S_k)$ and would appear as a by-product of the investigation. To determine this function would require using more values of S_k, of course, but in principle this is no problem.

The validity of functional measurement will depend on getting the same scales with different integration tasks. This follows Seward's (1955) requirement of constancy of the intervening variable, as well as the emphasis on converging operations of Garner, Hake, and Eriksen (1956). For this reason, it is desirable to develop alternative integration tasks. Simultaneous presentation has certain advantages over serial presentation. For length, both averaging and summing tasks may be used, with many variations of serial and simultaneous presentation. Few stimulus dimensions have this flexibility, but an averaging task appears feasible with almost any dimension. Indeed, traditional bisection judgments may be considered as averages of two stimuli, analogous to the present response at the second serial position. Although the present approach depends primarily on gross stimulus differences and a numerical response, choice data based on fine stimulus differences would also be of interest. There are various ways to begin the analysis of choice data, but the discriminant function approach of Rodwan and Hake (1964) has special interest and potential.

REFERENCES

ANDERSON, N. H. Application of an additive model to impression formation. *Science*, 1962, **138**, 817–818. (a)

ANDERSON, N. H. On the quantification of Miller's conflict theory. *Psychological Review*, 1962, **69**, 400–414. (b)

ANDERSON, N. H. Note on weighted sum and linear operator models. *Psychonomic Science*, 1964, **1**, 189–190. (a)

ANDERSON, N. H. Test of a model for number-averaging behavior. *Psychonomic Science*, 1964, 1, 191–192. (b)

ANDERSON, N. H. Application of a weighted average model to a psychophysical averaging task. *Psychonomic Science*, 1967, **8**, 227–228.

ANDERSON, N. H. Partial analysis of high-way factorial designs. *Behavior Research Methods and Instrumentation*, 1968, 1, 2–7.

ANDERSON, N. H., & JACOBSON, A. Further data on a weighted average model for judgment in a lifted weight task. *Perception and Psychophysics*, 1968, **4**, 81–84.

COCHRAN, W. G., & COX, G. M. *Experimental designs.* (2nd ed.) New York: Wiley, 1957.

GARNER, W. R. Context effects and the validity of loudness scales. *Journal of Experimental Psychology*, 1954, **48**, 218–224. (a)

GARNER, W. R. A technique and a scale for loudness measurement. *Journal of the Acoustical Society of America*, 1954, **26**, 73–88. (b)

GARNER, W. R., & CREELMAN, C. D. Problems and methods of psychological scaling. In H. Helson & W. Bevan (Eds.), *Contemporary approaches to psychology.* Princeton, N. J.: van Nostrand, 1967.

GARNER, W. R., HAKE, H. W., & ERIKSEN, C. W. Operationism and the concept of perception. *Psychological Review*, 1956, **63**, 149–159.

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

PARDUCCI, A., THALER, H., & ANDERSON, N. H. Stimulus averaging and the context for judgment. *Perception and Psychophysics*, 1968, **3**, 145–150.

RODWAN, A. S., & HAKE, H. W. The discriminant function as a model for perception. *American Journal of Psychology*, 1964, **77**, 380–392.

SEWARD, J. P. The constancy of the *I-V*: A critique of intervening variables. *Psychological Review*, 1955, **62**, 155–168.

SHANTEAU, J. C., & ANDERSON, N. H. Test of a conflict model for preference judgment. *Journal of Mathematical Psychology*, 1969, **6**, 312–325.

TREISMAN, M. Sensory scaling and the psychophysical law. *Quarterly Journal of Experimental Psychology*, 1964, **16**, 11–22.

(Received January 17, 1969)