JOURNAL OF VERBAL LEARNING AND VERBAL BEHAVIOR 14, 215–239 (1975)

The Role of Contradictions and Counterexamples in the Rejection of False Sentences[1]

KEITH J. HOLYOAK AND ARNOLD L. GLASS

Stanford University

Two hypotheses about the processes by which people can reject false but meaningful sentences quantified by *all* or *some* are outlined. These hypotheses distinguish between two basic types of false sentences: contradictory sentences (e.g., *All/Some birds are dogs*), which are rejected on the basis of a direct contradiction between the subject and predicate concepts; and counterexample sentences (e.g., *All birds are robins*), which are falsified when the person thinks of a counterexample to the assertion (e.g., *canary*). Experiment I demonstrated that people use contradictions to produce false completions of sentences. In Experiment II, the false production frequency norms obtained in Experiment I, together with the theoretical analysis of false sentences, were used to predict the time required to reject false sentences. The results supported the contradiction and counterexample hypotheses, and indicated that false sentences with subject and predicate words closely related in meaning can sometimes be disconfirmed relatively quickly. Experiment III extended the counterexample hypothesis to sentences containing the verb *have* (e.g., *All buildings have elevators*), and also provided some evidence that the process of exemplar search used to find counterexamples may also sometimes play a role in the verification of true generalizations.

Theories of semantic memory have often had difficulties in explaining how people decided that sentences are false (Collins & Quillian, 1972). The possible mechanisms underlying the rejection of false sentences are the central concern of the present paper. The specific hypotheses that were tested in the experiments below emerge from a theoretical framework common to several recent memory models (e.g., Anderson & Bower, 1973; Collins & Quillian, 1969). The basic processing assumption underlying these models is that facts in memory are accessed in a particular order, and that knowledge of that order will allow one to predict how quickly semantic decisions will be made. Sentence verification is described in terms of a search through a semantic network to find a pathway between the facts associated with the subject and predicate concepts. Since a person cannot know in advance whether a particular sentence will be true or false, any such search model must assume that the order in which information is retrieved will be (at least initially) independent of the truth value of the sentence. This suggests that for both true and false sentences, those sentences with subject and predicate concepts that are closely associated in memory will be verified relatively quickly. But in order to generate more specific experimental predictions, it is necessary to find some measure of the order in which information about specific concepts is retrieved.

One empirical measure of this sort was proposed by Glass, Holyoak, and O'Dell

[1] The ordering of authors is haphazard. We thank Edith Greene and Debbie Weinstein for their help in testing subjects, and Gordon H. Bower, Herbert H. Clark, Edward J. Shoben, Edward E. Smith, and Keith T. Wescourt for their extensive comments on the numerous drafts of this paper. This research was completed while K. Holyoak was supported by a Stanford University Fellowship and NIMH Grant MH20021 to Herbert H. Clark, and A. Glass held an N.S.F. graduate fellowship. It was supported by NIMH Grants MN13950-06 to Gordon H. Bower and MH20021 to Herbert H. Clark.

(1974). They asked subjects to provide true one-word completions for incomplete sentences of the form (*Quantifier*) *S are* ____, and tabulated the frequency with which different words were given as predicates. This constrained association technique is similar to the way one collects production frequency norms from subjects who are asked to produce different instances as responses to a category name (Battig & Montague, 1969). Glass *et al.* assumed that the frequency with which a word appeared as a completion reflected the speed with which that predicate can be accessed from the subject concept. Accordingly, people should be able to verify high-frequency true sentences (e.g., *All roses are flowers*) more quickly than corresponding low-frequency sentences (e.g., *All roses are plants*). And indeed, for five different quantifiers (*all*, *many*, *some*, *few* and *no*), production frequency norms successfully predicted true reaction time (RT). Similarly, a number of investigators have shown that those instances which are most frequently produced as responses to the category name can be most quickly classified as category members (Loftus, 1973; Wilkins, 1971). Other studies have shown that ratings of semantic relatedness (overlap in the meanings of subject and predicate words) can also be used to predict true RT; that is, true sentences with highly-related subject and predicate words are verified more quickly than sentences with low-related subject and predicate words (Rips, Shoben & Smith, 1973; Rosch, 1973). Rated relatedness has typically been highly correlated with production frequency in these studies; and where the effects of the two variables have been separated, production frequency has proved to be the better predictor of RT (Smith, Shoben & Rips, 1974). Consequently, these latter studies are also consistent with the claim that production frequency (hereafter abbreviated as PF) indexes the order in which the semantic information used for "true" decisions is retrieved from memory.

But what kind of information might be used to decide that a sentence is false? Collins and Quillian (1969, 1972) suggested that closely-related false sentences might be rejected relatively quickly, after rapid discovery of a contradiction; for example, *A collie is green* would be rejected on the basis of the information that collies are brown and white. However, their data offered no support for the contradiction hypothesis. Since the procedure by which Collins and Quillian estimated search order (using intuitively-defined logical hierarchies) is in general suspect (Conrad, 1972; Smith *et al.*, 1974), their result is not conclusive. However, later studies have found that high relatedness increases false RT (Schaeffer & Wallace, 1970). Several investigators have found that RT to reject meaningful (high-related) false sentences (e.g., *All stones are gems*) is longer than the RT to reject relatively anomalous sentences (e.g., *All typhoons are wheats*) (Kintsch, 1972; Meyer, 1970; Rips *et al.*, 1973; Wilkins, 1971). The prediction that high semantic relatedness will necessarily slow down negative decisions has been made a central feature of the models proposed by Schaeffer & Wallace (1970) and Smith *et al.* (1974). These latter models assume that verification time is largely determined by holistic comparison procedures, rather than by ordered retrieval of information.

However, only the Glass *et al.* (1974) study has addressed the question of how PF relates to false RT. They generated false but relatively meaningful sentences by substituting the positive quantifier *many* in sentences that were true when quantified by the negative *few*; for example, the true sentence *Few arrows are dull* was converted into the false sentence *Many arrows are dull.* They then found that false sentences with high-PF predicates, as measured by true *few*-completion norms (e.g., *Many arrows are dull*), were rejected *more* quickly than sentences with low-PF predicates (e.g., *Many arrows are wide*). However, minimally-related anomalous sentences (e.g., *Many arrows are intelligent*)

were rejected most rapidly of all. These results are not inconsistent with the earlier findings, since previous studies typically compared a mixture of relatively meaningful sentences, which presumably differed in PF, to anomalous (very low-related) sentences.

The Glass *et al.* results supported the hypothesis that the mechanisms by which meaningful false sentences are disconfirmed involve the discovery of a contradiction. In each case the predicate of a high-PF false sentence was contradicted by the predicate of a high-PF true sentence (e.g., *Many arrows are dull* vs. *sharp*); while the predicate of a low-PF false sentence was contradicted by the predicate of a low-PF true sentence (e.g., *Many arrows are wide* vs. *narrow*). Just as high PF led to relatively rapid rejection of false sentences, it also led to relatively rapid verification of true sentences. This parallel between true and false sentences, both in PF and RT, suggests that the same basic information is used for both positive and negative decisions; for example, the same information that confirms the proposition that *Many arrows are sharp* contradicts the proposition that *Many arrows are dull.*

A major purpose of the present paper was to extend the Glass *et al.* results for false property statements quantified by *many* to category-statements and other quantifiers, particularly *all* and *some*—that is, to the experimental domain with which semantic-memory models have generally been most closely concerned. Our hypothesis was that for meaningful sentences in which the subject and predicate are disjoint concepts (e.g., *All/Some men are women*), *false* PF—that is, the frequency with which people generate *women* as a *false* completion of *All/Some men are* ____—would index the speed with which a contradiction is accessed. Accordingly we again expected to find a parallel between true and false sentences, both in PF and RT. For this example, a person could reject the sentence after discovering that *men* and *women* are disjoint divisions of the category *human.* Since *All men are human* is a high-PF, quickly-verified true sentence, we expected that *All/Some men are women* would be a high-PF, quickly-rejected false sentence.

Note that meaningful false sentences such as the above example, which we will term *contradictory* sentences, are false whether quantified by *all* or by *some.* In contrast, other sentences are false only for *all*, while being true for *some* (e.g., *All birds are canaries*). Unlike the case for contradictory sentences, the latter type of false sentence does not produce a direct contradiction between the subject and predicate concepts. That is, *canary* does not contradict *bird* in the sense that *woman* contradicts *man.* How, then, can a sentence like *All birds are canaries* be rejected? One plausible mechanism might be for the person to access an instance of the concept *bird* which can not be a *canary*—that is, a counterexample, such as *robin.* According to this hypothesis, rejection of these sentences, which we all term *counterexample* sentences, also involves finding a contradiction—not between the subject and predicate concepts, but rather between the predicate and an *exemplar* of the subject. Note that we are not using the term "exemplar" to necessarily refer to an individual instance of the subject category. Throughout this paper we will use this term to refer to a concept at any level of abstraction (e.g., a particular canary, the class "canary," or the still larger class "songbird"), as long as it is necessarily included in the subject category.

The counterexample hypothesis also predicts a parallel between true and false PF and RT, but of a different sort than that predicted for contradictory sentences. If exemplars are accessed in the same order for both true and false sentences, then the frequency with which a predicate is generated as a false counterexample *all*-completion should correspond to the frequency with which it is given as a true *some*-completion. For example, the frequency with which *All birds are canaries* is generated as a false sentence should be positively

correlated with the frequency with which *Some birds are canaries* is generated as a true sentence. But since counterexample sentences do not involve a direct contradiction between the subject and predicate, false PF should not predict the time required to reach a "false" decision. Rather, the important variable should be the time required to access the first counterexample to the sentence; accordingly, the RT to reject *All birds are canaries* should depend on the PF with which the dominant counterexample is used to form a true *some*-statement (e.g., *Some birds are robins*).

These hypotheses were tested in three experiments. Experiment I concerned the collection and analysis of false predicate-completions for sentences quantified by *all* and *some*. In Experiment II the contradiction and counterexample hypotheses were tested in a sentence-verification experiment. Finally, Experiment III examined implications of the counterexample hypothesis for the processing of true sentences.

EXPERIMENT I

Method

Subjects in Experiment I were presented with incomplete quantified statements, and for each sentence were asked to supply nouns and adjectives which would form false completions. Twenty-four different nouns were each used as subject words twice, once in a sentence of the form *All S are* ____, and once in *Some S are* ____. These subject nouns were selected from two item sets for which true-completion norms were already compiled (Glass *et al.*, 1974; Glass & Holyoak, 1974.) The nouns were a subset of those for which a particular true predicate completion had been generated by 25% or more of the subjects in these previous studies. All the subject nouns had Thorndike-Lorge frequencies greater than 20 per million, and rated imagery value greater than 6.0 (on a 7-point scale) in the norms of Paivio, Yuille, and Madigan (1968) (with the exception of two, *fruit* and *gun*, for which imagery values were not available). *All-* and *some*-statements were ordered randomly in separate booklets. Each page of the booklets contained a single sentence, with the headings *Adjectives* and *Nouns* typed below it.

For each incomplete sentence, subjects were given 45 seconds to list under the appropriate headings as many nouns and as many adjectives as they could think of that would complete the sentence to form a false statement. At a signal from the experimenter they turned the page and began work on the next item. They were told that the experimenters were interested in the particular responses subjects gave for each item; apart from this, no attempt was made to elicit semantically-related rather than anomalous predicates. Subjects were shown examples of nouns and adjectives to ensure that they knew the distinction, and were asked to try to provide instances of both lexical classes as completions for each statement. Each subject completed both booklets; across subjects, each booklet-type was administered first and second equally often.

Thirty-two Stanford introductory psychology students participated in the experiment in order to satisfy a course requirement.

Results and Discussion

We will first describe various distributional properties of the false completions generated by subjects, and then examine the implications of the results for the contradiction and counterexample hypotheses. Figure 1 plots the distributions of noun responses, and Figure 2 plots the distribution of adjective responses. In both figures the height of the bars indicates the percentage of the grand total of all responses that fell into each of the indicated response classes. The values given in the two figures sum to 100%. The distributions are plotted separately for different frequency levels (the number of subjects generating tokens of a particular response

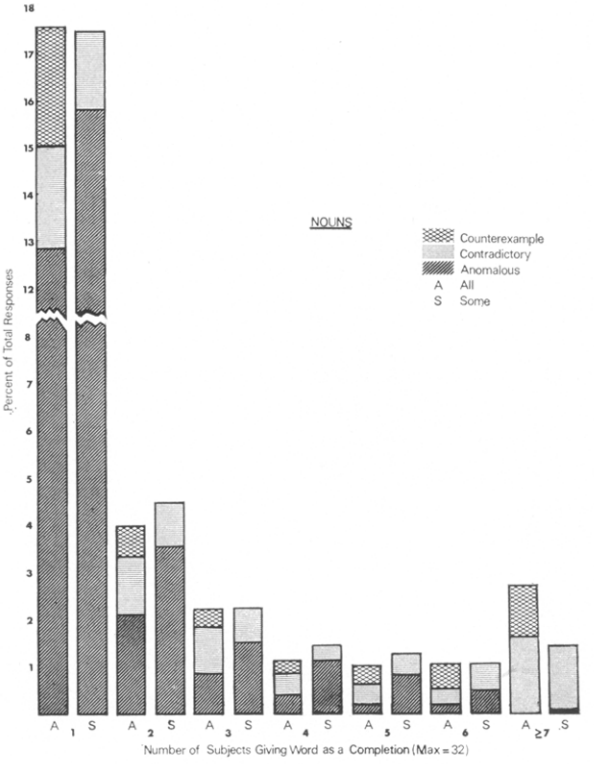

FIG. 1. Distribution of false noun completions.

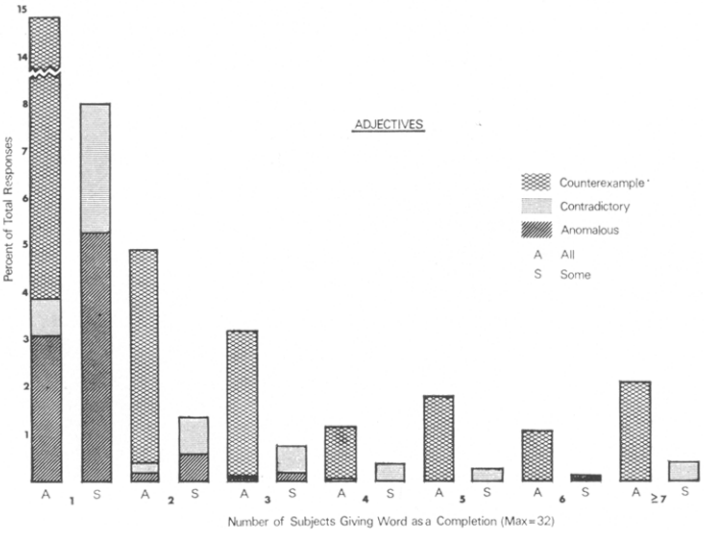

FIG. 2. Distribution of false adjective completions.

word), and for *all* and *some*. Completions that clearly did not form unambiguous false sentences (<3% of the data) were discarded prior to this tabulation. Overall, subjects gave a mean of 7.33 completions for each item. The number of completions for any one item by any one subject ranged from 0 to 27. An analysis of variance examined the number of noun and adjective completions as a function of quantifier (*all* vs. *some*) and booklet order. Both subjects and items were treated as random effects, and quasi *F*-ratios were calculated (Winer, 1971). Subjects provided significantly more *all-* than *some*-statement completions (4.28 vs. 3.05/item), $F'(1, 47) = 66.7, p < .001$. This difference reflects the fact that false *some*-statements logically form a proper subset of false *all*-statements. Overall, more nouns than adjectives were given as responses (4.41 vs. 2.92 per item), $F'(1, 41) = 20.1, p < .001$, though this difference was mainly contained in the *some*-statement data, $F'(1, 46) = 49.0, p < .001$. Neither the main effect nor any interactions involving booklet order (*all-* vs. *some*-statements completed first) achieved statistical significance.

The percentages in Figures 1 and 2 are also given separately for counterexample, contradictory and anomalous responses. The motivation for separating anomalous from contradictory completions was purely empirical—the results of Glass *et al.* (1974) indicated that the relationship between false PF and RT predicted by the contradiction hypothesis holds only for meaningful false sentences. The distinction between contradictory and anomalous predicates was not always clear-cut, and the percentages reported here are only meant to be approximate. Nevertheless, the criteria on which we based the distinction were fairly objective. False noun predicates were counted as contradictory if they were members of a category immediately superordinate to the subject concept (e.g., *All/Some birds are dogs* or *reptiles*); if the intersection between the subject and predicate concepts was any more remote, the completion was classified as anomalous (e.g., *All/Some birds are acorns* or *tables*). Adjective completions were counted as contradictory if they directly contradicted a true property of the subject concept (e.g., *All/Some horses are green*), and as anomalous if they referred to a dimension that was simply inapplicable to the subject concept (e.g., *All/Some horses are readable*). These criteria for separating contradictory and anomalous sentences are essentially the same as those introduced by Kintsch (1972). Evidence that others are sensitive to a distinction between contradictory and anomalous sentences was provided by relatedness ratings, reported in footnote 3 below (Experiment II), which indicated that people rate subject–predicate pairs drawn from anomalous sentences as much less related in meaning than pairs drawn from contradictory sentences. The criterion for classifying counterexamples was straightforward—these were predicates used as *all*-statement completions which, while false for the quantifier *all*, would have been true for *some* (e.g., *All horses are stallions* or *black*).

The frequency with which each word was given as a sentence completion was tabulated, summing across subjects. The number of completions which occurred at different frequency levels, partitioned into contradictory, counterexample, and anomalous responses, is presented graphically in Figures 1 and 2. Two striking results are immediately apparent. The first is simply that certain completions were given by a sizeable minority (20% or more) of the subjects. Given the extremely loose constraints that the false production task imposed on subjects, there was no *a priori* reason to expect any consistency at all in the responses given by different subjects. For any sentence fragment presented (e.g., *All birds are* ____), virtually any word in the English language would produce a false sentence. The second striking fact is that virtually all the high-PF completions were semantically related to the subject noun. While people produced enor-

mous numbers of anomalous responses, the overwhelming majority of these were given by only one or two of the 32 subjects. As Figures 1 and 2 indicate, a much greater percentage of these anomalous sentences were nouns (67%), as opposed to adjectives (24%), suggesting that people produce such completions primarily by retrieving category names without regard for the presented context, testing each word to see if it produces a false sentence, and writing down those that do. But this kind of strategy produced no consistency across subjects in the particular completions used. When the consistent, high-PF completions were examined, these almost all proved to result in meaningful sentences. This result clearly suggests that the high-PF completions were produced by strategies based on the meaning of the quantified subject concept.

Further analyses tested the relationships between true and false completions predicted by the contradiction and counterexample hypotheses. Since few adjectives achieved high PF (i.e., were produced by 20% or more of the subjects), these analyses (and subsequent experiments) dealt only with noun completions. Recall that according to the contradiction hypothesis, contradictory predicates will be generated by accessing a superordinate of the subject concept, and then using a disjoint subset of the superordinate category as a response; for example, the fact that all men are humans can be used to generate the false sentence *All men are women.* Accordingly, there should be a close relationship between the frequency with which *All/Some men are women* is produced as a false sentence, and the frequency with which *All men are humans* is produced as a true sentence. Specifically, we expected that each high-PF true completion of *All S are* ____, from the norms of Glass and Holyoak (1974) and Glass *et al.* (1974), would determine some high-PF contradictory completion of both *All S are* ____ and *Some S are* ____.

As the examples given in the top of Table 1 illustrate, this prediction was confirmed. Fourteen of the 16 highest frequency true *all*-statements, produced by between 35 and 78% of the respondents, corresponded to contradictory sentences produced by from 19 to 56% of respondents. Less than 1% of the false completions fell in this highest frequency range. Other evidence suggests that contradictory completions are generated by the same process, regardless of the quantifier. First, inspection of Figure 1 shows that the absolute number of contradictory noun responses was comparable for the two quantifiers. (This equality did not hold for adjectives, Figure 2, since counterexample adjective completions were apparently generated so easily for *all* that contradictory and anomalous

TABLE 1

EXAMPLES OF RELATIONSHIPS BETWEEN HIGH-FREQUENCY TRUE AND FALSE SENTENCES

True version	False version
Contradictory falses	
All birds are animals	All/Some birds are dogs
All chairs are furniture	All/Some chairs are tables
All women are humans/females	All/Some women are males
All diamonds are stones	All/Some diamonds are emeralds
Counterexample falses	
Some flowers are roses	All flowers are roses
Some prisoners are men	All prisoners are men
Some books are novels	All books are novels
Some teachers are professors	All teachers are professors

responses became relatively infrequent.) Second, we calculated the correlation between the frequency with which particular contradictory responses were used as completions with *all* and with *some*. Since all subjects provided both *all-* and *some*-completions, there was a danger that subjects would sometimes have remembered completions from one booklet to the next. To avoid this difficulty we only considered first-booklet responses. This reduced the number of respondents for each item to 16, so that the frequency estimates used in calculating the correlation were undoubtedly subject to considerable error variance. To avoid having the correlation spuriously inflated by including the large number of responses given just once to one quantifier and never to the other, only completions given at least twice as responses to one of the two quantifiers were included (a total of 65 different nouns). Despite these difficulties, a modest but highly reliable correlation, $r = .48$, was obtained between the frequency of contradictory responses given to *all* and to *some*, $F(1, 63) = 19.2$, $p < .001$. These results are at least consistent with the hypothesis that the same strategy is used to generate contradictory completions for both quantifiers.

According to the counterexample hypothesis, counterexample *all*-statements (e.g., *All birds are canaries*) should be generated on the basis of the corresponding true proposition quantified by *some* (e.g., *Some birds are canaries*). Our prediction, therefore, was that each high-PF true *some*-statement (from the earlier norms of Glass and Holyoak, 1974, and Glass *et al.*, 1974) would correspond to a frequent counterexample *all*-statement in the present experiment when subjects were generating false completions. This prediction was also confirmed. The bottom of Table 1 lists four examples of the 22 most frequent true *some*-statements from earlier norms, given by from 22 to 89% of respondents. Without exception, each of these 22 *some*-statements (e.g., *Some flowers are roses*) corresponded to a counter-example sentence (e.g., *All flowers are roses*) given by from 16 to 53% of respondents. Again, less than 1% of the false completions fell into this highest frequency range.

A final analysis was directed at the relationship between PF and output order. Since we wish to relate PF to order of information retrieval, it was important to demonstrate that completions given with high frequency also tend to be the earliest responses given by subjects. Accordingly, we calculated the mean output position for each completion given by seven or more (22% or more) of the 32 respondents (a total of 39 items) and in each case compared this mean to the mean output rank for all completions of that sentence fragment. The mean rank of the high-PF completions was 1.94, as compared to the overall mean rank of 2.82. This difference was significant by a sign test, $p < .001$. This result extends to false productions the correlation between rank and PF established for true productions by Battig and Montague (1969), Bousfield and Barclay (1950), and Loftus and Scheff (1971). Further, it has long been known that PF is highly correlated with speed of generation in a free or constrained word association task (Woodworth & Schlosberg, 1950, pp. 61–66). These results support the hypothesis that PF is an index of the order in which information is accessed in memory.

Experiment II

The contradiction and counterexample hypotheses predict that disconfirmation of meaningful false sentences (i.e., contradictory and counterexample sentences) requires discovery of a contradiction. Assuming that search is self-terminating, it follows that the sooner the person can access a fact which brings out a contradiction between the subject and predicate, the quicker such sentences will be rejected. For both contradictory and counterexample sentences, the order in which contradictions are discovered should be

predicted by PF norms; however, the variable which determines false RT should be quite different for these two kinds of false sentences.

Contradictory sentences (e.g., *All/Some birds are dogs*) contain predicates that directly contradict the subject. For such a sentence, its false PF was taken as an index of the speed of accessing the contradiction between the subject and predicate concepts. For contradictory sentences, then, false sentences that have high frequency in the norms should be rejected more quickly than false sentences given with low frequency in the norms. In contrast, for counterexample sentences (e.g., *All birds are canaries*), the predicate does not directly contradict the subject. In order to reject this type of sentence, the person must discover some subset or exemplar of the subject concept (e.g., *robin*) that contradicts the predicate. Therefore, the RT should be fastest for those sentences for which a disconfirming counterexample was produced most frequently as a true *some*-completion. Specifically, the RT to reject a counterexample sentence such as *All birds are canaries* should be faster the higher the frequency with which the most common counterexample (e g., *robin*) was given as a true completion of the sentence *Some birds are* ____.[2] Since there is no direct contradiction between the subject and predicate for this type of sentence, the PF of the sentence itself (*bird* to *canary*) should have no appreciable effect upon the time to reject it.

Predictions of the Feature-Comparison Model

The predictions of the contradiction and counterexample hypotheses can be contrasted with those derived from the two-stage feature-comparison model proposed by Smith *et al.* (1974). The predictions of the feature-comparison model are based on ratings of the semantic relatedness of subject and predicate words, rather than on frequency measures. Numerous investigators have found relatedness and production frequency to be positively correlated for true sentences (Rips *et al.*, 1973; Rosch, 1973; Smith *et al.*, 1974). However, the relationship between relatedness and production frequency has not been examined for false sentences. Accordingly, we obtained ratings of the subject-predicate word pairs for each of the false sentences selected for the present experiment. Twenty-two Stanford undergraduates rated each pair as to "how closely you feel the two words are associated in meaning." A 7-point scale was used, with 7 indicating maximum relatedness. For contradiction and counterexample sentences, these ratings showed a small but highly significant positive correlation, $r = .32$, between PF and relatedness, $F(1, 98) = 11.4$, $p < .001$.[3]

Since production frequency is positively correlated with rating-scale measures of the relatedness of subject and predicate, the feature-comparison model predicts that high-PF true sentences will be verified more quickly than low-PF sentences. This prediction, of

[2] The total frequency of all possible counterexamples would not be a reliable measure of the speed of accessing the first counterexample, since numerous low frequency completions (e.g , *albatross*, *heron*) which a person is unlikely to ever think of immediately may never be the first counterexample assessed. Evidence for this assumption is provided by the Battig and Montague (1969) norms, which indicate that only a small subset of the total instances given as responses for a category are ever the first response.

[3] The mean ratings of high-PF contradictory, low-PF contradictory, and anomalous sentences were 4.88, 4.47, and 1.77, respectively. The greater relatedness of high-PF as opposed to low-PF sentences was highly significant treating subjects as a random effect, $t(42) = 3.11$, $p < .01$, and marginally significant treating items as a random effect, $t(48) = 2.24$, $p < .05$. Note that only the significance level across subjects is strictly relevant in assessing the predictions of the feature-comparison model for our particular item set. The marginal significance of this difference when tested against item variability only indicates that other high- and low-PF contradictory sentences might not show the same difference in relatedness (see Clark, 1973). For counterexample sentences, relatedness was clearly higher for the high-PF than for the low-PF sentences (5.65 vs. 4.61), $F'(1, 52) = 11.4$, $p < .001$. No other relatedness differences achieved significance treating both subjects and items as random effects.

course, is the same as that derived from the type of ordered search model we have been assuming. But since relatedness increases with PF for false sentences as well, the predictions concerning false RT differ radically. The feature-comparison model predicts that high relatedness will *slow down* false RT (see Smith *et al.*, 1974). Accordingly, the feature-comparison model predicts that RT should be slower for high-PF than for low-PF false sentences. For contradictory sentences, this is precisely opposite to the prediction of the contradiction hypothesis.

For counterexample sentences the situation is slightly more complex, since the counterexample hypothesis specifies counterexample frequency, rather than the frequency of the sentence itself, as the critical factor in determining RT. For sentence pairs in which the high-PF sentences are associated with low-frequency counterexamples and the low-PF sentences with high-frequency counterexamples, the predictions of the feature-comparison model and of the counterexample hypothesis coincide, both predicting that high-PF sentences will be rejected more slowly than low-PF sentences. But for sentence pairs in which the high-PF sentences are associated with high-frequency counterexamples, and low-PF sentences with low-frequency counterexamples, the feature-comparison model continues to predict relatively slow RT for the high-PF sentences, whereas the counterexample hypothesis predicts just the opposite, since in this case a counterexample should be found more quickly for the high-PF than for the low-PF sentences.

All these differing predictions were tested in a sentence verification experiment, using items selected from the false PF norms collected in Experiment I, and the true PF norms of Glass and Holyoak (1974) and Glass *et al.* (1974). As well as examining contradictory and counterexample false sentences, the experiment also compared the time to reject anomalous sentences with the time to reject meaningful contradictory sentenc es The results of Glass *et al.* indicated that the predictions of the contradiction hypothesis hold only for meaningful false sentences. But while no theoretical predictions were made concerning anomalous sentences, previous research (e.g., Glass *et al.*, 1974) led us to expect them to be rejected quickly. Since the subject and predicate pairs in the anomalous sentences were given very low relatedness ratings, the feature-comparison model predicts that such sentences will be rejected rapidly, after a quick assessment of overall relatedness.

Method

The present experiment contained two sub-experiments. One sub-experiment focused on the relative speeds for subjects to reject high- and low-PF contradictory and anomalous sentences; the second one focused on the speed of rejecting counterexample false *all*-statements, depending on the frequency of a counterexample. Each sub-experiment included an equal number of true statements. These differed in production frequency, and it was expected that high frequency would lead to quick verification. In the full experiment, items from these two sub-experiments were randomly mixed for presentation. Table 2 presents examples of the various sentence types used.

Materials. The entire item set contained 126 true and 126 false sentences. The first sub-experiment comprised 156 items. To form the 78 anomalous and contradictory false test sentences, 13 subject nouns quantified by *all* and 13 quantified by *some* (of which nine were the same for both quantifiers) were each paired with three noun completions. In each case, one predicate (noun) was a high-PF contradictory (e.g., *All fruits are vegetables*) given by over 19% of respondents (a mean of 35%), one was a low-PF contradictory (e.g., *All fruits are flowers*) given by less than 7% of the respondents (a mean of 5%), and one was anomalous (e.g., *All fruits are hills*) given by a mean of 4% of respondents. All such contradictories are sentences that are

TABLE 2

TWO EXAMPLES OF EACH SENTENCE TYPE USED IN EXPERIMENT II

True sentences	High-frequency	Medium-frequency	Low-frequency
All doctors are...	people	physicians	graduates
All boys are...	males	mammals	organisms
Some animals are...	mammals	primates	deer
Some chairs are...	rockers	stools	thrones
False sentences	**High-frequency contradictory**	**Low-frequency contradictory**	**Anomalous**
All fruits are...	vegetables	flowers	hills
All valleys are...	mountains	lakes	organisms
Some boys are...	girls	monkeys	houses
Some chairs are...	tables	beds	stars
	High-frequency counterexample[a]		**Low-frequency counterexample[a]**
	Low-frequency sentence		High-frequency sentence
All flowers are...	pansies (roses)		roses (pansies)
All buildings are...	jails (homes)		homes (factories)
	High-frequency sentence		Low-frequency sentence
All prisoners are...	women (men)		thieves (innocent)
All animals are...	birds (mammals)		males (females)

[a] The most frequent possible counterexample (from true *some*-completion norms) is given in parentheses.

false whether quantified by *all* or by *some*. These 26 triplets of false sentences were matched with 26 triplets of true sentences, half quantified by *all* and half by *some*. For these sentences, the three true predicates assigned to each subject word represented three levels of production frequency as indexed by the "true" completion norms of Glass and Holyoak (1974) and Glass *et al.* (1974). The predicates of high-, medium-, and low-PF true sentences were given by a mean of 47%, 14% and 4% of respondents, respectively.

The remaining 96 test sentences were true and false *all*-statements used to test our predictions for counterexample sentences. The 48 false counterexample *all*-statements would have been true had the quantifier been *some*. As the examples at the bottom of Table 2 illustrate, the frequency of the sentence itself was varied orthogonally with the frequency of the most common counterexample, producing four groups of 12 sentences each. To construct the counterexample sentences, 24 subject nouns quantified by *all* were each paired with two false predicates, one of which had been given with high frequency (mean of 24% of respondents), the other with low frequency (mean of 3% of respondents) in our false PF norms. For half of these pairs, the most dominant counterexample to the low-PF sentence appeared with high frequency (given by a mean of 32% of the respondents) as a true completion of *Some S are* ___ in the norms of Glass and Holyoak and Glass *et al.*; while the most common counterexample to the high-PF sentence was given less often (by a mean of 10% of respondents). For example, *All buildings are jails* was a low-PF counterexample sentence, and it has a high-frequency counterexample, since the sentence can be disconfirmed by the very common *some-*

completion *house*. On the other hand, *All buildings are houses* is a high-PF sentence which has only a low-frequency counterexample. For the remaining 12 false sentence pairs within this sub-experiment, this relationship was reversed. For instance, our false norms indicate that *All animals are birds* is a high-PF false sentence and it has a high-frequency counterexample (*mammals*); on the other hand, *All animals are males* is a low-PF sentence and it has a low-frequency counterexample (*females*).[4] For this group of items the high-frequency counterexamples were given by a mean of 42% of respondents, and the low-frequency counterexamples by a mean of 0%. The 48 false counterexample *all*-statements were matched with an equal number of true *all*-statements, the latter divided into two levels of production frequency.

Procedure. Sentences were presented by means of a tachistoscope. At a signal from the experimenter, the subject initiated a trial by pressing a start button. A dot then appeared in the viewer indicating where the sentence would begin. After 1 second the sentence was presented. The subject then pressed one of two decision buttons indicating whether the sentence was true or false; this response stopped a timer and removed the sentence. Before the next trial began the experimenter informed the subject whether his response had been correct. Assignment of hand to true response button was counterbalanced across subjects. Subjects were instructed to respond as quickly as they could, but to make as few errors as possible.

Test sentences were presented in seven blocks of 36 items; all experimental conditions were represented equally often in each block. Presentation order within these item blocks was random, and the order of the seven blocks was changed for each subject. Twenty different practice sentences were given prior to the test items.

Subjects were 14 Stanford undergraduates (7 males) who participated either for pay or course credit. Data from one subject who made errors on over 18% of his responses was excluded, and another subject was tested as a replacement.

Results and Discussion

Mean correct RT and error rate for each condition is presented in Table 3, along with the mean relatedness ratings for the false conditions. RTs which exceeded the subject's mean RT for that item type by 2 seconds (less than 0.3% of responses) were counted as errors. The overall error rate was 6.8%, and across conditions error rates were positively correlated with RT. For purposes of analysis of variance, errors were replaced by the subject's mean RT for that cell of the design. Both items and subjects were treated as random effects in all analyses of variance, and quasi *F*-ratios were calculated (Winer, 1971). The symbol F' denotes quasi *F*-ratios, while t' will denote the related quasi *t*-statistic (see Clark, 1973).

In Table 3, the pattern of RT differences gives striking confirmation to every prediction based on the relationship between PF and search order. First, for the true sentences matched with anomalous and contradictory falses, RT decreased monotonically as produc-

[4] For the latter 12 sentence pairs, the sentences with high- vs. low-frequency counterexamples also tended to differ in the "set relation" of the subject and predicate, as defined by Meyer (1970). Eight of the 12 sentences with high-frequency counterexamples were "superset" statements—those in which all predicate exemplars are also subject exemplars (e.g., *All birds are canaries*); whereas just three of the 12 sentences with low-frequency counterexamples were of this type. In each case the remaining items were "overlap" statements—those in which only some predicate exemplars are subject exemplars (e.g., *All birds are pets*). This confounding would be problematic only if it were shown that superset statements are intrinsically easier to reject than overlap statements. This seems unlikely, since Meyer's data suggests just the opposite. Moreover, since Meyer did not control for counterexample frequency, it is possible that the differences he found in RT to reject superset vs. overlap statements were due to variations in availability of a counterexample.

TABLE 3

MEAN RT AND ERROR RATE[a] FOR TRUE AND FALSE SENTENCES, AND RELATEDNESS RATINGS[b] FOR FALSE SENTENCES

True statements

	Quantifier	
	All RT	Some RT
High-PF	1364(7.2)	1340(2.8)
Medium-PF	1436(8.8)	1422(5.5)
Low-PF	1454(10.4)	1427(4.4)

Contradictory and anomalous statements

	All RT	Relatedness	Some RT	Relatedness
High-PF contradictory	1292(1.6)	4.97	1346(3.3)	4.79
Low-PF contradictory	1444(6.1)	4.85	1492(13.9)	4.09
Anomalous	1289(1.1)	1.74	1315(1.1)	1.79

Counterexample all-statements

	All RT	Relatedness
High-PF with low-frequency counterexample	1529(18.7)	5.89
Low-PF with high-frequency counterexample	1421(4.8)	4.92
High-PF with high-frequency counterexample	1373(6.5)	5.41
Low-PF with low-frequency counterexample	1482(13.0)	4.30

[a] Percent error is given in parentheses.
[b] Maximum = 7.

tion frequency increased across three levels, $t'(74) = 2.69$, $p < .01$. Neither the residual systematic variance associated with production frequency nor the difference between the two quantifiers achieved significance. For the true *all*-statements matched with counterexample falses, high-PF sentences were also verified more quickly than low-PF sentences, $F'(1, 33) = 9.00$, $p < .001$. These results replicate the positive effect of production frequency on true RT reported in previous work (Glass & Holyoak, 1974; Glass *et al.*, 1974; Loftus, 1973; Wilkins, 1971).

Second, consider the analysis of the RT data for the contradictory and anomalous false conditions. The three different types of completions produced significant differences in false RT, $F'(2, 70) = 9.17$, $p < .001$. As predicted by the contradiction hypothesis, high-PF contradictory sentences were rejected more quickly than low-PF contradictory sentences. This result extends the similar findings for false *many*-statements reported by Glass *et al.* (1974), and disconfirms the opposing prediction made by the feature-comparison model. Anomalous sentences were rejected a non-significant 17 msec more quickly than the high-PF Contradictory sentences.[5] Consideration of possible accounts of the rapid rejection of anomalous sentences will be postponed until the final discussion section. The difference in RT between *all-* and *some*-statements was not statistically significant in any cell of the design.

Another analysis of variance was performed on the RT data for false counterexample *all*-statements. As predicted by the counter-

[5] Edward E. Smith has recently repeated a portion of the above experiment. While he replicated our results for contradictory sentences he found that anomalous sentences were rejected significantly more quickly than high-PF contradictories. If the non-significance of the difference in our study is attributable only to experimental error, this would bring the present results more closely in line with those of Glass *et al.* (1974), who also found that anomalous sentences were rejected most quickly.

example hypothesis, sentences with high-frequency counterexamples were rejected more quickly than those with only low-frequency counterexamples, $F'(1, 54) = 7.34$, $p < .01$. Production frequency of the test sentence itself did not significantly affect RT to reject counterexample sentences, either as a main effect, $F' < 1$, or through the interaction with counterexample frequency, $F'(1, 48) = 1.43$, $p > .20$. The prediction of the feature-comparison model, that the high-PF sentences would be relatively slow to reject regardless of counterexample frequency, is thus contradicted. Nevertheless, a multiple regression analysis performed on the counterexample data did reveal a significant residual effect of rated relatedness (as opposed to production frequency) in the direction predicted by the feature-comparison model. After the variance due to counterexample frequency was accounted for, relatedness had a partial correlation of .43 with RT, $F(1, 45) = 9.97$, $p < .01$. Thus while relatedness clearly does not account for the effect of counterexample frequency, the possibility that the former variable has some independent influence on RT for this particular class of sentences cannot be ruled out.

Further analyses tested a possible alternative explanation of our false RT results. The production frequencies of the predicate words from our test sentences turned out to be positively correlated ($r = .46$ for the contradictory sentences, $r = .32$ for the counterexample sentences) with their frequencies in printed text (word frequency). We therefore used multiple regression analysis to compare the usefulness of word frequency versus associative production frequency as predictors of RT to reject contradictory sentences; a separate regression analysis compared word frequency and counterexample frequency as predictors of RT to reject counterexample sentences. In order to make the distributions of these three variables approximate normal distributions more closely, logarithmic transformations were applied. As expected, production frequency (PF) was the better predictor of false RT for contradictory sentences, while counterexample frequency was the better predictor of RT to reject counterexample sentences. When the variance attributable to these variables was accounted for, the partial correlations between word frequency and RT were negligible and nonsignificant ($r = -.06$ in the contradictory analysis, and $r = -.10$ in the counterexample analysis). In contrast, when the variance due to word frequency was removed, the partial correlation between PF and RT in the contradictory analysis was $r = -.42$, $F(1, 49) = 10.4$, $p < .01$; while that between counterexample frequency and RT in the counterexample analysis was $r = -.44$, $F(1, 45) = 10.3$, $p < .01$. The present results are therefore clearly not attributable to the effects of word frequency.

Experiment III

A major finding of Experiment II was that certain false *all*-statements (e.g., *All birds are canaries*) appear to be disconfirmed by finding a counterexample (e.g., *robin*) that contradicts the predicate. This strategy must involve a search through exemplars of the subject. However, the person clearly does not know in advance whether a particular sentence will be false, so if he searches exemplars at all, he must do so for true as well as false sentences. Can an exemplar search play a role in the verification of true as well as false *all*-statements?

So far we have been tacitly assuming that all true *all*-statements are verified by finding that the predicate is stored in memory directly with the subject. For instance, we would assume that the sentence *All birds are animals* is verified by simply finding that *bird* entails *animal*. For such sentences there should be no need for the person to examine exemplars of *bird* at all. However, it seems possible that a person might encounter an *all*-statement for which the predicate is *not* stored as a superordinate of the subject, and yet he might

nevertheless have sufficient information in memory to convince him that the sentence is true. For example, suppose a person is asked whether he believes that all birds lay eggs. He may never have previously thought about this proposition. Nevertheless, he may recall some exemplars of *bird*, and find that each exemplar lays eggs. Since he can recall only confirmatory evidence, and no counter-example, he may therefore conclude that the *all*-statement is true. This type of decision process would be a simple kind of inductive reasoning. Note that if the sentence were false (e.g., *All birds can fly*), this same procedure would lead to the discovery of a counter-example (such as *ostrich*), and consequently to a "false" decision.

Tversky and Kahneman (1973) have shown that people use this type of exemplar search in a variety of situations in order to estimate frequency or probability. People appear to conclude that a class of events or objects is relatively frequent if they can rapidly bring to mind many exemplars. For example, in one experiment people first listened to lists of names, and then were asked to judge whether more of the names had been male or female. On some lists the female names were famous (e.g., *Elizabeth Taylor*), while the male names were less famous (e.g., *William Fullbright*); while on other lists the relationship between sex and fame was reversed (e.g., *Richard Nixon* vs. *Lana Turner*). In either case, a large majority of the subjects erroneously judged the sex consisting of the more famous names (which an independent group of subjects found easier to recall) to be more frequent. These and other results reported by Tversky and Kahneman suggest that people rely on the availability of exemplars to judge the frequency of a class. Our hypothesis extends this principle to sentence verification: If people can recall exemplars which support some proposition, and none that are counter-examples, then they are likely to accept the proposition as universally true.

Experiment III was designed to test whether it is possible to distinguish *all*-statements that can be verified directly from those that require the retrieval of information about exemplars, and at the same time to extend the counter-example hypothesis to sentences with a different structure than those studied previously. We selected true and false sentences of the form *All/Some S have P*. All the true sentences were designed to be true whether quantified by *all* or by *some*. Half of these contained predicates that we expected to be stored directly as easily-accessed facts about the subject concept (e.g., *All/Some knives have blades*). We will refer to this subset simply as the "easy" sentences. The other half of the sentences, the "difficult" sentences, contained predicates that we hoped would not be stored directly with the subject concept, but could nevertheless be verified by considering exemplars (e.g., *All/Some knives have handles*).

According to our hypothesis, the relationship between the processes involved in verifying these two classes of true sentences should differ as a function of the quantifier used. When quantified by *all*, the easy sentences should be verified directly on the basis of the subject concept alone, without searching exemplars. Similarly, when the easy sentences are quantified by *some* they can still be verified directly on the basis of the subject concept, or else on the basis of whatever exemplar is considered first. Accordingly, there should be little difference in the RT to verify easy sentences as a function of quantifier. However, the situation should be quite different with the difficult sentences. When these sentences are presented with *some*, they should still be verified on the basis of the first exemplar, although perhaps after a somewhat longer search. But when the difficult sentences are quantified by *all*, they can be verified neither by direct access from the subject concept nor by considering a single exemplar. In fact, no matter how many exemplars of the subject are found that satisfy the predicate, it will still be possible that further search will retrieve a

counterexample. But if people apply the availability heuristic they should consider a number of exemplars, find that all are compatible with the predicate, and consequently decide to respond "true." Clearly this type of extended search process for difficult *all*-statements should drastically increase their RT relative to difficult *some*-statements.

The false *all*-statements were again selected to test the counterexample hypothesis. These sentences expressed propositions that could be falsified either by accessible exemplars (e.g., *All animals have wings*, for which the common mammals are counterexamples), or less accessible exemplars (e.g., *All animals have legs*, for which *snake* might serve as a counterexample). Our prediction, of course, was that the sentences with more accessible counterexamples would be more easily rejected.

Method

Subjects were timed as they responded true or false to visually-presented sentences of the form *All/Some S have P*.

Materials. The complete item set consisted of 60 true and 40 false sentences. Of these, 40 of the true sentences and 20 of the false sentences were of experimental interest; the remainder of the items served as fillers. All of the 60 critical items are listed in Table 4. To form the 40 true sentences in Table 4, 10 subject words were each paired with two predicates that formed an easy and a difficult sentence, respectively, and with both *all* and some (e.g., *All/Some knives have blades, All/Some knives have handles.*) The easy and difficult sentences were constructed on the basis of the experimenters' intuitions. While no actual measures were taken, it should be

TABLE 4

TEST SENTENCES USED IN EXPERIMENT III

True *all*- and *some*-statements	Easy predicates	Difficult predicates
1. All/Some knives have	blades	handles
2. All/Some cars have	wheels	dashboards
3. All/Some volumes have	pages	bindings
4. All/Some oceans have	water	islands
5. All/Some countries have	people	police
6. All/Some forests have	trees	insects
7. All/Some birds have	feathers	gizzards
8. All/Some cities have	buildings	animals
9. All/Some bicycles have	wheels	spokes
10. All/Some planes have	wings	tails

Counterexample *all*-statements[a]	Predicates with highly accessible counterexamples	Predicates with less accessible counterexamples
1. All countries have	presidents (England)	ports (Switzerland)
2. All planets have	people (Mars)	moons (Mercury)
3. All animals have	wings (dog)	legs (snake)
4. All buildings have	elevators (school)	windows (fallout shelter)
5. All continents have	kingdoms (North America)	cities (Antarctica)
6. All rooms have	beds (kitchen)	windows (prison cell)
7. All plants have	thorns (ivy)	leaves (cactus)
8. All shoes have	buckles (tennis shoes)	laces (loafers)
9. All clocks have	chimes (alarm clock)	hands (digital clock)
10. All cups have	decorations (plastic cup)	handles (dixie cup)

[a] The most frequent counterexample is given in parentheses.

clear from inspection of the items that the easy sentences would be generated with relatively high production frequency, and contain highly-related subject and predicate words; while the difficult sentences would undoubtedly almost never be produced in a PF task, and contain less related subject and predicate words. Our aim was to select easy sentences which could be verified directly on the basis of a connection between the subject and predicate concepts. The difficult sentences, on the other hand, were intended to be propositions that would not be stored in memory directly, but which people would nevertheless believe to be true on the basis of indirect evidence, such as consideration of exemplars. Note that if our intuitions were wrong, and the easy and difficult sentences actually did not differ in their probability of being stored directly, then our critical prediction—that difficult sentences would take longer to verify when quantified by *all* rather than by *some*, while easy sentences would not differ in RT across quantifiers—would be expected to fail.

The 20 false sentences listed in Table 4 were counterexample *all*-statements, selected to again test the counterexample hypothesis. Ten subject words were each paired with two predicates, each of which formed a false statement. One of these statements could be rejected on the basis of an obvious counterexample (e.g., *All plants have thorns*, which is falsified by most fruits and vegetables), while the other could only be falsified by less quickly accessed counterexamples (e.g., *All plants have leaves*, which is falsified by moss, evergreen trees, etc.). The 20 sentences were constructed by the experimenters, and then two empirical measures of counterexample availability were taken. First, 20 undergraduate subjects were asked to rate each sentence from 1 to 7 on the basis of "how difficult it is to think of a counterexample that shows that the sentence is false" (with 7 indicating maximal difficulty), and also to write down the first counterexample that they thought of for each sentence. Subjects were first shown an example of a false statement (*All mammals are hairy*) along with possible counterexamples (*whales* and *dolphins*). While subjects tended to use only the lower end of the scale, for each of the 10 pairs of sentences they gave a higher mean rating of the difficulty of thinking of a counterexample to the sentence selected by the experimenters to have a less available counterexample. The mean rating was 1.48 for the sentences with highly accessible counterexamples, and 2.87 for the sentences with less accessible counterexamples.

Second, a production frequency (PF) measure was also used to assess counterexample availability. A different group of 21 undergraduate subjects were presented with the 10 categories used as subject words in the counterexample sentences (e.g., *country*) and asked to list the first three instances of the category that came to mind. These data were then used to calculate the mean frequencies of the counterexamples listed for each sentence by the first group of subjects. The highly accessible counterexamples were produced by a mean of 23% of the respondents in the instance production task, while the less accessible counterexamples were given by a mean of only 6% of the respondents. For each of the 10 sentence pairs, mean counterexample frequency was higher for the sentence chosen to have the more accessible counterexample. The most common word provided as a counterexample is given for each counterexample sentence in Table 4.

Measures were also taken for two possible confounding variables. Length of the predicate words (in both letters and syllables) tended to be somewhat greater for the sentences with more accessible counterexamples, while their word frequency (from Kučera & Francis, 1967) tended to be lower. To the extent that these factors have an effect on RT, they would therefore weigh against the prediction of the counterexample hypothesis.

These 40 false *all*-statements were balanced by 40 relatively meaningful false *some*-

statements (e.g., *Some lizards have feathers*). These *some*-statements were essentially filler items, of no particular experimental interest. All of the *some*-statements so far discussed would have had the same truth value if quantified by *all*. It would therefore have been possible for subjects to develop a strategy of ignoring the presented quantifier, and processing every sentence as though it were quantified by *all*. To prevent the use of this strategy an additional 10 *some*-statements were included in the design. These were filler items that were true only with the quantifier *some*, and not with *all* (e.g., *Some buildings have elevators*). Finally, an additional 10 true *all*-statements were included in order to maintain an equal number of true *all*- and *some*-statements.

Procedure. The sentences were presented tachistoscopically in the same manner as in Experiment II. In addition to the usual RT-task instructions, subjects were given several explicit instructions on how to evaluate the sentences. They were told that any sentence that would be true for *all* would also be true for *some* (e.g., *Some cows are mammals*). They were instructed to consider only "normal" exemplars in evaluating generalization; for example, *All people have two legs* was to be considered true despite the existence of amputees. Also, they were told that generalizations that are true only "accidentally," rather than by definition (e.g., *All bachelors are over a foot tall*), should be considered true. Finally, subjects were asked to inform the experimenter if on any trial they were forced to simply guess at the answer. In addition, on any trial on which the subject made an error, the experimenter ascertained whether the subject in fact knew that his answer was incorrect. Usually the subject would immediately realize he had made an error. If not, he was questioned further. If he had responded "true" to a counterexample sentence (e.g., *All continents have cities*), he was asked whether a particular counterexample in fact made the sentence false (e.g., Antarctica). If he had responded "false" to a true generalization (e.g., *All oceans have islands*), he was asked whether he could think of a counterexample to the sentence. If after questioning the subject indicated either that he simply had not known the answer, or that he actually disagreed with the experimenter's view of the truth value of the sentence, that fact was recorded.

The sentences were presented in random order, with the restriction that a minimum of 20 items intervened between repetitions of the same sentence frame with different quantifiers. All subjects made "true" responses with their right hand. Fourteen Stanford graduate students served as volunteer subjects.

Results and Discussion

An important methodological problem presented itself in analyzing the results of Experiment III. Our hypothesis concerning "true" decisions required that the study include statements that people would believe to be true, even though they had not stored the information directly. Accordingly, it was important to be sure that subjects in fact agreed with the experimenter's assessment of the truth value of these sentences. Clearly it would be no surprise if people required longer to verify a difficult statement when it was quantified by *all* rather than *some*, if they knew that the *some*-statement was true but had no idea whether the corresponding *all*-statement was true or not. Similarly, it would be of no interest to show that the false sentences with less accessible counterexamples take a long time to reject if subjects simply had never heard of the relevant counterexamples. Note that what is critical is not whether the sentences used were *actually* true or false descriptions of the world, but whether our subjects *believed* them to be true or false on the basis of information they possessed at the time they were required to make a decision.

As described above, the experimenter noted those trials when the subject simply did not

know the truth value of the sentence or actually disagreed with the experimenter. Each of these trials was eliminated from the data analysis, together with the corresponding trial in the relevant comparison condition. For instance, if a subject did not believe *All cars have dashboards* to be true, then his response to that sentence, as well as his response to *Some cars have dashboards*, was simply discarded. For the counterexample sentences, if the subject did not believe that a sentence with a less accessible counterexample was false (e.g., *All continents have cities*), then his response to the matched sentence with a more accessible counterexample (*All continents have kingdoms*) was also eliminated. This procedure was designed to avoid biasing our results by the inclusion of comparisons involving sentences that subjects simply were unable to evaluate correctly. Overall, subjects disagreed with 1% of the easy true statements, 15% of the difficult true statements, 0% of the false statements with highly accessible counterexamples, and 4% of the false statements with less accessible counterexamples. In evaluating the RT and error rate data reported below, it is important to remember that all comparisons involving these disputed sentences were discarded prior to any further data analysis. In particular, note that the reported mean error rates are based only on those trials on which subjects agreed that their response had been incorrect.

TABLE 5

MEAN CORRECT RT AND ERROR RATE FOR TRUE AND FALSE SENTENCES IN EXPERIMENT III

True sentences	All		Some	
	RT	% Error	RT	% Error
Easy	1296	4	1190	1
Difficult	1594	20	1297	2

False sentences	RT	% Error
All-statements with highly-accessible counterexamples	1437	15
All-statements with less accessible counterexamples	1617	43
Some-statements	1427	4

The mean RT and error rate for each condition of experimental interest is reported in Table 5. Response times more than 2000 msec over the subject's mean RT for that condition (2% of the responses) were discarded and counted as errors. Subjects made only 4% errors on the filler true *some*-statements that would have been false if quantified by *all* (e.g., *Some buildings have elevators*), indicating that they processed all sentences in terms of the presented quantifier.

As the data in Table 5 indicate, our predictions for the true statements were essentially confirmed. Analyses of variance were performed on the RT data, and minimum quasi F-ratios were calculated using the methods described by Clark (1973). Overall, subjects required less time to verify easy than difficult sentences, $minF'(1, 18) = 9.37$, $p < .01$, and they verified sentences quantified by *some* more quickly than sentences quantified by *all*, $minF'(1, 28) = 6.34$, $p < .05$. But as predicted, the relative advantage of *some-* as opposed to *all*-statements was greater (by 191 msec) for the difficult than for the easy statements. The overall interaction was significant both across subjects, $F(1, 13) = 10.1$, $p < .01$, and across items, $F(1, 18) = 5.38$, $p < .05$, although the conservative minimum quasi F-statistic fell short of formal significance, $minF'(1, 25) = 3.51$, $p < .10$. The easy statements were verified a nonsignificant 106 msec more slowly when quantified by *all* as opposed to *some*, $minF'(1, 27) = 1.19$, $p > .25$; while the difficult generalizations required an extra 297 msec to verify when quantified by *all*, $minF'(1, 27) = 9.33$, $p < .01$. Essentially the same pattern of results appears in the error rate data. These results support the hypothesis that the predicates of easy sentences are stored directly with the subject concepts, but that verification of difficult statements quantified by *all* requires a more indirect, inductive evaluation process, possibly involving a

search for information about exemplars of the subject.

For the false sentences, the prediction of the counterexample hypothesis was confirmed. Statements falsified by highly accessible counterexamples (e.g., *All buildings have elevators*) were rejected 180 msec more quickly than were generalizations falsified by less accessible counterexamples (e.g., *All buildings have windows*), $minF'(1, 22) = 7.24$, $p < .025$. This RT difference was paralleled by a dramatic increase in error rate, from 15 to 43%. Subjects clearly experienced great difficulty in reaching correct decisions about those sentences for which it was difficult to think of a counterexample under the time pressure of an RT task, even though immediately after making an error the subjects realized that they in fact did know of relevant counterexamples.

One result which requires further explanation is the nonsignificant trend for easy true statements to be verified more quickly when quantified by *some* rather than *all*. If these generalizations are stored directly, they should be verified equally easily with either quantifier. Indeed, subjects often reported that *some* statements clearly true for *all* (e.g., *Some oceans have water*) seemed unnatural. Previous research in fact indicates that high-PF *all*-statements take *longer* to verify when quantified by *some* (Glass & Holyoak, 1974). Why, then, do the present results indicate a trend (even if it is non-significant) for these *some*-statements to be verified more quickly? A straightforward explanation for this tendency was spontaneously offered by a number of our subjects. Since the present study included a large number of extremely difficult false *all*-statements, which subjects often erroneously judged to be true, the subjects tended to become cautious in responding "true" to any *all*-statement. In particular, note that the false *all*-statements were considerably more difficult, in terms of both RT and error rate, than were the false *some*-statements. The previous results indicating that high-PF *all*-statements are verified more slowly when quantified by *some* were obtained when the difficulty of false sentences was equalized for the two quantifiers (Glass & Holyoak, 1974). What is important in interpreting the present results is that the advantage of *some* as opposed to *all* becomes both considerably larger and more reliable when the difficult sentences are considered.

General Discussion

Let us recapitulate our major findings. Experiment I demonstrated that words which are given frequently as false predicate-completions of the sentence frame *All/Some S are* ____ generally bear a clear semantic relationship to frequent true completions. In Experiment II the predictions of the contradiction and counterexample hypotheses were tested in a verification experiment. For contradictory sentences—those false whether quantified by *all* or by *some*, but with related subject and predicate words (e.g., *All fruits are vegetables*)—high predicate PF produced fast false RT. For counterexample sentences—those false when quantified by *all*, though true for *some* (e.g., *All fruits are oranges*)—PF had no effect on RT. For these sentences RT was determined by how closely a possible counterexample (e.g., *apple*) was associated with the subject noun. For all these sentences high PF was positively correlated with rated relatedness of subject and predicate words. These results therefore demonstrated that an increasing monotonic relationship does not necessarily hold between false RT and semantic relatedness. Experiment III extended the counterexample hypothesis to sentences with the verb *have* (e.g., *All buildings have elevators*), and also provided some evidence that exemplar search may play a role in the verification of some true *all*-statements (e.g., *All knives have handles*).

Counterexamples and Generalizations

Several interesting theoretical issues are raised by the results of Experiment III. Our results for true sentences suggest that sentences with predicates that have never been stored directly with the subject can still be verified on the basis of an inductive reasoning process that is qualitatively different from the more deductive process by which high-PF true sentences are verified. However, the present results do not clearly indicate the nature of this inductive process. We have suggested that at least one aspect of this decision strategy involves the retrieval and evaluation of exemplars, a process that is also the basic mechanism for disconfirming counterexample sentences. For instance, in order to decide that all knives have handles, the subject may retrieve a number of exemplars of the class of knives, find that each exemplar supports the generalization, and hence decide to respond "true." Many subjects in fact reported using this kind of decision process. But a different kind of strategy was also commonly mentioned. Subjects sometimes described performing what may be best described as a "theoretical analysis" of the presented generalization. For instance, in order to decide whether all oceans have islands, the subject might reason that oceans are huge bodies of water, that islands are common everywhere in the world, and hence that the generalization is virtually certain to be true. Interestingly, this type of reasoning was perhaps most commonly reported on trials when the subject made an error. For the above example, the subject might also reason that nothing in his knowledge about oceans requires that they have islands, and therefore decide to respond "false," even though he might later realize that in fact he knew of islands in every ocean. When successful, this type of theoretical analysis appeared to invoke information over and above the definition of the subject word, including judgments of magnitude and numerosity.

The process of theoretical analysis may be related to another decision-making heuristic, "representativeness," that is discussed by Kahneman and Tversky (1972, 1973). Kahneman and Tversky found that people judge the probability that an instance will fall into a particular class (e.g., that some student will be an engineer) on the basis of how well it fits their stereotype of a member of that class. While exemplar search and the availability heuristic involve searching a number of instances of a category, decisions based on theoretical analysis or the representativeness heuristic involve consideration of only the general concept, or perhaps a typical exemplar of the category.

Clearly the strategies of exemplar search and of theoretical analysis are not mutually exclusive. Indeed, subjects often reported considering information about one or more exemplars as part of a more general reasoning strategy. The problem of determining when exemplar search as opposed to theoretical analysis will be applied to a sentence may be similar to the problem of identifying the variables that influence people's choice between the availability and the representativeness heuristics (Tversky & Kahneman, 1973). It seems possible that there may be reliable differences among different items in the type of reasoning strategies they trigger. Subjects appeared more likely to report using a theoretical analysis during the verification of sentences for which it intuitively seemed relatively difficult to rapidly think of large, well-defined classes of exemplars of the subject (e.g., *All cities have animals*). The systematic investigation of the variables that determine the use of different reasoning strategies in the evaluation of generalizations would appear to be an important area for future research.

A related issue raised by the present results concerns the distinction between knowledge about word meanings and knowledge about the referents of words. Intuitively, the fact that all oceans have water appears to be true

by virtue of the definition of *ocean*, while the fact that all oceans have islands seems to be an accident of geography. A long tradition in semantic theory has attempted to distinguish between these two types of propositions, the analytic and the synthetic (see Katz, 1972). Psychologists investigating semantic memory have tended to simply ignore this distinction, leaving it unclear whether their theories were about the representation of word meanings or of facts (see Glass & Holyoak, in press). In philosophy, it has been debated whether or not such a distinction should be drawn at all (Quine, 1953).

We believe that the question of whether or how the analytic–synthetic distinction should be incorporated into a psychological model is one of the major theoretical problems involved in the study of semantic memory. This question cannot be decided *a priori*; rather, the answer will hopefully emerge as an integrated theory of semantic decisions is developed on the basis of empirical results. So while the present results are not at all conclusive on this issue, it is still important to assess the implications of our findings for the analytic–synthetic puzzle. Two points appear to suggest that a basis for the distinction may be found. First, the notion of a contradiction provides a mechanism for representing necessary falsehood, the counterpart of necessary truth. Second, the results for true sentences in Experiment III suggest a possible structural distinction between the representations of analytic and synthetic propositions. These results suggested that the easy sentences (which intuitively seem analytic) can be verified by searching directly from the subject to the predicate concepts, while the difficult sentences (which intuitively seem synthetic) require consideration of exemplars. A natural hypothesis, therefore, is that a necessary condition for a proposition to be analytic is that it must contain a predicate that is linked to the subject directly, rather than only to exemplars of the subject. However, this structural condition on analyticity is clearly not a sufficient definition. For instance, the propositions *All birds are feathered* and *All bachelors are male* may both be stored as directly connected subject and predicate concepts. Furthermore, people may be equally confident of the truth of either sentence. Nevertheless, one can perhaps conceive of a bird species without feathers (such as penguin-like creatures with fur) that could still fall under the concept "bird"; but it seems more difficult to imagine the possibility of a non-male bachelor. This suggests that the intuition of analyticity depends on structural features of the memory representation that are yet to be specified.

However the distinction between factual and semantic knowledge may be drawn, it appears that both kinds of information can be used in a single verification decision. Consider the counterexample sentence *All men are husbands*. Nothing in the meaning of *man* entails that some men are unmarried, so accessing the exemplar *bachelor* must involve factual knowledge. On the other hand, recognizing that *bachelor* contradicts *husband* would appear to depend only on information about word meanings. The evaluation of such sentences appears to involve an integrated process of decomposing the meaning of the subject word, searching exemplars, and checking for contradictions. The available evidence thus suggests that factual and strictly semantic knowledge are organized together in memory, and that a common set of retrieval and decision procedures operates on the entire knowledge store.

Anomalous Sentences

While the present study demonstrated that the contradiction and counterexample hypotheses can successfully predict RT to reject false but meaningful sentences, these hypotheses as presently formulated are still inadequate when applied to the entire range

of false sentences. What mechanism might account for the rapid rejection of anomalous sentences, such as *All birds are chairs*? One possibility is that such sentences are rejected by some process qualitatively different from any we have yet considered. For instance, in terms of the two-stage model of Smith *et al.* (1974), anomalous sentences will almost invariably be rejected after some kind of fast holistic comparison reveals that the subject and predicate words are very unrelated in meaning.

However, it would be more parsimonious if the contradiction hypothesis could be extended to account for the disconfirmation of anomalous as well as meaningful contradictory sentences. In terms of the contradiction mechanism, the quick rejection of anomalous sentences suggests that certain abstract types of information which differentiate between almost all words (such as the distinction between "living" and "nonliving") are uniformly accessed quickly. Contradictions found at this level would lead to the quick disconfirmation of anomalous sentences (e.g., *All birds are chairs*). A similar analysis of the rejection of anomalous sentences has been offered by Kintsch (1972). The present hypothesis suggests that the same basic mechanism —discovery of a contradiction—is used to reject both anomalous and contradictory sentences. This hypothesis suggests that there is no qualitative distinction between these two types of sentences; rather, there is simply a continuum from highly related to minimally related contradictory sentences. But since the first information retrieved will include some facts which are closely related to the specific subject and predicate concepts as well as some information which is very general, both high-PF and anomalous sentences will be rejected relatively quickly. Consequently, the relationship between false RT and relatedness will be nonmonotonic over this range of sentences. The relative speed with which high-PF and anomalous sentences are disconfirmed will depend on the relative probabilities of first accessing closely-related as opposed to very general properties of the subject and predicate concepts.

While this analysis of anomalous sentences is consistent with the present results, it is admittedly ad hoc. As Experiment I demonstrated, there is a close relationship between high-PF contradictory sentences (e.g., *All birds are dogs*) and high-PF true sentences (e.g., *All birds are animals*). But the assumed true parallel to an anomalous sentence such as *All birds are chairs* would be *All birds are living things*, which is presumably a low-PF sentence. However, the low frequency of such a sentence may simply reflect the fact that the category "living thing" is not a single lexical item and appears relatively infrequently in spoken or written English. In other words, production frequency may not be a valid measure of the association strength of such abstract properties. A clear goal for future research should be to provide an independent test of the hypothesis that abstract markers are accessed quickly. Specifically, it will be necessary to create experimental conditions under which fast "true" decisions might be made about propositions involving abstract concepts like "living." Some recent data obtained by Shoben (1974) in fact provides some preliminary support for this prediction. He found that subjects were relatively quick to respond "same" to two exemplars that shared the same value on such abstract dimensions as living vs. nonliving, concrete vs. abstract, and count vs. mass, as compared to the time they required to perform the same task using less general dimensions, such as size and predacity.

In conclusion, people seem to have available a variety of interrelated heuristics for evaluating the validity of sentences. The discovery and analysis of these heuristics appears to be an essential prerequisite to the development of an integrated description of semantic decision-making.

APPENDIX

False Sentences Used in Experiment II

Contradictory All-statements

Subject word	Hi-PF predicate	Lo-PF predicate	anomalous
1. blossoms	trees	grass	teeth
2. chairs	tables	desks	pencils
3. diamonds	emeralds	glass	nails
4. fires	water	smoke	cities
5. horses	cows	mules	rocks
6. arrows	bows	spears	doctors
7. horses	birds	donkeys	money
8. valleys	mountains	lakes	organisms
9. women	men	babies	clocks
10. flowers	trees	foods	towns
11. boys	girls	fathers	trains
12. tables	chairs	benches	rats
13. fruits	vegetables	flowers	hills

Contradictory Some-statements

1. birds	dogs	reptiles	tables
2. blossoms	trees	leaves	stones
3. animals	plants	doctors	rocks
4. chairs	tables	beds	stars
5. horses	cows	donkeys	rivers
6. valleys	mountains	lakes	horses
7. fruits	vegetables	flowers	chairs
8. women	men	babies	trees
9. flowers	trees	foods	tigers
10. women	boys	babies	cars
11. boys	girls	monkeys	houses
12. boys	women	apes	bicycles
13. tables	chairs	beds	organisms

Counterexample All-statements

Subject word	Hi-PF predicate low-frequency counterexample	Lo-PF predicate high-frequency counterexample
1. snakes	rattlesnakes	vipers
2. women	mothers	writers
3. houses	mansions	cottages
4. buildings	homes	jails
5. chairs	rockers	thrones
6. forests	parks	jungles
7. buildings	houses	libraries
8. prisoners	criminals	lawyers
9. flowers	roses	pansies
10. houses	homes	churches
11. teachers	professors	janitors
12. books	novels	mysteries

Counterexample All-statements

Subject word	Hi-PF predicate high-frequency counterexample	Lo-PF predicate low-frequency counterexample
1. animals	birds	males
2. animals	reptiles	females
3. prisoners	women	thieves
4. prisoners	men	crooks
5. fruits	apples	citrus
6. fruits	oranges	spheres
7. teachers	men	parents
8. teachers	women	friends
9. gems	diamonds	necklaces
10. gems	rubies	earrings
11. birds	robins	flyers
12. birds	eagles	swimmers

REFERENCES

ANDERSON, J. R., & BOWER, G. *Human associative memory*. Washington, DC: V. H. Winston & Sons, 1973.

BATTIG, W. F., & MONTAGUE, W. E. Category norms for verbal items in 56 categories: A replication and extension of the Connecticut Category Norms. *Journal of Experimental Psychology Monograph*, 1969, **80** (3, pt. 2).

BOUSFIELD, W. A., & BARCLAY, W. D. The relationship between order and frequency of occurrence of restricted associative responses. *Journal of Experimental Psychology*, 1950, **40**, 643–647.

CLARK, H. H. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 1973, **12**, 335–359.

COLLINS, A. M., & QUILLIAN, M. R. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 1969, **8**, 240–248.

COLLINS, A. M., & QUILLIAN, M. R. Experiments on semantic memory and language comprehension. In L. W. Gregg (Ed.), *Cognition in learning and memory*. New York: Wiley, 1972.

CONRAD, C. Cognitive economy in semantic memory. *Journal of Experimental Psychology*, 1972, **92**, 149–154.

GLASS, A. L., & HOLYOAK, K. J. The effect of *some* and *all* on reaction time for semantic decisions. *Memory & 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.

GLASS, A. L., & HOLYOAK, K. J. Alternative conceptions of semantic memory. *Cognition*, 1975, in press.

KAHNEMAN, D., & TVERSKY, A. Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 1972, **3**, 430–454.

KAHNEMAN, D., & TVERSKY, A. On the psychology of prediction. *Psychological Review*, 1973, **80**, 237–251.

KATZ, J. J. *Semantic theory*. New York: Harper & Row, 1972.

KINTSCH, W. Notes on the semantic structure of memory. In E. Tulving & W. Donaldson (Eds.), *Organization and memory*. New York: Academic Press, 1972.

KUČERA, H., & FRANCIS, W. N. *Computational analysis of present-day English*. Providence, RI: Brown University Press, 1967.

LOFTUS, E. F. Category dominance, instance dominance, and categorization time. *Journal of Experimental Psychology*, 1973, **97**, 70–94.

LOFTUS, E. F., & SCHEFF, R. W. Categorization norms for 50 representative instances. *Journal of Experimental Psychology*, 1971, **91**, 355–364.

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

PAIVIO, A., YUILLE, J. C., & MADIGAN, S. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, 1968, **76**, Monograph Supplement No. 1, Part 2.

QUINE, W. Two dogmas of empiricism. In W. Quine, *From a logical point of view*. Cambridge: Harvard University Press, 1953.

RIPS, L. J., SHOBEN, E. J., & SMITH, E. E. Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*, 1973, **12**, 1–20.

ROSCH, E. R. On the internal structure of perceptual and semantic categories. In T. M. Moore (Ed.), *Cognitive development and acquisition of language*. New York: Academic Press, 1973.

SCHAEFFER, F., & WALLACE, R. The comparison of word meanings. *Journal of Experimental Psychology*, 1970, **86**, 144–152.

SHOBEN, E. J. Semantic features in semantic memory. Unpublished doctoral dissertation, Stanford University, 1974.

SMITH, E. E., SHOBEN, E. J., & RIPS, L. J. Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, 1974, **81**, 214–241.

TVERSKY, A., & KAHNEMAN, D. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 1973, **5**, 207–232.

WILKINS, A. T. Conjoint frequency, category size, and categorization time. *Journal of Verbal Learning and Verbal Behavior*, 1971, **10**, 382–385.

WINER, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1971.

WOODWORTH, R. S., & SCHLOSBERG, H. *Experimental psychology*. New York: Holt, Rinehart & Winston, 1950.

(Received November 29, 1974)