Memory & Cognition
1989, 17 (4), 463-473

A semantic memory sentence verification model based on relative judgment theory

PAUL J. CASEY
Riverina-Murray Institute of Higher Education, Wagga Wagga, New South Wales, Australia

and

RICHARD A. HEATH
University of Newcastle, New South Wales, Australia

A subjective referent model of sentence verification in semantic memory tasks based on the relative judgment theory of Link and Heath (1975), together with the derivation of a discriminability index, are presented in this paper. An attractive feature of the model is its consideration of both error rates and response times (RTs) in the calculation of the discriminability index. The model is also able to account for the frequent finding in semantic memory tasks that error RTs are longer than correct RTs. A partial replication of Experiment 2 of McCloskey and Glucksberg's (1979) sentence verification context effect studies, in which we employed 44 subjects and 28 categories, and controlled for item familiarity, revealed that error RTs were consistently longer than correct RTs—a finding inconsistent with the McCloskey and Glucksberg property comparison model, but in accord with the subjective referent model. An important fortuitous result was the detection of a context effect by the discriminability measure, an effect not detected by the RT data alone. The discriminability measures yielded a near perfect correlation with estimates of the mean step size of the random walk obtained by application of the parameter estimation program FITTRW (Heath, 1983).

Shoben (1982), in the error-management section of his tutorial on semantic memory and lexical studies, stated, "The important issue is to examine one's error data seriously" (p. 304). However, no techniques for examining error data were proposed. Chang (1986), in contrast with Shoben, dismissed error-making rather lightly, claiming that errors seldom occur with the speeded sentence verification task, and that they are "more likely due to misreadings and the like than to memory failure " (p. 200).

The reported but unanalyzed error rates from many semantic memory categorization experiments are of concern, since both error rate and response time (RT) reflect the operation of the categorization process. In most of the semantic memory literature, the frequent neglect of error data is likely due to a paucity of available techniques for analyzing error data. However, this situation is unsatisfactory. Pachella (1974) has shown that even when there are small error rates, the interpretation of the RT data becomes difficult, because small differences in error rates are often associated with large differences in RT. Moreover, error RT data do not always exhibit a speed-accuracy tradeoff, since longer mean RTs are often accompanied by a greater proportion of errors.

In this paper, we present a subjective referent model of sentence verification, together with the derivation of a discriminability index based on relative judgment theory (Link & Heath, 1975). A key feature of the subjective referent model is that it predicts that error RTs can be longer than correct RTs, while the main attraction of the discriminability index is the combination of mean RT and error rate into the one measure. The derivation of the discriminability index will be described, and then the central features of the model and measure will be illustrated in a partial replication of one of the context-effect experiments of McCloskey and Glucksberg (Experiment 2, 1979).

A Model for Semantic Judgments Based on a Random Walk Decision Process

The model proposed is the subjective referent model, according to which the categorization process is viewed as a discrimination task, the decision process is analyzed as a random walk of the kind proposed by Link and Heath (1975). The model is similar in many ways to that of McCloskey and Glucksberg (1979), who proposed that people verify category membership statements by comparing properties of the subject and predicate and making a decision when the accumulated evidence exceeds some criterion. However, the subjective referent model differs in its conceptualization of the decision process, its

This project was assisted in part by a grant from the Riverina-Murray Institute of Higher Education to P. J. Casey, and a grant from the Australian Research Grants Scheme to R. A. Heath (Grant A28416077). We wish to thank several anonymous reviewers for advice on earlier versions of the manuscript. We wish to thank the subjects for their cooperation, as well as Geoff Fellows, who assisted with data entry. This research formed part of P. J. Casey's dissertation in the Department of Psychology at the University of Newcastle, NSW. Requests for reprints should be addressed to: Paul Casey, Riverina-Murray Institute of Higher Education, P.O. Box 588, Wagga Wagga, NSW 2650 Australia.

expectations of the relationship between correct and error RTs, and its treatment of error data in general.

Link and Heath (1975) presented "a theory of discrimination which assumes that subjects compare *psychological values* [italics added] evoked by a stimulus to a subjective referent" (p. 77). The psychological values in the sentence verification task consist of measures of relational strength between category and instance. The term *relational strength* is used in a broad sense, to refer to category-instance associative strength, termed *instance dominance*, as measured by category norms (e.g., see Battig & Montague, 1969); to the level of typicality relationship between an instance and a category as measured by typicality ratings (e.g., see Rosch, 1973); or to the instance-category associative strength relationship, termed *category dominance* (e.g., see Loftus, 1973). The measure must be unidimensional.

In a sentence verification task, the subjective referent, or adaptation level, is developed during the practice session; it represents a weighted average of the kinds of items composing the sentence list. Each trial generates an internal representation of a difference between the relational strength of the stimulus pair and a subjective reference relational strength. This situation is conceptually equivalent to a line length discrimination task in which both lines are presented simultaneously. A subject compares successive relational strength samples with referent samples in order to give a series of strength differences. These strength differences are accumulated in a single counter. The making of a decision about category membership can be considered as a process of computing associative strengths, which accumulate until one of two response boundaries is exceeded. This process can be analyzed as a random-walk decision process. The measure of item discriminability that is proposed requires for its calculation only the knowledge of mean RT and error rate for a particular set of a subject's test items. The assumptions behind the measure, and its derivation, follow.

Assumptions. The assumptions behind the Link and Heath (1975) model for decision-making in psychological judgment tasks are:

1. Stimulus-derived information is a random variable representing the difference between a stimulus-derived quantity and a subjective reference value, the latter varying with the subject's experience of similar stimuli.

2. The subject can set response thresholds at positions equidistant from the origin on a dimension representing accumulated stimulus differences. The starting point for the decision process can be anywhere between these two boundaries, but it will be assumed to be set equal to zero, midway between these response thresholds. This assumption, which seems reasonable, since each stimulus in a two-choice task is equally probable, simplifies the computation of the discriminability index. It can be relaxed when a parameter estimation program such as FITTRW (Heath, 1983) is used.

3. During each sampling period of arbitrarily small duration, the stimulus difference information is added to a single counter, and a decision is made when the contents of the counter first reach or exceed one of the two response thresholds. The time taken to attain this condition is the decision time for that response. We assume that the response time equals the decision time plus a residual time, the latter representing the time consumed by other processes such as stimulus transduction and response emission.

Derivation of the discriminability index. The subjective referent model can be readily applied to decision-making in a categorization task. On each trial, two stimuli, an instance and a category, are presented, and the subject is required to judge whether the instance belongs to that category. We assume that the subject computes the strength of the relationship between the instance and the category. For example, in the McCloskey and Glucksberg (1979) context experiments, where the categorization RT for a particular sentence type was shown to be related to the other sentence types included in the test set, the strength of the relationship between the instance and the category will be high for the true, highly related stimulus pairs, termed here *true-high* (e.g., "All sparrows are birds"), and low for the *false-low pairs* (e.g., "All houses are birds"). The strengths of the relationships for *true-low* (e.g., "All chickens are birds") and *false-high* pairs (e.g., "All whales are fish") assume intermediate values. For each stimulus pair within each of the four relatedness categories, namely true-high, true-low, false-high, and false-low, relational strength is a random variable. It is also assumed that the subject has access to a reference relational strength that assumes intermediate values along the relational strength dimension. The hypothesized distributions are shown in Figure 1. The means are *M*(false-low), *M*(false-high), *M*(referent), *M*(true-low), and

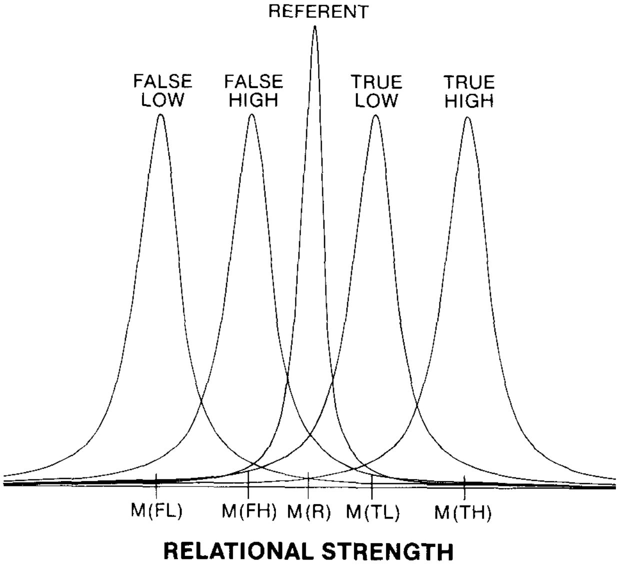

Figure 1. The distribution of relational strengths for false-low (FL), false-high (FH), subjective referent (R), true-low (TL), and true-high (TH) associations. The mean of each distrubution is indicated by M(FL), M(FH), M(R), M(TL), and M(TH) for the FL, FH, R, TL, and TH conditions, respectively.

M(true-high) for the false-low, false-high, referent, true-low, and true-high distributions, respectively.

On each trial, the subject sets the response thresholds at A and $-A$ and the counter is set at a value C, $-A < C < A$, at time $t = 0$. C serves as a response bias parameter, being closer to the expected response threshold and approximately equal to zero when each response is equiprobable prior to stimulus presentation. During each sampling period of duration, Δt, the relational strength of the stimulus pair presented, is compared with a sample from the referent to yield a strength difference on the ith sampling period:

$$d_{s_i} = a_{s_i} - a_{r_i},$$

where s = false-low, false-high, true-low, or true-high, r = referent, a_{s_i} = relational strength for stimulus pair s in the ith time period, and a_{r_i} = relational strength for the referent in the ith time period.

Let

$$D_{SN} = \sum_{i=1}^{N} d_{s_i} + C$$

be the sum of the accumulated strength differences during N sampling periods. The following decision rule determines whether the sampling terminates after N sampling periods, or continues:

If $D_{SN} \geq A$, respond "True"
If $D_{SN} \leq -A$, respond "False"
Otherwise continue sampling.

Once one of the response thresholds has been attained, the RT is given by

$$\mathrm{RT} = N\Delta t + \mathrm{RT}_0,$$

where RT_0 is a residual RT component.

A measure of mean difference in relational strength can be computed, using the relationship between mean RT pooled over both "True" and "False" responses and Pr (True/s), the proportion of "True" responses given stimulus pair s. Using a similar argument to that given in Link and Heath (1975), mean RT can be computed, using:

$$\overline{\mathrm{RT}}(s) = \frac{(A-C)\mathrm{Pr(True}/s) - (A+C)[1-\mathrm{Pr(True}/s)]}{\mu(s)} + \mathrm{RT}_0,$$

where $\mu(s) = M(s) - M(R)$, the mean random walk increment during each sampling period, is the distance between the means of the distribution s and the referent R.

Simplifying the above expression yields:

$$\overline{\mathrm{RT}}(s) = \frac{A[2\mathrm{Pr(True}/s)-1] - C}{\mu(s)} + \mathrm{RT}_0,$$

where RT_0 is the mean residual RT.

If there are equal numbers of true and false relationships and the subject is instructed appropriately, then we might expect the response bias parameter, C, to be close to zero. If we let $C = 0$, the analysis becomes substantially simplified. We now have:

$$\overline{\mathrm{RT}}(s) = \frac{A\theta(s)[2\mathrm{Pr(True}/s)-1]}{\mu(s)\theta(s)} + \mathrm{RT}_0, \quad (1)$$

where $\theta(s)$ is derived from the moment-generating function of the distribution of random walk step sizes generated by stimulus type s. $\mu(s)$ increases with stimulus discriminability and for all practical purposes may be considered proportional to $\theta(s)$.

For normally distributed steps, for example:

$$\theta(s) = \frac{2\mu(s)}{\sigma^2},$$

where σ^2 is the step size variance (Cox & Miller, 1965).

Now $A\,\theta(s)$ can be estimated directly from the data. We will assume that the starting point of the random walk is zero, a reasonable assumption when there are equal numbers of positive and negative items and subjects are instructed properly. This simplifies the mathematical representation of the discriminability index. The proportion of "True" responses is given by:

$$\mathrm{Pr(True}/s) = \frac{e^{A\theta(s)}-1}{e^{A\theta(s)}-e^{-A\theta(s)}}$$

$$= \frac{e^{A\theta(s)}}{e^{A\theta(s)}+1}$$

(Link, 1978, Equation 2). Hence,

$$A\theta(s) = \ln\left[\frac{\mathrm{Pr(True}/s)}{1-\mathrm{Pr(True}/s)}\right] \quad (2)$$

(this is akin to the logistic psychometric function employed by Link, 1978).

Substituting Equation 2 into Equation 1 yields:

$$\overline{\mathrm{RT}} = \frac{[2\mathrm{Pr(True}/s)-1]\ \ln\left[\frac{\mathrm{Pr(True}/s)}{1-\mathrm{Pr(True}/s)}\right]}{\mu(s)\theta(s)} + \mathrm{RT}_0.$$

If $\theta(s) = 2\alpha^2\mu(s)$, where α^2 is a scale constant, then simple algebraic manipulation yields:

$$\mu(s) = \sqrt{\frac{[2\mathrm{Pr(True}/s)-1]\ \ln\left[\frac{\mathrm{Pr(True}/s)}{1-\mathrm{Pr(True}/s)}\right]}{2(\overline{\mathrm{RT}}-\mathrm{RT}_0)}}\Bigg/\alpha. \quad (3)$$

The sign of the discriminability estimate is determined by the relative positions of the mean and the referent, usually being positive when the sentence is true and negative when the sentence is false.

This equation yields an estimate of the mean step size for the random walk up to an arbitrary scale factor, $(\sqrt{2}\alpha)^{-1}$. If a suitable value for the residual mean RT, namely RT_0, cannot be obtained, and only an estimate of stimulus discriminability accurate to a monotonic transformation is required—e.g., in an analysis of variance—then RT_0 can safely be ignored. Note that the discriminability index has no absolute meaning but is a relative measure of the signal/noise ratios for each kind of instance-category pair.

The derivation of Equation 3 depends on an approximate solution of a complex random walk equation, as is explained in Link and Heath (1975). In fact $\mu(s)$ will tend to underestimate the true drift rate of the random walk (Vickers & Smith, 1985). Nevertheless, this discrepancy is unlikely to affect the relative values of $\mu(s)$ computed for the various conditions in this research. Provided that the step size variance is sufficiently small, the discrepancy will be negligible.[1]

The mean step size for this random walk depends on the mean difference between the relational strength of the instance-category pair for a given trial and the stored reference level of associative strength. The discriminability index is computed for each type of instance-category presentation—e.g., true-high—and in no way does it involve all instance-category types. Indeed, there is no need to assume that the random walks generated by the various types of instance-category stimuli are symmetric. A general asymmetric version of relative judgment theory is quite appropriate, and a more general parameter estimation procedure is provided by the program FITTRW (Heath, 1983).

The setting of the subjective referent may be illustrated by reference to one of the context experiments of McCloskey and Glucksberg (1979). McCloskey and Glucksberg (1979, Experiment 2) embedded true-high and true-low sentences in a context of false-low sentences in one condition, the *unrelated false condition*, and embedded the same true sentence sets in a context of false-high sentences in another condition, the *related false condition*. McCloskey and Glucksberg (1979) found that the sentences in the unrelated false condition were categorized significantly faster than the sentences in the related false condition, a context effect. In a condition in which true sentences are embedded in a context of false-high sentences, the subjective referent distribution should move in a position direction (i.e., to the right in Figure 1). This implies that the discriminability measures for the true-low and true-high items in the related false condition should be less than the discriminability measures for these items in the unrelated false condition. Hence, according to this model, the McCloskey and Glucksberg (1979) context effect is explained by an adaptive shift in the subjective referent, which represents a change in the null point of the relational strength scale due to experience.

Predictions of the Subjective Referent and Property Comparison Models

Like McCloskey and Glucksberg's (1979) property comparison model, the subjective referent model predicts a context effect.The subjective referent model, however, unlike the property comparison model, allows error RTs to be longer than correct RTs. Vickers' (1980) review of the relationship between correct and error RTs showed that, for instructions emphasizing caution, and also for slow responders when the instructions stress both speed and accuracy, error RTs are longer than correct RTs. In general, in tasks in which imprecise instances require a relatively long categorization time, error RTs should be longer than correct RTs. The freedom allowed for variations in the relationship between correct and error RTs is the distinguishing mark of the subjective referent model in comparison with the property comparison model.

In the property comparison model, the subject accumulates log likelihood ratio evidence—that is, the logarithm of the ratio $p(T/E)/p(F/E)$, the probability of a true response divided by the probability of a false response, given the available evidence E—so that errors are possible and the technology developed by Laming (1968) can be employed to predict the relationship between RT and response accuracy. Laming's model serves as a suitable formulation of the property comparison model, but relative judgment theory is more general than Laming's random walk model, since it can accommodate a wider variety of predictions. (Note that the property comparison model may appear to be an accumulator model in which evidence is accumulated in two separate counters. However, evidence toward one response reduces the probability of the alternate response, and hence there are not two independent counters.)

One other difference between the subjective referent model and the property comparison model concerns the expectation of an interaction effect between sentence type and context. True-high sentences are categorized more quickly than true-low sentences, an effect termed variously as the *relatedness effect for positives* or the *typicality effect*. McCloskey and Glucksberg (1979) predicted an increase in the size of the typicality effect when the distracting false sentences were made more related. However, reference to Figure 1 shows that for the subjective referent model the distance between the criterion distribution and the true-high and true-low distributions decreases as the criterion distribution is moved to the right. This decrease in distance, which represents a decrease in discriminability, is equal for both the true-high and the true-low items. When false-high items are included in the list in place of false-low items, the criterion distribution moves closer to the other distributions, discriminability is reduced, and hence a context effect occurs; but since this reduction in discriminability is the same for both the true-high and true-low item sets no interaction occurs. These are not necessarily competing predictions, since the McCloskey and Glucksberg model is tested by correct RTs, whereas the subjective referent model is tested by the discriminability measure, which is calculated from both mean RT and error rate. The interaction may be sensitive to the choice of dependent variable.

Fits of the Subjective Referent Model to Data from McCloskey and Glucksberg (1979)

The mean RTs and error rates for the true-high and true-low items in the unrelated false and related false conditions of Experiment 2 of McCloskey and Glucksberg (1979, Table 2, p. 21), together with the estimated discriminability scores calculated as in Equation 3, are shown in Table 1. (The discriminability index is calculated usually with overall mean RT, whereas the reported McCloskey and Glucksberg mean RTs were for correct responses only. In this approximate analysis, we assume that the mean RTs for correct and erroneous responses are equal.)

There was a context effect for the discriminability data, indicated by the lower discriminability scores for the sentences in the related false condition relative to the unrelated false condition. However, there was no interaction for the discriminability data, the size of the relatedness effect for positives (the term used by McCloskey and Glucksberg as the equivalent of the *typicality* effect) being approximately the same for the related false as for the unrelated false condition. In contrast, the relatedness effect for positives measured by RT alone showed an interaction, there being a much larger relatedness effect for the related false condition than for the unrelated false condition. Hence, since the RT data fit the interaction prediction of the property comparison model and the discriminability data fit the subjective referent model, these data were insufficient to discriminate between the goodness of fit of the models.

The relative lengths of the correct and error RTs therefore become crucial in distinguishing between the models. McCloskey and Glucksberg (1979) reported the error RT data for the 8 subjects in the related false condition. Correct response RTs for true and false were 1,216 and 1,340 msec, respectively, while the corresponding error RTs were 1,383 and 1,391 msec, respectively, with the differences not being significant. Since 4 of the 8 subjects had faster RTs for errors, while the other 4 subjects were slower, McCloskey and Glucksberg (1979) interpreted their data as supporting the prediction of the property comparison model that, for a given type of response, RTs should be the same for correct and incorrect responses. However, for no subject did RT for correct responses equal RT for incorrect responses. This is a crucial point, and it is difficult to see how the error RT data can be interpreted as supporting the property comparison model for individual subjects.

The Discriminability Index and Error Rates

Concern has already been expressed over the reported but unanalyzed error rates from many semantic memory categorization experiments, since both error rate and RT should reflect the operation of the same categorization process. The discriminability index offers an attractive alternative to RT as the dependent variable in speeded categorization tasks. The measure can accommodate speed-accuracy tradeoffs, since the discriminability index should be invariant with changes in the selection of the response threshold under instructions to vary overall mean RT. It serves as a measure of performance in sensitive situations in which both error rate and RT are higher in one condition than in another, but in which the RT differences are not sufficient by themselves to demonstrate significant effects.

Data from McCloskey and Glucksberg's (1979) Experiment 1 illustrate the potential of the discriminability index for utilizing mean RT and error rates to detect effects that may not be obvious when only mean RT for correct responses is used. McCloskey and Glucksberg failed to find the robust relatedness for negatives effect in the highly related false condition. The mean correct RT and error rate data for their negative items are shown in Table 2. The discriminability estimates for these data, calculated using Equation 3, are given in the third column of the table.

The discriminability index is calculated usually with overall mean RT, whereas the McCloskey and Glucksberg (1979) mean RTs are for correct responses only. For illustrative purposes, it was assumed that mean RTs for correct and erroneous responses were equal, so that Equation 3 could be employed to compute the discriminability index. Although no tests for significance could be carried out because data on variability within and between subjects were unavailable, the discriminability scores are nonetheless in the order normally found for negatives. The disjoint sentences are the most discriminable, then the low-related category-member sentences, with the high-related category-member sentences being the least discriminable. Hence, the discriminability measure suggested the presence of a well-established effect that was not detected by the RT data alone.

The Discriminability Index and FITTRW

A more complex and rigorous alternative to calculating a discriminability index is to estimate the mean step size or drift of the random walk generated by a particular set of stimuli, using the FITTRW estimation program (Heath, 1983). FITTRW was designed to estimate parameters of a general class of random walk models for

Table 1
Mean RTs (in msec) and Error Rates from Experiment 2 of McCloskey and Glucksberg (1979), and Estimated Discriminability Scores

	List Condition					
	Unrelated False			Related False		
Sentence Type	RT	Error Rate	Discriminability	RT	Error Rate	Discriminability
true-high	877	1.3%	1.55	1,062	4.6%	1.14.
true-low	1,022	6.3%	1.07	1,369	12.9%	0.72
Size of Effect	145		.48	307		.42

Table 2

Mean RT (in msec), Error Percentages (EP), and Discriminability Estimates (D) for the Negative Sentences of the Highly Related False Condition of McCloskey and Glucksberg's (1979) Experiment 1

Negative Sentence Type	RT	EP	D
Disjoint false sentences (e.g., "All shoes are birds")	1,037	1.7%	−1.37
High-related superset false sentences (e.g., "All birds are sparrows")	1,029	6.7%	−1.05
Low-related superset false sentences (e.g., "All birds are geese")	1,074	3.3%	−1.21

two-choice response time. It uses the STEPIT function minimization routine (Chandler, 1975), and requires response proportion and mean RT for both correct and error responses, with standard errors of the mean RTs being desirable. The parameters that are estimated include the response threshold, the subject's response bias, the mean step size of the random walk, and the residual mean RT (nondecision components). An estimate of the step size moment generating function asymmetry, γ, is also provided. If $\gamma < 1$, mean RTs for errors are longer than mean RTs for correct responses. If $\gamma > 1$, then mean RTs for errors are shorter than mean RTs for correct responses. If $\gamma = 1$, then mean RTs for errors are equal to mean RTs for correct responses. For example, if the random walk steps are normally distributed, then $\gamma = 1$ and mean RT for errors equals mean RT for correct responses.

If the goodness of fit of the FITTRW parameter estimates is acceptable, then one may compare the estimates of the mean drift rates obtained for particular item sets with the discriminability measures derived using Equation 3. If the drift rate parameter estimates and the corresponding discriminability measures are highly correlated, then one may confidently use the computationally simpler discriminability measure. The data from the experiments of McCloskey and Glucksberg (1979) cannot be fit with FITTRW, since their summary data do not include mean RTs for error responses.

EXPERIMENTAL EVALUATION OF THE MODELS

In the experiment that follows, data from a partial replication of McCloskey and Glucksberg's (1979) Experiment 2 are reported, which enabled testing of the predictions of both the property comparison model and the subjective referent model regarding the relative durations of error and correct RTs. The McCloskey and Glucksberg context effect paradigm was chosen because of the wide range of relatedness of subjects and predicates in the test items, ensuring a useful range of error rates for model evaluation. Moreover, McCloskey (1980) reported an absence of control for familiarity in the positive items used in McCloskey and Glucksberg's (1979) Experiments 1 and 2, whereas we found a significant variability in familiarity in the items used as negatives in their Experiment 2. Hence, these experiments offered an opportunity to test for context effects with appropriate control for item familiarity. In order to obtain a greater amount of data for testing the correct versus error RTs hypotheses, and to increase the generality of the context effect, we extended the number of categories from 15 to 28, each of these 28 categories being able to be classified as either *natural* (categories of objects that exist independently of human activities) or *functional* (categories of objects that are products of human activities). This division of categories enables the exploration of possible differences in the categorizing of members of these two category classes.

Summary of Predictions of the Subjective Referent and Property Comparison Models

This experiment replicated the McCloskey and Glucksberg (1979) context effect experiment, with item familiarity controlled across levels of relatedness. A major prediction of the subjective referent model, as for the property comparison model, was that there would be a context effect. However, the subjective referent model predicts additivity for the discriminability index rather than an interaction between typicality and context. Furthermore, the subjective referent model allows mean error RTs to be longer than mean correct RTs. Longer error RTs might be expected for items to which subjects respond more slowly, such as items low on typicality.

Method

Subjects. The subjects were 44 introductory psychology students for the RT experiment, plus an additional pool of subjects for a series of rating tasks. The subjects, approximately 75% females, ranged in age from 18 to 30 years.

Materials. In the absence of any known normative data for Australian subjects, sets of categories and instances were developed as follows. Nineteen subjects generated category names. These subjects then classified categories as being *natural*, *functional*, or *other*. The 14 categories most commonly rated as natural categories were: tree, flower, bird, vegetable, metal, landform, mammal, gas, gemstone, reptile, fish, fruit, insect, and rodent. The 14 categories most commonly rated as functional categories were: furniture, clothes, dwelling, tool, container, seafood, fuel, artwork, drink, drug, ornament, sport, weapon, and vehicle. Fifteen subjects then generated instances of these categories. False-high lists of nonmembers of each category were also prepared by the experimenters. Separate sets of nonmembers for each category were also prepared by scrambling the false-high probes and allocating them at random to categories to form false-low probes.

Ratings for category members on typicality, nonmembers on relatedness, and all words, including category titles, on familiarity were carried out by groups of 10 subjects. Test sets were then composed for four context conditions, namely unrelated false and related false for natural categories, and unrelated false and related false for functional categories. Note that the terms *unrelated false* and *related false* refer to experimental conditions, while the terms *true-high*, *true-low*, *false-high*, and *false-low* refer to sentence types *within* a condition. The organization of conditions and sentence types is shown in Table 3.

The unrelated false and related false conditions were structured so that the positive items were the same for both unrelated false and related false conditions; the negative items were also the same for each condition, but they were highly related for the related false conditions (e.g., bat-bird, wheel-vehicle) and scrambled so as to

Table 3
Test Sets for Negative Contexts

Category Type	Condition Unrelated False	Related False
Natural	true-high	true-high
	true-low	true-low
	false-low	*false-high*
Functional	true-high	true-high
	true-low	true-low
	false-low	*false-high*

be lowly related for the unrelated false conditions (e.g., bat-flower, wheel-dwelling). For each condition, test sets consisted of 56 true and 56 false sentences, and each subset of true items contained 28 true-high and 28 true-low items. The measures obtained for the experimental control for typicality, familiarity, relatedness, and word length are given in Table 4.

There were no significant differences between groups on factors that had to be balanced, namely familiarity and word length across all items, typicality for positive items, and relatedness for negative items. However, typicality levels of true-high items were significantly higher than those of true-low items ($p < .001$), and relatedness levels of false-high items were significantly higher than those of false-low items ($p < .001$). Hence, the stimuli were relatively homogeneous for word familiarity and word length, but they differed in typicality, which was an independent variable in this study.

In addition to the test sets, practice sets of 32 items were composed, which contained the same proportions of true and false items and the same levels of typicality and familiarity as the test sets.

Design and Procedure. Category type (natural and functional) and relatedness of the negatives (unrelated false and related false) were between-subjects factors; typicality (true-high and true-low) was a within-subject factor. Each of the 44 subjects was allocated at random to one of the four test conditions: unrelated false natural, unrelated false functional, related false natural, and related false functional.

Each subject was seated before the video display unit of an Apple II microcomputer. The response keys on each side of this unit were operated with the index fingers. Response keys were balanced for handedness, so that for 50% of the subjects a true response was signaled with the dominant hand, whereas the reverse mapping applied to the other subjects. The subjects, who were seated approximately 30 cm from the screen, were told that a fixation point would appear on the screen for 2 sec. It would then be replaced by a pair of words, such as *chair-furniture*, which was to be interpreted as the assertion, "All chairs are furniture." The subjects were to affirm or deny the assertion by pressing the appropriate response key. They were advised to respond as quickly and as accurately as possible. This response was followed immediately by feedback on the accuracy of the response, "correct" or "wrong" appearing on the screen. The appropriate 32-item practice set for each condition was given at the start of the session, and was followed by the test sets for each condition. A rest period of 30 sec was given after each block of 28 test questions. Response latency was measured to the nearest millisecond, with a machine language program.

Results and Discussion

The application of the subjective referent model was evaluated, by means of fitting it to response accuracy and mean RTs for correct and erroneous responses with the parameter estimation program FITTRW (Heath, 1983). Discriminability scores, calculated as shown in Equation 3, and mean correct RTs were also analyzed. Responses to positive items were analyzed separately from responses to negative items. RTs greater than 4 sec (0.5% of the data) were omitted from the analysis but were not counted as errors.

FITTRW. The RT data, uncensored for errors, were pooled over the natural and functional conditions to provide a better estimate of error RT. The FITTRW analysis resulted in an estimate of response threshold, $\hat{A} = 17.74 \pm .064$, and starting point, $\hat{C} = 0.99 \pm 0.50$, the latter being sufficiently close to zero to justify the use of the $C = 0$ assumption in the computation of the discriminability index. The residual mean RT, 258±60 msec, was close to estimates computed for psychophysical tasks (Luce, 1986). The estimate of the step size moment generating function asymmetry, $\hat{\gamma} = 0.90 \pm 0.08$, was less than 1, as was expected with error RTs longer than correct RTs (Link & Heath, 1975).

The estimation of the mean step size for the random walk $\mu(s)$ for condition s, was simplified by the observations that (1) $\theta(s)$ was linearly related $\mu(s)$ with slope 9.01 and intercept −0.0074, and (2) $\mu(s)$ for the unrelated false conditions was linearly related to $\mu(s)$ for the related false conditions, with a slope of 1.39 and intercept −0.09. When FITTRW was rerun with these linear constraints,

Table 4
Control Factors for Item Selection (Ratings for Typicality, Relatedness, and Familiarity on a 1-7 Scale, Word Length as Number of Letters)

	Condition					
	Unrelated False			Related False		
	True-High	True-Low	False-Low	True-High	True-Low	False-High
Typicality						
Natural	6.4	3.9		6.4	3.9	
Functional	6.5	3.9		6.5	3.9	
Relatedness						
Natural			1.1			3.0
Functional			1.2			3.1
Familiarity						
Natural	6.9	6.8	6.9	6.9	6.8	6.9
Functional	6.8	6.8	6.9	6.8	6.8	6.9
Word Length						
Natural	5.6	5.4	5.4	5.6	5.4	5.4
Functional	5.3	5.6	5.4	5.3	5.6	5.4

Table 5
FITTRW Estimates of the Mean Step Sizes (μ) and Standard Errors (*SE*) for Each Sentence Type, Using the Data Pooled Over the Natural and Functional Categories

Sentence	List Condition		
	Unrelated False	Related False	
Type	$\hat{\mu}$	$\hat{\mu}$	*SE*
true-high	.022	.016	.002
true-low	.015	.011	.001
false-high		−.012	.007
false-low	−.018		

Note—Since the estimates for the unrelated false condition were fixed by linear constraint, there are no *SE*s.

Table 6
Observed and Estimated RTs (in msec) for True and False Responses for Each Sentence Type in Each Condition, with the Data Collapsed Across Natural and Functional Categories

	Respond "True"		Respond "False"	
Sentence Type	Obtained	Estimated	Obtained	Estimated
	Unrelated False Condition			
true-high	1,002±43	952	1,309±392	1,115
true-low	1,131±49	1,197	1,344±274	1,421
false-low	1,296±257	1,224	1,204±35	1,205
	Related False Condition			
true-high	1,146±49	1,146	1,362±178	1,357
true-low	1,297±59	1,293	1,543±155	1,564
false-high	1,496±127	1,507	1,448±37	1,447

Note—The error data were obtained from the "True" responses to either false-low or false-high sentences and the "False" responses to true-high or true-low sentences. The 95% confidence intervals for the observed RTs are shown.

the estimates of $\mu(s)$ shown in Table 5 were obtained. Since the estimates for the unrelated false context were fixed by the second linear relation above, no standard errors could be computed.

A test for goodness of fit yielded $\chi^2(10) = 1.92$, n.s. The correlation between the FITTRW estimates of the mean step size or drift of the random walk for each sentence type in the related false and unrelated false conditions and the corresponding discriminability scores was almost perfect. The obtained and FITTRW estimated RTs are shown in Table 6. The only response sets for which the predicted RT values lay outside the 95% confidence intervals for the observed mean RT were true responses for the true-high and true-low sentences in the unrelated false condition.

Discriminability. Mean discriminability scores and their standard errors are shown in Table 7. For the positive items, min *F′* analyses yielded significant context and typicality effects, but no interaction. Specifically, the context effect was significant [min $F'(1,87) = 6.34$, $p < .025$], and the typicality effect was also significant [min $F'(1,81) = 16.41, p < .001$]. There was no effect for category type (min $F' < 1$). The interaction between category type and typicality yielded a nonsignificant min $F'(1,86) = 2.24$, and no other interactions approached significance. The analysis of the data for negatives showed only one significant result, namely a significant main effect for relatedness of the negatives [min $F'(1,69) = 29.21, p < .001$], but no significant effects for category type or interaction of category type and relatedness.

Correct RT. The analysis of the correct RT data for positive items showed a significant main effect for typicality [min $F'(1,90) = 36.74, p < .001$], but no effects for context or category type (min $F' < 1$). The RT data were in the direction predicted for a context effect, but the large variability in the RT data masked the context effect. There were also no interaction effects. The usual relatedness effect was found for the negative RT data [min $F'(1,46) = 6.14, p < .025$], but there was no effect for category type and no interaction. Mean RTs for all groups, with error rates in parentheses, are shown in Table 8.

Error rates. While the RT data failed to reveal a context effect, an analysis of the error rates for the positive items found such an effect [min $F'(1,92) = 4.21$, $p < .05$]. The error rate data also showed the expected typicality effect [min $F'(1,78) = 8.68, p < .005$]. There were no other significant error rate effects for the positive items. The finding of a context effect in the error rate data but not in the RT data emphasized the need to take note of error data. However, at times error rate data alone are insufficient to enable meaningful analysis, and thus the discriminability measure, which employs both error rate and RT data, is recommended.

The error rates for the negative items demonstrated the expected relatedness effect for negatives [min $F'(1,89) = 17.91, p < .001$]. The negative items also revealed

Table 7
Mean Discriminability Scores (*SE*s) for the Unrelated False and Related False Conditions for Natural and Functional Categories

		List Condition			
		Unrelated False		Related False	
Category Type	Sentence Type	*M*	*SE*	*M*	*SE*
Natural	true-high	1.25	.09	1.08	.08
	true-low	1.11	.06	0.84	.08
	false-high			−0.56	.05
	false-low	−1.24	.06		
Functional	true-high	1.30	.10	1.13	.05
	true-low	1.00	.10	0.67	.06
	false-high			−0.83	.05
	false-low	−1.11	.12		

Table 8
Mean Correct RTs (in msec), Standard Errors (*SE*s), and Error Rates for the Unrelated and Related False Conditions for Natural and Functional Categories

		List Condition					
		Unrelated False			Related False		
Category Type	Sentence Type	*M*	*SE*	Error Rate	*M*	*SE*	Error Rate
Natural	true-high	1,039	84	3.6%	1,194	92	5.3%
	true-low	1,184	99	4.0%	1,351	112	10.3%
	false-high				1,537	114	18.9%
	false-low	1,245	97	2.4%			
Fuctional	true-high	993	100	3.3%	1,136	99	4.6%
	true-low	1,136	86	8.9%	1,294	92	16.3%
	false-high				1,466	77	8.5%
	false-low	1,227	119	6.3%			

an interaction such that the increase in errors from false-low to false-high sentences was significantly greater for the natural than the functional categories [min $F'(1,85) = 6.24$, $p < .025$]. The different patterns of error rates between the natural and functional categories, shown in Table 8, suggest that subjects adopted a stricter criterion of membership for functional than for natural categories. However, the generality of this finding must be questioned, for McCloskey and Glucksberg (1978) have demonstrated both between-subjects disagreement and within-subjects inconsistency for the categorization of items of uncertain membership.

Error RT. While a variety of sequential processing models may describe the decision process, the error data are crucial in deciding which random walk model best fits the data. There were 11 subjects in each of the four conditions, and three sentence types in each condition, making a total of 132 data sets, 66 from the unrelated false conditions, and 66 from the related false conditions. Errors occurred in 47 of the 66 unrelated false data sets. In 25 of these sets, mean correct RTs were slower than mean error RTs, and in 22 data sets, the reverse occurred. In contrast, for the related false conditions, in which errors occurred in 58 of the 66 data sets, mean error RTs were slower than mean correct RTs in 42 data sets, but the reverse occurred in only 16 data sets ($p < .05$, using a binomial test).

When the overall mean RTs for each sentence type in each condition (three sentence types in each of four conditions) were calculated, error RTs were found to be longer than correct RTs in all 12 data sets. The differences in mean correct and error RTs for each sentence type in each condition were tested for significance levels by t tests that included the Satterthwaite solution to the Behrens-Fisher problem of heterogeneity of variance (Howell, 1982, p. 137). This resulted in significant differences for four of the data sets. The mean correct and error RTs, together with standard errors, and the error rates are shown in Table 9, for all 12 data sets. Significantly different mean RTs are indicated.

The McCloskey and Glucksberg (1979) model predicts no difference between mean correct and error RTs, whereas the subjective referent model based on relative judgment theory (Link & Heath, 1975) can accommodate the result that as more caution is exercised, the mean RTs for errors will be longer than correct RTs. The latter is clearly the case here.[2]

The replication of McCloskey and Glucksberg's (1979) Experiment 2 produced two important results, one fortuitous and the other predicted. The fortuitous result was the failure of the RT data alone to detect a significant context effect, an effect revealed by the use of the discriminability measure, which combined both error and RT data. Furthermore, the discriminability measures correlated

Table 9
Mean Correct (*C*) and Error (*E*) RTs (in msec), Standard Errors (*SE*), and Error Rates for all Conditions

		List Condition									
		Unrelated False					Related False				
Category Type	Sentence Type	*C*	*SE*	*E*	*SE*	Error Rate	*C*	*SE*	*E*	*SE*	Error Rate
Natural	true-high	1,037	22	1,305	205	3.6%	1,177	25	1,513	83‡	5.3%
	true-low	1,161	26	1,382	188	4.0%	1,329	30	1,523	86	10.3%
	false-high						1,473	20	1,501	57	18.9%
	false-low	1,229	17	1,316	176	2.4%					
Functional	true-high	967	23	1,314	194*	3.3%	1,115	24	1,188	99	4.6%
	true-low	1,100	24	1,327	91†	8.9%	1,263	29	1,556	71‡	16.3%
	false-high						1,426	17	1,484	73	8.5%
	false-low	1,178	18	1,289	86	6.3%					

Note—These grand mean correct RTs differ slightly from those in Table 8, which are the means of the subjects' means. $^*p < .05$. $^\dagger p < .01$. $^\ddagger p < .001$.

highly with the FITTRW estimates of the mean step size of the random walk. Although context effects may occur for RTs, errors, or both, it seems more desirable to have one measure that detects such an effect, especially if this discriminability index amplifies the effect in the raw RT and error measures. The predicted result concerned the fact that the mean RTs for errors were longer than the mean RTs for correct responses. The longer RTs for errors supported the relative judgment version of the random walk model rather than the McCloskey and Glucksberg (1979) model, which is similar to the sequential probability ratio test model of Laming (1968). The latter model cannot account for low errors without substantial modification (Heath, 1981).

Process versus structure models. Kounios, Osman, and Meyer (1987) recently proposed a new methodology, speed-accuracy decomposition, for collecting and analyzing data from semantic memory categorization tasks. A major aim was to develop a method to assist in deciding whether search or the gradual accumulation of information better represented the way the sentence verification task was carried out. Their data favored models that opted for a gradual accumulation of information (such as the model in McCloskey & Glucksberg, 1979). Moreover, such models usually posit a relatively unstructured memory base. Hence the subjective referent model has received recent empirical support for its assumptions regarding both the structure of memory and the manner in which information is retrieved from memory. Moreover, this model is an improvement on McCloskey and Glucksberg's (1979) property comparison model, which Kounios et al. (1987) cite as typifying the gradual accumulation of information approach, because the subjective referent model allows error RTs to be longer than correct RTs.

CONCLUSION

A model has been presented for the analysis of semantic memory categorization data in terms of Link and Heath's (1975) random walk model of discrimination judgments. The model belongs to that class of models that require no strong assumptions about the structure of memory, and it is focused mainly on decision processes. The model is primarily distinguished from the McCloskey and Glucksberg (1979) model in terms of its prediction that error RTs can exceed correct RTs. The analysis of the categorization task in terms of a random walk model allows the derivation of a theoretically based discriminability index, which combines both RTs and error rates and makes no assumptions about the step size distribution—that is, it is a distribution-free estimate, unlike the d' of signal detection theory. The experimental results have illustrated the utility of employing a measure that combines both error rates and RT. Although a difference between discriminability and RT results is not in itself a recommendation for the superiority of the discriminability measure, the McCloskey and Glucksberg context effect was not apparent when only correct RT data were used, but it was verified with the discriminability measure. Hence, the applicability of this measure in analyzing sentence verification RT data is to be recommended. The subjective referent model seems applicable in a wide range of discrimination tasks, whether psychophysical, lexical, or semantic, where speed and accuracy either may be traded or may covary. It is based on more general assumptions than those of the property comparison model, and it can handle a wider range of findings. A task for the future is to compare results from the discriminability model with those from a recently developed, tractable version of the accumulator model (Heath, 1984; Vickers & Smith, 1985), and to apply these stochastic decision models in a variety of cognitive tasks.

REFERENCES

Battig, W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology Monographs*, **80** (3, Pt. 2).

Chandler, J. P. (June, 1975). *Subroutine STEPIT* (Quantum Chemistry Program Exchange No. 307). Stillwater: Oklahoma State University, Computer Science Department.

Chang, T. M. (1986). Semantic memory: Facts and models. *Psychological Bulletin*, **99**, 199-220.

Cox, D. R., & Miller, H. D. (1965). *The theory of stochastic processes*. London: Methuen.

Heath, R. A. (1981). A tandem random walk model for psychological discrimination. *British Journal of Mathematical & Statistical Psychology*, **34**, 76-92.

Heath, R. A. (1983). FITTRW: A parameter estimation program for a general random walk model analysis of two-choice response time (2CRT) data. *Behavior Research Methods & Instrumentation*, **15**, 95-96.

Heath, R. A. (1984). Random walk and accumulator models of psychophysical discrimination: A critical evaluation. *Perception*, **13**, 57-65.

Howell, D. C. (1982). *Statistical methods for psychology*. Boston: Duxbury Press.

Kounios, J., Osman, A. M., & Meyer, D. E. (1987). Structure and process in semantic memory: New evidence based on speed-accuracy decomposition. *Journal of Experimental Psychology: General*, **116**, 3-25.

Laming, D. (1968). *Information theory of choice reaction time*. New York: Wiley.

Link, S. W. (1978). The relative judgment theory of the psychometric function. In J. Requin (Ed.), *Attention and performance* VII (pp. 619-630). Hillsdale, NJ: Erlbaum.

Link, S. W., & Heath, R. A. (1975). A sequential theory of psychological discrimination. *Psychometrika*, **40**, 77-105.

Loftus, E. F. (1973). Category dominance, instance dominance, and categorization time. *Journal of Experimental Psychology*, **97**, 70-74.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

McCloskey, M. E. (1980). The stimulus familiarity problem in semantic memory research. *Journal of Verbal Learning & Verbal Behavior*, **19**, 485-502.

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, **6**, 462-472.

McCloskey, M. E., & Glucksberg, S. (1979). Decision processes in verifying category membership statements: Implications for models of semantic memory. *Cognitive Psychology*, **11**, 1-37.

Pachella, R. G. (1974). The interpretation of reaction times in information-processing research. In B. H. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition* (pp. 41-82). Hillsdale, NJ: Erlbaum.

Rosch, E. H. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language* (pp. 111-144). New York: Academic Press.

Shoben, E. (1982). Semantic and lexical decisions. In C. R. Puff (Ed.), *Handbook of research methods in human memory and cognition* (pp. 287-314). New York: Academic Press.

Vickers, D. (1980). Discrimination. In A. T. Welford (Ed.), *Reaction times* (pp. 25-72). London: Academic Press.

Vickers, D., & Smith, P. (1985). Accumulator and random walk models of psychological discrimination: A counter-evaluation. *Perception*, **14**, 471-497.

NOTES

1. A simulation of 1,000 trials of a random walk with normally distributed steps showed that, for a mean step size, μ, equal to 0.10, and variance equal to 1, the discrepancy between observed and predicted μ was negligible, provided $A > 10$. In the FITTRW analysis that follows, the approximation was found to be quite adequate, since $A = 17.74$.

2. We performed a partial replication of this experiment, using the same test items and a different pool of 44 subjects, the one change being that subjects received no feedback on the accuracy of their responses. Again, error RTs were consistently longer than correct RTs. An interesting feature of the data was the absence of a context effect in both the discriminability and RT results, a result attributable to the absence of feedback.

(Manuscript received January 21, 1985;
revision accepted for publication November 7, 1988.)