Recognition Confusions among Quantifiers

KEITH J. HOLYOAK

University of Michigan

AND

ARNOLD L. GLASS

Rutgers University

Subjects listened to a story containing sentences with five different quantifiers (*all, many, some, a few,* and *none*). They were then given a recognition test in which for each sentence they were asked to rate all the quantifiers as to how certain they were that each had appeared in the sentence. The degree of confusion between any two quantifiers (measured by both mean certainty ratings and first choice frequencies) declined monotonically with the separation of the two terms in a linear order. The confusion pattern was identical regardless of whether subjects were given intentional or incidental learning instructions. Successive interval scaling and the Luce biased choice model were both used to derive unidimensional scales of quantifier values that accurately described the obtained confusion matrix. A semantic feature model was also able to provide a fairly good overall account of the data, except that it had difficulty explaining the high frequency of confusions between *all* and *many*. The results raise the possibility that a major component of the memory trace for quantifiers is an essentially analog representation of magnitude.

English and other natural languages contain a number of words that express degrees of quantity (e.g., *all, some, several, often*). Most of these words are frequently used and apparently easily understood by speakers of the language. However, underlying their apparent simplicity is a high degree of semantic complexity, which is evidenced by numerous subtle inferences and ambiguities. Recent psychological approaches to language and

memory (Anderson, 1976; Anderson & Bower, 1973; Kintsch, 1974) have had difficulty developing a representation system for quantifiers sufficient to explain the linguistic puzzles they create. The importance of quantification in everyday language use and the difficulties it poses for the psychology of language suggest the need for detailed studies of how people understand, use, and remember quantifiers.

Psychological studies of quantification have focused on the related areas of reasoning and sentence verification. A large body of work deals with the kind of inferences people draw in syllogistic reasoning problems (Erickson, 1974; Revlis, 1975; Wason & Johnson-Laird, 1972). Other studies have examined the processes by which people verify quantified sentences on the basis of information in pictures (Just & Carpenter, 1971) or in semantic memory (Glass & Holyoak, 1974; Glass,

Holyoak, & O'Dell, 1974; Holyoak & Glass, 1975; Meyer, 1970). The dependent measures in most of this work have been either the probability of drawing particular valid or invalid inferences or the verification reaction time. The present study investigated a different aspect of the psychological representation of quantifiers, the way in which they are coded in episodic memory; and it employed a different experimental method, the analysis of recognition confusions among a set of five related quantifiers (all, many, some, a few, and none).

Numerous studies have used recognition confusions to investigate the form in which sentences are stored in memory (Bransford, Barclay, & Franks, 1972; Fillenbaum, 1966; Sachs, 1967). These studies have consistently found that recognition confusions are largely determined by similarity in meaning. However, few studies of word meaning have examined confusion matrices based on a set of words from the same semantic field. One exception is a series of studies by Clark and his associates (Clark & Card, 1969; Clark & Clark, 1968; Clark & Stafford, 1969). Their data were based on recall errors rather than recognition confusions. For example, Clark and Stafford examined error matrices for verbs that systematically differed in tense and aspect. Although few studies of memory for word meanings have been based on confusion matrices, such analyses have been used extensively in areas such as immediate memory (Conrad, 1964) and letter perception (e.g., Pachella, Smith, & Stanovich, 1977).

POSSIBLE PATTERNS OF CONFUSION AMONG QUANTIFIERS

The fundamental assumption underlying the use of confusion matrices to explore memory for word meanings is that the probability of selecting $word_j$ when $word_i$ has been presented depends on the similarity, at the time of the recognition test, between the meaning of $word_j$ and the memory trace for the meaning of $word_i$. The more similar the psychological representations of $word_i$ and $word_j$, the more likely they are to be confusable at the time of testing. While other factors, such as response biases, may also influence confusion probabilities, the confusion matrix should in part reflect the similarity structure inherent in the mental representations of the critical words.

Given this general framework, there are a number of possible hypotheses about the manner in which quantifier meanings may influence memory confusions. Three general hypotheses will be outlined.

All-or-none hypothesis. A relatively uninteresting but not entirely implausible hypothesis is that recognition memory for the five quantifiers will be all or none. That is, either the subject will remember the word that was presented and respond correctly, or he will forget it and guess among all five alternatives with equal probability (except for response biases). This is equivalent to assuming that the memory traces for all the quantifiers are equally similar to one another. Since the set of items being studied is indeed very homogeneous in meaning, it is possible that confusion data may be insensitive to the relatively small differences in the representations of the five words, producing a confusion pattern consistent with the all-or-none hypothesis.

Semantic overlap hypothesis. A more interesting possibility is that confusion probabilities will reflect the semantic structure of the quantifiers, as indicated by linguistic analyses. A possible analysis of the quantifiers used in the present study, based mainly on Leech (1969), is presented in Table 1. This analysis is not meant to be definitive, but only to give an outline of the relations among the meanings of the five words.

The two basic terms in the set are the universal quantifier *all* and the existential *some.* These terms form an inverse opposition (Leech, 1974), and must be represented with opposing values on a common attribute in order to explain semantic equivalences of the following form.

TABLE 1
SEMANTIC RELATIONS AMONG THE QUANTIFIERS

Quantifier	Semantic structure
All	$\text{Quantifier}_{\text{universal}}$
Some	$\text{Quantifier}_{\text{existential}}$
Many	$\text{Some} + \text{number}_{\text{large}}$
A few	$\text{Some} + \text{number}_{\text{small}}$
None	Neg(some)

(1) Some of the cats don't like fish = Not all of the cats like fish.
(2) All of the cats don't like fish = None of the cats like fish.

That is, for a sentence containing a negative, interchanging *some* and *all* and changing the position of the negative produces a sentence with the same meaning.

The other three terms can be derived from the meaning of *some*. *Many* and *a few* are both existential quantifiers that are further specified for numerosity: Roughly, *many* indicates a "large number" and *a few* indicates a "small number."[1] Thus, each contains the meaning of *some*, plus an additional attribute of "number" with the value "large" or else "small." Finally, *none* is the negation of *some*, represented as neg(Some), e.g., *none* is the surface realization of "not some" in the right-hand sentence of (2) above.

If the semantic structures indicated in Table 1 form the memory code for the quantifiers, it is possible to make a number of predictions about recognition confusions. First, *many* and *a few* should both be confusable with *some*, from which they differ by just one additional feature. Also, the semantic feature hypothesis as it has been applied to other language domains stresses the principle that features can be lost from memory more easily than they can be added (Clark & Clark, 1968). It follows that a number of confusion asymmetries should be observed. *Many* and *a few* should be

[1] The quantifier *a few* used in the present study should not be confused with the negative *few*, which would be represented as neg(Many), i.e., "not a large number" (Glass et al., 1974; Just & Carpenter, 1971).

confused with *some* more often than *some* is confused with *many* or *a few*. Similarly, *none* should be confused with *some* (if the negation feature is lost) more often than the reverse. Since neither *all* nor *some* is more basic than the other, no asymmetry need be observed in the confusions between these two words.

The analysis presented in Table 1 does not completely specify a predicted confusion pattern. For example, whether *many* or *none* is more confusable with *some* should depend on whether the number or negation feature is more easily lost, which the hypothesis does not directly predict. Also, the addition of further features to the analysis could change some predictions. For example, Table 1 suggests that *all* will be more often confused with *some* (one feature change) than with *many* (the same feature change plus a feature addition). However, it is possible that people may take *all* to imply the feature "$\text{number}_{\text{large}}$," providing a possible link between *all* and *many*. Similarly, "$\text{number}_{\text{small}}$" may be added to the representation of *none*, increasing its similarity to *a few*. It is even possible that *all* and *none*, which unlike the other three terms specify definite and absolute quantities, will therefore share a common feature and hence be confusable to some degree. However, these additional speculations lack the linguistic justification of the analysis presented in Table 1. A formal model of confusion frequencies based on the analysis in Table 1 will be developed and tested below.

Linear order hypothesis. This third hypothesis hinges on the fact that the five terms form an ordinal quantity scale: *all*, *many*, *some*, *a few*, *none*. (While *some* can be used to cover the range of *many* and *a few*, in conversation *some* would most often indicate a quantity intermediate between *many* and *a few*.) The linear order hypothesis assumes that the memory representation of a quantifier can be represented as a value on an internal scale. Given variability in the scale values, quantifiers will sometimes be confused with other

quantifiers with similar scale values. In fact, the linear order hypothesis predicts that the obtained confusion matrix will fit a simple pattern of single-peaked functions (ignoring possible response biases): The probability of a recognition confusion between two quantifiers will decrease monotonically as the separation between the two scale values increases. It should therefore be possible to use the confusion data to recover a unidimensional scale on which the stimuli can be placed to account for the degree of confusability between all stimulus pairs. The quantifiers need not be spaced at equal intervals on the scale; in fact, the resulting scale should yield information about the metric properties of the quantifier ordering.

The linear order hypothesis is closely related to current models of the processing of symbolically ordered stimuli, such as digits (Holyoak, 1978; Moyer & Landauer, 1967), objects ordered in size on the basis of memory (Holyoak, 1977; Moyer, 1973; Paivio, 1975), and experimentally taught orderings (Potts, 1972, 1974; Trabasso & Riley, 1975; Woocher, Glass, & Holyoak, 1978). Much of this work suggests that symbolic orderings are represented in memory as analog magnitude values or positions along an internal array. For all of the above types of stimuli, subjects can select the greater of two stimuli from memory more quickly the larger their subjective magnitude difference, suggesting a possible "internal psychophysics" (Moyer, 1973). This "symbolic distance effect" has also been found with terms from semantic orderings, such as time terms (e.g., *second, minute, hour,* etc.) (Holyoak & Walker, 1976). Since the quantifiers used in the present study form a semantic ordering, it is possible that they may be represented in memory in a similar fashion. However, none of the previous linear order studies has examined recognition confusions among ordered concepts.

Note that the semantic overlap hypothesis and the linear order hypothesis differ in their predictions about confusion asymmetries. As

discussed above, the semantic overlap hypothesis predicts that confusions between a given pair will sometimes be more frequent in one direction than another, since a semantically complex term will be more readily transformed into a simpler term than vice versa. Pair similarities should therefore depend on direction. In contrast, the linear order hypothesis predicts that similarity (and hence confusability) will depend on a distance metric, and, since distance is inherently symmetric, confusion asymmetries should only result from some form of response bias.

METHOD

Recognition test. The five critical quantifiers (*all, many, some, a few, none*) were presented to subjects in the context of the following brief "story" (our apologies to George Orwell):

The Citizens of Animal Farm

This is a story about a very strange farm. On this farm the animals had taken over complete control of their own affairs. Many different kinds of animals lived together in harmony. There were the large animals, the cows and horses; and the small farm animals, pigs, goats, chickens and geese. Dogs and cats were also included, as well as normally wild animals, the mice and squirrels.

The customs and habits of the citizens of Animal Farm were very diverse. They all enjoyed different pastimes. Q_1 of the cows watched television. Q_2 of the horses read books. Q_3 of the chickens listened to the opera. Q_4 of the pigs went dancing. Q_5 of the cats played football.

Their lifestyles also varied a great deal. Q_5 of the squirrels sat on the farm council. Q_4 of the mice worked in the local bank. Q_3 of the goats attended college. Q_2 of the dogs were reliable mechanics. Q_1 of the geese sang with a band.

It might seem that with this degree of cultural diversity, friction among the various groups would be inevitable. But strangely enough, that was not the case. When problems arose, someone would ask whether they would prefer it if the humans were running things again. That was always enough to bring about a compromise.

The quantifiers appeared at the positions Q_1–Q_5 in the second and third paragraphs

(i.e., each quantifier appeared twice). The order of the quantifiers always differed between the two paragraphs. Across five different groups of subjects, each quantifier was placed equally often in each of the 10 sentences, in order to control for possible variations in the memorability of particular sentences or serial positions. It was hoped that the story framework would induce subjects to attempt to comprehend the material, but the critical sentences were sufficiently bizarre that it was unlikely that subjects' background knowledge would make a particular quantifier seem especially probable or improbable in a given sentence. The quantifiers all appeared in the second and third of four paragraphs so as to minimize effects of serial position and recall from immediate memory.

The story was tape-recorded and played to groups of subjects. Half of the subjects were given intentional memory instructions. They were told to listen to the story carefully as they would later be tested on their memory for its content. The other half of the subjects were given incidental instructions. They were told to listen to the story and to prepare to answer questions about how boring or interesting it was. None of the subjects was told that the quantifiers were especially important. After they had listened to the tape, all subjects received a recognition test. The 10 critical sentences which had contained the quantifiers were presented in random order on a sheet of paper, with the quantifier deleted. For each sentence subjects rated all five quantifiers as to how certain they were that each had appeared in the sentence. A rating of 6 indicated certainty that the quantifier had appeared in the sentence, and a rating of 1 indicated certainty that it had not appeared. A total of 100 Rutgers undergraduates served in the experiment in order to satisfy a course requirement.

Quantifier difference ratings. In order to obtain an independent measure of the differences among the quantities expressed by the five quantifiers, 25 University of Michigan students participated in a rating task. The 10 possible pairs of quantifiers were presented in random order on a sheet of paper. All quantifiers were presented in the same phrase (e.g., "*all* of the horses"), with the quantifier underlined. Instructions included the following: "If you hear the sentence '*Many of the horses were brown*', the word *many* gives you some notion of the number of horses that were brown. Different words convey a different idea of the quantity involved (e.g., *many of the horses* vs. *a few of the horses*). For each of the pairs below, please write down a number from 1 to 9 reflecting how big a difference there is between the quantities suggested by the two underlined words." A rating of 1 indicated that the difference was minimal, whereas a rating of 9 indicated that the difference was as large as possible.

RESULTS

Certainty Ratings

For each quantifier that was presented, the mean certainty rating for each of the five possible responses was calculated. The resulting 5×5 matrix of values is presented in Table 2, both for all subjects together and separately for the two instruction conditions. The left–right diagonal entries are the ratings for the correct responses. There are clear differences in the certainty ratings for incorrect responses, contrary to the all-or-none hypothesis. In fact, with only minor exceptions, the rows of values in Table 2 are single-peaked functions, with certainty ratings decreasing with distance from the correct response. The quantifier *none* appears to be the least confusable; when it is presented, the correct response produced the highest certainty rating in Table 2, whereas the incorrect responses received the lowest ratings. Also, several asymmetries are apparent. The certainty rating for the response *some*, when each of the other four quantifiers was presented, was greater than the certainty rating for the other quantifiers when *some* was presented. The semantic overlap hypothesis

TABLE 2

MEAN CERTAINTY RATINGS FOR QUANTIFIERS[a]

Quantifier presented	Response				
	All	Many	Some	A few	None
All subjects combined					
All	3.70	3.06	2.90	2.26	1.41
Many	2.76	3.35	3.32	2.59	1.48
Some	2.38	3.08	3.62	2.72	1.59
A few	2.31	2.82	3.68	3.06	1.55
None	1.72	1.84	2.22	2.10	4.27
Intentional instructions					
All	3.69	2.89	2.78	2.31	1.55
Many	2.53	3.62	3.52	2.87	1.44
Some	2.36	3.28	3.85	2.92	1.71
A few	2.43	2.93	3.67	3.16	1.65
None	1.68	2.03	2.25	2.24	4.26
Incidental instructions					
All	3.71	3.23	3.02	2.20	1.27
Many	2.98	3.07	3.11	2.31	1.51
Some	2.40	2.88	3.39	2.51	1.46
A few	2.19	2.70	3.68	2.95	1.44
None	1.75	1.65	2.18	1.95	4.28

[a] Note: A rating of 6 indicates certainty that the quantifier was presented.

predicts such an asymmetry for *some* vs. *many*, *a few*, and *none*, but not necessarily for *some* vs. *all*. The fact that the asymmetries were obtained in each case suggests that subjects may have responded with a response bias favoring *some*. Note that when *a few* was presented, *some* actually obtained a higher certainty rating than the correct response. The pattern of certainty ratings did not differ between the intentional and incidental learning conditions, $F(16, 1440) = 1.26, p > .20$.

Confusion Frequencies

It is possible that the single-peaked functions in Table 2 do not reflect genuine recognition confusions, but rather a simple response strategy. Subjects had to rate each quantifier as a response to each sentence. They therefore may have often given their highest rating to the correct response and then assigned ratings to the remaining alternatives in proportion to their subjective distances from the correct choice. If the pattern apparent in Table 2 resulted from this strategy, an analysis of first choices might still support the all-or-none hypothesis. In addition, more powerful data analysis techniques are available for analyzing confusion frequencies than certainty ratings. Accordingly, first choice frequencies were derived and used for the major data analyses.

Table 3 gives the frequencies with which the various responses were given as the first choice to each of the presented quantifiers. The data in Table 3 are presented separately for the intentional and incidental conditions; the confusion matrix obtained by collapsing over both instructional groups is given at the top of Table 4. The first choice frequencies were obtained by selecting the response with the highest certainty rating for each test sentence; ties (18.2% of the cases) were broken ran-

TABLE 3

QUANTIFIER CONFUSION MATRICES (FIRST CHOICES) FOR INTENTIONAL AND INCIDENTAL CONDITIONS[a]

Quantifier presented	Response				
	All	Many	Some	A few	None
Intentional instructions					
All	43	22	19	10	6
Many	17	30	29	22	2
Some	11	26	39	17	7
A few	15	16	36	30	3
None	4	10	14	13	59
Incidental instructions					
All	46	16	28	8	2
Many	28	23	28	13	8
Some	17	23	37	17	6
A few	14	19	34	28	5
None	7	6	15	10	62

[a] Note: Row frequencies sum to 100 responses.

TABLE 4

OBTAINED AND PREDICTED CONFUSION MATRIX (ALL SUBJECTS)

Quantifier presented	Response				
	All	Many	Some	A few	None
Obtained[a]					
All	89	38	47	18	8
Many	45	53	57	35	10
Some	28	49	76	34	13
A few	29	35	70	58	8
None	11	16	29	23	121
Predicted by successive interval scaling					
All	88.6	41.4	41.8	20.4	7.8
Many	46.4	47.8	62.6	33.2	10.0
Some	29.8	46.4	67.2	40.2	10.4
A few	24.2	42.8	65.2	45.2	12.6
None	13.8	11.6	23.6	31.8	119.2
Predicted by Luce choice model					
All	89.0	39.0	43.8	22.3	6.0
Many	41.4	55.7	62.6	31.9	8.5
Some	31.7	42.6	76.4	38.9	10.4
A few	28.0	37.7	67.6	52.6	14.1
None	11.9	16.0	28.7	22.3	121.0

[a] Note: Row frequencies sum to 200 responses.

TABLE 5

SECOND CHOICE RESPONSES AS A FUNCTION OF THE FIRST CHOICE

First choice	Second choice				
	All	Many	Some	A few	None
All	—	101	36	33	32
Many	56	—	87	37	11
Some	33	118	—	103	25
A few	19	37	95	—	17
None	31	27	46	56	—

domly. The pattern apparent in the frequency data is extremely similar to that of the certainty ratings (Table 2), providing further support for the linear order hypothesis. The only deviations from single-peaked functions involve especially high frequencies of *some* responses, reflecting an apparent response bias. When *many* or *a few* was presented, *some* was the most frequent first choice as a response. The quantifier *none* produced the least number of confusions. As in the rating data, the two learning instructions produced no significant differences, either in the number of correct responses or in the pattern of confusions, $F < 1$. The frequency matrix for high confidence first choices (ratings of 5 and 6) was also examined. The pattern of confusions among these high confidence choices was very similar to the pattern found using all of the responses.

Table 5 presents the frequencies of the second choice responses (i.e., the responses with the second highest certainty ratings) as a function of the first choice (collapsing over which quantifier had actually been presented). There were a large number of ties for second choice (61.1% of the cases) which were broken randomly. The frequencies in Table 5 should therefore be interpreted with caution. However, a clear pattern is apparent. Second choice frequencies consistently decline monotonically as distance from the first choice increases, providing additional support for the view that quantifier similarities are based on a linear ordering. The second choice data were

also broken down according to which response was correct; however, the resulting cell frequencies were generally too small to be reliable.

Derivation of Quantifier Scale Values

The linear order hypothesis predicts that it should be possible to derive a unidimensional scale of quantifier values that can account for the obtained pattern of recognition confusions. Two methods of deriving such a scale from confusion matrices were attempted. The first method, successive interval scaling (Bock & Jones, 1968), provides a simultaneous normalization of the distribution of responses to each stimulus. The five response alternatives were treated as a category scale, and parameters for the means and variances of the five stimulus words and for the boundaries between response categories were estimated by a least squares technique. The estimated stimulus means provide an interval scale for the quantifiers. The obtained quantifier scale, normalized between 0 and 1, is given in the first row of Table 6. By far the largest distance between adjacent terms lies between *a few* and *none*, reflecting the small number of confusions obtained when the stimulus word was *none*. The next largest gap is between *all* and *many*. The three middle terms, *many*, *some*, and *a few*, are clustered very close together, with a slightly larger distance between *many* and *some* than between *some* and *a few*. In fact, the means for these three quantifiers all fell within the boundaries of the response category *some*.

TABLE 6

QUANTIFIER SCALE VALUES DERIVED BY THREE TECHNIQUES

	All	Many	Some	A few	None
Successive interval scaling (confusion data)	0	.207	.281	.320	1.000
Luce choice model (confusion data)	0	.224	.316	.401	1.000
Pair ratings of quantity differences	0	.206	.522	.717	1.000

The confusion frequencies predicted by the successive intervals model are presented in the middle of Table 4. The model provides a reasonable description of the data, accounting for 97% of the variance; however, the residual deviations are significant, with a χ^2 value of 19.4 for eight degrees of freedom, $p < .05$.

The second method used to derive a quantifier scale was based on the Luce (1959, 1963) biased choice model. The Luce model separates parameters representing the subjects' bias to make response j, regardless of the stimulus presented (β_j), and parameters representing the similarity between each pair of stimuli (η_{ij}). It is assumed that the probability of response j to stimulus i is proportional to the relative magnitude of the product of the corresponding bias and similarity parameters:

$$p_{ij} = \frac{\beta_j \eta_{ij}}{\sum_j \beta_j \eta_{ij}} .$$

The model further assumes that similarities are symmetric $(\eta_{ij} = \eta_{ji})$, and that the similarity of a stimulus to itself is equal to one $(\eta_{ii} = 1)$. The bias and similarity parameters for a confusion matrix can be estimated by maximum likelihood procedures (Bishop, Fienberg, & Holland, 1975, p. 289).

In their general form the Luce similarity parameters need not yield a unidimensional or any other distance structure. The parameters η_{ij} for each stimulus–response pair are entirely independent. One might expect, however, that high similarity values should reflect small psychological distances, whereas low similarity values should reflect large distances. Luce (1963) suggests that a similarity value can be mapped onto a distance measure by the transformation,

$$d_{ij} = -\log \eta_{ij}.$$

If the pair similarities depend on an underlying unidimensional scale of stimulus distances, it follows that the similarity values for pairs of ordered stimuli should combine multiplicatively:

$$\eta_{ik} = \eta_{ij} \times \eta_{jk}.$$

The above constraint on the similarity parameters was therefore imposed in analyzing the quantifier confusion matrix. This constraint reduced the number of similarity parameters to be estimated from 10 (all pairs ij of quantifiers, $i \neq j$) to 4 (only adjacent pairs). With the addition of four independent bias parameters, a total of eight parameters was used to fit the choice model to the data.

The response frequencies predicted by this linear version of the Luce model are presented in the bottom of Table 4. The model provided an excellent description of the data, accounting for 99% of the variance. The residual variance was not significant, with a χ^2 value of 9.91 for 12 degrees of freedom, $p > .50$. The fit of the linear model was not significantly different from that of the general Luce model, which required six additional similarity parameters. In a further analysis the linear model, without any changes in the parameters, was fitted to the

two confusion matrices produced by the intentional and incidental learning groups (Table 3). The model accounted for 96% of the variance among the 50 cell frequencies, and the residual variance was again nonsignificant, with a χ^2 of 30.3 for 32 degrees of freedom, $p > .25$. This result confirms that the two instruction groups produced essentially identical confusion patterns.

The scale of quantifier distances derived from the Luce model is given in the middle row of Table 6. This scale is very similar to that derived by successive interval scaling. The largest separation is between *none* and all the other quantifiers, followed by the distance between *all* and *many*. The three middle quantifiers again form a relatively compact cluster. The relative response biases for the five quantifiers, ordered from *all* to *none*, were .244, .187, .289, .115, and .166 (normalized to sum to 1). The biases tended to favor responses of *some* and also *all* and to weigh against the response of *a few*. The fit of the model significantly decreased when the null hypothesis of equal β values was tested.

For comparison purposes, a quantifier scale was also derived from the ratings of quantity differences between the 10 possible word pairs obtained from an independent group of subjects. Since the number of pairs was too small to constrain adequately a solution derived by purely ordinal methods, the interval properties of the ratings were used to construct a scale, and the rank order of the ratings was used as a check on the solution. Since it was unlikely that subjects would assign equal intervals to the nine-point response scale, the response frequencies were submitted to successive interval scaling. This had the effect of "stretching" the end categories relative to those in the middle. However, the rank order of the 10 values of pair differences obtained from successive interval scaling remained identical to that of the mean ratings. These 10 difference values, d_{ij}, ranged from 0.00 for the pair *some–a few* to 6.60 for the pair *all–none*. The values were then entered below the left–right

diagonal in a 5×5 matrix of quantifier pairs. The diagonal entries, d_{ij}, were set to zero. Since all differences between nonidentical quantifiers should presumably be greater than the zero values assigned to pairs d_{ii}, a constant value (a somewhat arbitrary 2.60) was added to the 10 d_{ij} values for which $i \neq j$. Finally, each of the 10 empty cells d_{ji} in the matrix was set equal to $-d_{ij}$. The column sums were then used as interval scale values for the five quantifiers. This technique (suggested by Elliot Noma) is essentially the same as the least squares solution for paired comparisons outlined by Mosteller (1951). The obtained unidimensional scale preserved the original rank order of the 10 distance ratings quite well, with just one reversal of two differences adjacent in the ordering. In particular, the scaled distances agree with the rank order of the rated differences for the four adjacent pairs.

The quantifier scale based on the rating data is given in the third row of Table 6, again normalized between 0 and 1. The order of the five terms of course agrees with the order derived from the confusion data; however, the relative interval sizes are considerably different. In particular, the end terms (especially *none*) are relatively close to the three middle terms in the scale based on rated differences. Also the distance *many–some* is larger than the distance *some–a few* in the third scale; these two distances are about equal in the Luce scale, although there is a small difference in the same direction in the successive intervals scale. On the other hand, in all three scales, the distance *a few–none* is greater than the distance *all–many*, and *some* and *a few* form the closest adjacent pair.

Formalization of the Semantic Overlap Hypothesis

The linear Luce model, which assumes that pairwise similarities are symmetrical, was able to account for the obtained confusion matrix extremely accurately. However, it is possible that the response bias parameters of

the Luce model masked the kind of asymmetries predicted by the semantic overlap hypothesis. It was therefore desirable to develop and test a formal model of confusion frequencies based on the semantic overlap hypothesis. This was done using a type of log-linear model (Bishop et al., 1975).[2] Log-linear models, of which the Luce choice model is a special case, are so called because the logarithm of the predicted probability of response j to stimulus i (log p_{ij}) is a linear combination of log parameter values. The semantic feature models tested here take the form,

$$p_{ij} = \frac{\eta_{ij}}{\sum_j \eta_{ij}}.$$

These models differ from the Luce model in two respects: (1) Bias parameters are eliminated (at least initially); and (2) the restriction that similarities be symmetrical ($\eta_{ij} = \eta_{ji}$) is no longer maintained. Removing the symmetry restriction presents a problem in constraining the number of required similarity parameters, since any confusion matrix could be fit perfectly by a vacuous model that required a separate parameter for each stimulus–response combination. It is therefore necessary to derive the η_{ij} values from a smaller set of more elementary parameters.

The underlying assumption of the semantic overlap hypothesis is that recognition confusions result when the semantic features of the stimulus word are altered in the direction of an alternative response. That is, features may be lost from or added to the representation of the stimulus, so that it comes to resemble an incorrect response choice. It therefore seems reasonable to assume that the value of η_{ij} should be directly proportional to the ease of transforming stimulus$_i$ into response$_j$. The ease of a particular stimulus-to-response transformation should in turn be related to the ease of the necessary com-

ponent feature changes. The models tested here assume that each η_{ij} is equal to the product of parameters (γ_k) reflecting the ease of these feature changes, i.e., in log-linear form,

$$\log \eta_{ij} = \sum_k \log \gamma_k,$$

where k feature changes are required to transform stimulus$_i$ into response$_j$.

Table 7 presents the derivation of the simplest feature model that was able to provide a reasonable description of the obtained confusion matrix. The semantic representation of the quantifiers as outlined in Table 1 contains three pairs of contrastive features: "positive" vs. "negative"; "existential" (exi) vs. "universal" (uni); and "large" vs. "small." In Table 7 each quantifier is represented as a feature list (e.g., *a few* contains the features "pos," "exi," and "small"). Only one addition to the Table 1 analysis improved the fit of the model: It is assumed that people infer the feature "large" for *all* and add it to the memory representation for that quantifier. This feature is bracketed in Table 7 to indicate that it is present in the representation of *all*, but need not be added to the representation of another quantifier in order to transform the latter into *all* (i.e., the features "pos" and "uni" are sufficient to define *all*, and "large" is redundant).

Five feature changes (represented by parameters γ_1 to γ_5) were sufficient to represent the transformations required to show the relations among the feature descriptions of the quantifiers: "exi" can be changed to "uni" or vice versa (γ_1); "large" or "small" can be lost (γ_2) or added (γ_4); and "neg" can be lost (γ_3) or added (γ_5). As in the negation models of Clark and Chase (1972) and others, "pos" is assumed to be a default feature present when the representation does not contain "neg." The model in Table 7 also assumes that the ease of each feature change is independent of the particular word that contains the relevant feature (e.g., the feature "exi" can be changed as easily in *many* as in *some*). The derivation

TABLE 7

THE SEMANTIC OVERLAP HYPOTHESIS AS A LOG-LINEAR MODEL

Features defining each quantifier	Feature change parameters	Estimated parameter values
All = pos, uni, (large)	γ_1 = Change exi to uni (or vice versa)	0.467
Many = pos, exi, large	γ_2 = Lose large or small	1.024
Some = pos, exi	γ_3 = Lose neg	0.250
A few = pos, exi, small	γ_4 = Add large or small	0.565
None = neg, exi	γ_5 = Add neg	0.164

Quantifier presented	Response				
	All	Many	Some	A few	None
Derivation of η_{ij} for each stimulus–response combination					
All	1	γ_1	$\gamma_1\gamma_2$	$\gamma_1\gamma_2\gamma_4$	$\gamma_1\gamma_2\gamma_5$
Many	γ_1	1	γ_2	$\gamma_2\gamma_4$	$\gamma_2\gamma_5$
Some	γ_1	γ_4	1	γ_4	γ_5
A few	$\gamma_1\gamma_2$	$\gamma_2\gamma_4$	γ_2	1	$\gamma_2\gamma_5$
None	$\gamma_1\gamma_3$	$\gamma_3\gamma_4$	γ_3	$\gamma_3\gamma_4$	1
Predicted confusion matrix[a]					
All	85.8	41.6	42.0	23.8	6.9
Many	30.0	61.8	62.6	35.4	10.2
Some	34.9	40.7	72.0	40.7	11.8
A few	30.3	35.3	62.5	61.8	10.1
None	14.6	17.1	30.2	17.1	121.0

[a] Note: Obtained values are presented at the top of Table 4.

of the η_{ij} values for the model is presented in the middle of Table 7. As in the Luce model, the diagonal elements, η_{ii}, are set equal to 1. Each of the other η_{ij} values is defined as the product of the component feature changes required to convert the stimulus into the response. For example, transforming *all* into *none* requires changing "uni" to "exi" (γ_1), losing "large" (γ_2), and adding "neg" (γ_5).

The confusion frequencies predicted by this semantic feature model are presented at the bottom of Table 7. The overall fit of the model, based on just five parameters, is quite good, accounting for 96% of the variance with a χ^2 of 20.2 for 15 degrees of freedom, $p > .25$. The estimated parameter values, given in the top right corner of Table 7, are readily interpretable in the framework of the semantic overlap

hypothesis. Higher parameter values indicate greater ease of a feature change. An orderly pattern of asymmetries is apparent: In each case the estimated parameters indicate that it is easier to lose a particular feature than to add it. The negation feature is least likely to be lost or added, whereas the quantity features are most likely to be altered.

However, there is one substantial error in the predictions of the feature model: An under-prediction of the number of *all* responses to the stimulus *many* (30 predicted vs. 45 obtained). This deviation is significant ($p < .025$) by a Freeman–Tukey test (Bishop et al., 1975, p. 137). This shortcoming of the model is particularly important in that the pair *all–many* gives rise to the major distinction between the semantic feature and linear order

analyses of the quantifiers. In terms of the semantic analysis in Table 1, *all* and *many* are separated by a major feature division (universal vs. existential), but, in spite of this, they are relatively often confused with each other. Without the slightly ad hoc device of adding the feature "large" to *all*, the feature model also significantly underpredicts the number of *many* responses to *all*. It appears difficult to find a principled way to modify the feature model so as to reduce the error in its prediction for the *many–all* cell. Several possibilities were tested. For example, the number of parameters was increased to eight by "splitting" each of γ_1, γ_2, and γ_4 into two parameters (i.e., allowing γ_1 to vary with direction of change, and γ_2 and γ_4 to differ for "large" vs. "small"). Another approach was to augment the feature model with bias parameters so that the predicted and obtained column sums were equalized. However, none of these modifications significantly improved the overall fit of the model, nor did any eliminate the significant error in the *many–all* cell.

DISCUSSION

The results of the present study indicate that the pattern of recognition confusions among the five quantifiers can be related to a unidimensional scale of quantifier values. This is true both for subjects told to expect a memory test and for subjects given only incidental instructions. The results clearly reject the all-or-none hypothesis: Confusion frequencies are in fact highly sensitive to differences in the similarity of quantifier pairs. The semantic overlap hypothesis, explicitly formulated as a log-linear model, provided a reasonable overall account of the data. However, the feature model encountered difficulty in explaining the high frequency of confusions between *all* and *many*, which represent a major feature division among the quantifiers. The linear order hypothesis easily

accounts for the confusability of this pair, since these terms are adjacent in the quantifier ordering.

The linear order hypothesis is strongly supported by the existence of a unidimensional quantifier scale that describes the obtained confusion pattern. The only apparent problem for this hypothesis is that, while the scales derived from the confusion data order the stimuli in the same way as a scale of subjective quantity, the metric properties of the two types of scales appear to differ (Table 6). While the linear order hypothesis does not demand that the memory scale for quantifiers should directly reflect the corresponding quantity scale, this would have been an attractively simple result.

However, it is possible that the discrepancies between the metric properties of the two scale types are less troublesome than they at first appear. The major difference in the scales is the fact that the two end terms, *all* and *none*, are relatively further from the middle terms in the memory scales than in the quantity scale. It may be that subjects in the rating task tended to consider just the ordinal relations among the five terms. This could have made the intervals appear more equally spaced than they really are.

It is also possible that the large distances between the end terms and the middle ones in the memory scales actually reflect differences in the variability of the memory representations of the quantifiers. If the quantifiers are represented in memory as distributions on a unidimensional quantity scale, then by Thurstonian assumptions the confusability of two terms will vary inversely with the difference between their means and directly with the size of their variances. The derived scale of quantity differences does not indicate the variability of the scale values; but, intuitively, the quantities expressed by *all* and *none* seem less variable than those expressed by the three middle terms. It is possible that the large distances at the ends of the scales derived from successive interval scaling and the choice

model in part reflect the relatively low variability of the memory traces for *all* and *none*.[3]

Finally, it is by no means necessary to view the semantic overlap and linear order hypotheses as mutually exclusive alternatives. It is possible that the memory representation for a quantifier contains some information about semantic features in addition to an essentially analog representation of quantity. In particular, the negative component in *none* may be maintained in the memory trace with a high probability, serving to distinguish the correct response from the other alternatives on the recognition test. Even if the linear order hypothesis is accepted, it would not imply that a semantic analysis of the sort presented in Table 1 is incorrect or lacking in "psychological reality." Some such analysis appears to be needed in any case to account for the complex inferences people can make on the basis of quantified sentences.

The present results certainly do not rule out the semantic overlap hypothesis for quantifiers, even though its prediction was substantially in error for one cell of the confusion matrix. The predictions of the semantic overlap and linear order models are simply too highly correlated to be definitively distinguished for the terms that were investigated. In addition, it is always possible that

---

[3] The Luce similarity parameters do not separate effects of mean distance versus variability. The successive intervals model provided estimates of the sigmas for the five quantifiers, but these were actually largest for the end terms, reflecting the fact that a significant number of confusions were made over the entire range of quantifiers, even between *all* and *none*. However, this assumes that a normal distribution underlies the responses; the true distribution may have a sharper "peak" than the normal. Also, neither of the models has a provision for pure guessing, which must surely have occurred to some extent. The remote confusions between *all* and *none* may simply reflect guessing in the absence of any information, rather than true confusions based on highly variable memory representations. As an aside, it should be noted that none of the models employed here includes parameters for pure guessing. It may be necessary to introduce such parameters in further work on memory confusion matrices (see Pachella et al., 1977).

some alternative semantic feature model could account for the data more adequately than the models tested here. For example, if *many* were defined in terms of *all* instead of *some* (e.g., "all with a few exceptions"), this would presumably predict more confusions between *many* and *all* as required by the data (although other problems might be created). The most important aspect of the present study is not any strong substantive conclusion, but rather its demonstration that the general class of log-linear models can be used to derive detailed quantitative predictions about recognition confusions from hypotheses that are conceptually very different. Such fine-grained analyses of confusion data may provide valuable information about the way word meanings are remembered. In particular, the kinds of models discussed in the present paper can be tested for other sets of terms, drawn from linear orderings and also other types of semantic fields.

While the present results are not conclusive, they do support the possibility that a major component of the memory trace for quantifiers corresponds to a value on an internal quantity scale. In one sense this possibility seems almost trivial—after all, it is only common sense that what people should understand from a quantifier is some notion of quantity. But in another sense such a conclusion would have important implications for memory models. It is certainly not logically necessary that the memory trace for quantifiers should resemble an analog to subjective quantity as closely as it appears to do so. The memory representation might well have proved to be much more abstract. Indeed, none of the propositional models of language now current in psychology explicitly includes continuous magnitude values in the semantic representation of "logical" words such as quantifiers. However, the present results, together with other recent work (Holyoak & Walker, 1976; Moyer, 1973; Walker, 1975) and earlier findings (Cliff, 1959), suggest that magnitude information must be incorporated into any psychological theory of meaning.

REFERENCES

ANDERSON, J. R. *Language, memory and thought.* Hillsdale, New Jersey: Erlbaum, 1976.

ANDERSON, J. R., & BOWER, G. *Human associative memory.* Washington, D.C.: V. E. Winston & Sons, 1973.

BISHOP, Y. M. M., FIENBERG, S. E., & HOLLAND, P. W. *Discrete multivariate analysis: Theory and practice.* Cambridge: MIT Press, 1975.

BOCK, R. D., & JONES, L. V. *The measurement and prediction of judgment and choice.* San Francisco: Holden–Day, 1968.

BRANSFORD, J. D., BARCLAY, J. R., & FRANKS, J. J. Sentence memory: A constructive vs. interpretive approach. *Cognitive Psychology,* 1972, **3,** 193–209.

CLARK, H. H., & CARD, S. K. The role of semantics in remembering comparative sentences. *Journal of Experimental Psychology,* 1969, **82,** 545–552.

CLARK, H. H., & CHASE, W. G. On the process of comparing sentences against pictures. *Cognitive Psychology,* 1972, **3,** 472–517.

CLARK, H. H., & CLARK, E. V. Semantic distinctions and memory for complex sentences. *Quarterly Journal of Experimental Psychology,* 1968, **20,** 129–138.

CLARK, H. H., & STAFFORD, R. A. Memory for semantic features in the verb. *Journal of Experimental Psychology,* 1969, **80,** 326–334.

CLIFF, N. Adverbs as multipliers. *Psychological Review,* 1959, **66,** 27–44.

CONRAD, R. Acoustic confusions in immediate memory. *British Journal of Psychology,* 1964, **55,** 75–84.

ERICKSON, J. R. A set analysis of behavior in formal syllogistic reasoning tasks. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola Symposium.* Hillsdale, New Jersey: Erlbaum, 1974.

FILLENBAUM, S. Memory for gist: Some relevant variables. *Language and Speech,* 1966, **9,** 217–227.

GLASS, A. L., & HOLYOAK, K. J. The effect of *some* and *all* on reaction time for semantic decisions. *Memory and Cognition,* 1974, **2,** 436–440.

GLASS, A. L., HOLYOAK, K. J., & O'DELL, C. Production frequency and the verification of quantified statements. *Journal of Verbal Learning and Verbal Behavior,* 1974, **13,** 237–254.

HOLYOAK, K. J. The form of analog size information in memory. *Cognitive Psychology,* 1977, **9,** 31–51.

HOLYOAK, K. J. Comparative judgments with numerical reference points. *Cognitive Psychology,* 1978, **10,** 203–243.

HOLYOAK, K. J. & GLASS, A. L. The role of contradictions and counter-examples in the rejection of false sentences. *Journal of Verbal Learning and Verbal Behavior,* 1975, **14,** 215–239.

HOLYOAK, K. J., & WALKER, J. H. Subjective magnitude information in semantic orderings. *Journal of Verbal Learning and Verbal Behavior,* 1976, **15,** 287–299.

JUST, M. A., & CARPENTER, P. A. Comprehension of negation with quantification. *Journal of Verbal Learning and Verbal Behavior,* 1971, **10,** 244–253.

KINTSCH, W. *The representation of meaning in memory.* Hillsdale, New Jersey: Erlbaum, 1974.

LEECH, G. *Semantics.* Harmondsworth: Penguin Books, 1974.

LEECH, G. N. *Towards a semantic description of English.* Bloomington, Indiana: Indiana University Press, 1969.

LUCE, R. D. *Individual choice behavior.* New York: Wiley, 1959.

LUCE, R. D. Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology,* Vol. I. New York: Wiley, 1963.

MEYER, D. E. On the representation and retrieval of stored semantic information. *Cognitive Psychology,* 1970, **1,** 242–300.

MOSTELLER, F. Remarks on the method of paired comparisons. I. The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika,* 1951, **16,** 3–11.

MOYER, R. S. Comparing objects in memory: Evidence suggesting an internal psychophysics. *Perception and Psychophysics,* 1973, **13,** 180–184.

MOYER, R. S., & LANDAUER, T. K. Time required for judgments of numerical inequality. *Nature (London),* 1967, **215,** 1519–1520.

PACHELLA, R. G., SMITH, J. E. K., & STANOVICH, K. E. Qualitative error analysis and speeded classification. In J. Castellan & F. Restle (Eds.), *Cognitive theory,* Vol. III. Hillsdale, New Jersey: Erlbaum, 1977.

PAIVIO, A. Perceptual comparisons through the mind's eye. *Memory and Cognition,* 1975, **3,** 635–647.

POTTS, G. R. Information processing strategies used in the encoding of linear orderings. *Journal of Verbal Learning and Verbal Behavior,* 1972, **11,** 727–740.

POTTS, G. R. Storing and retrieving information about ordered relationships. *Journal of Experimental Psychology,* 1974, **103,** 431–439.

REVLIS, R. Syllogistic reasoning: Logical decisions from a complex data base. In R. J. Falmagne (Ed.), *Reasoning: Representation and process.* Hillsdale, New Jersey: Erlbaum, 1975.

SACHS, J. S. Recognition memory for syntactic and semantic aspects of connected discourse. *Perception and Psychophysics,* 1967, **2,** 437–442.

TRABASSO, T., & RILEY, C. A. On the construction and

use of representations involving linear order. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola Symposium.* Hillsdale, New Jersey: Erlbaum, 1975.

WALKER, J. H. Real-world variability, reasonableness judgments, and memory representations for concepts. *Journal of Verbal Learning and Verbal Behavior,* 1975, **14**, 241–252.

WASON, P. C., & JOHNSON-LAIRD, P. N. *Psychology of reasoning: Structure and content.* Cambridge: Harvard University Press, 1972.

WOOCHER, F. D., GLASS, A. L., & HOLYOAK, K. J. Positional discriminability in linear orderings. *Memory and Cognition,* 1978, **6**, 165–173.