Perception & Psychophysics
1989, 45 (6), 557-566

The psychophysics of numerical comparison: A reexamination of apparently incompatible data

STANISLAS DEHAENE
Laboratoire de Sciences Cognitives et Psycholinguistique, Paris, France

Reaction-time studies of numerical comparison have used essentially two paradigms: *classification*, in which a target number must be labelled "larger" or "smaller" in comparison to a fixed standard, and *selection*, in which the larger (or smaller) number of a pair must be picked out. In previous studies, classification has yielded only a distance effect in RTs, whereas selection has also revealed *magnitude* (or *minimum*) and *congruity* effects. We used two experiments with two-digit number comparisons to find the reason for this discrepancy. In Experiment 1, we used a variant of the classification task with the standard changing on each trial. RTs increased along with the standard for "smaller" responses and decreased along with the standard for "larger" responses, in a manner reminiscent of magnitude and congruity effects. In Experiment 2, we again used classification, but the fixed standard 75 was not at the center of the range of target numbers (20, 21, ... 99). Close to the standard, RTs were faster for "larger" than for "smaller" responses, again a congruity effect. Our data show that magnitude and congruity effects can be obtained with two-digit numbers in classification as well as in selection tasks. A single equation, which implies that numbers are compared with respect to reference points at both ends of the continuum, describes the results from both tasks.

We have little insight into the algorithm that enables us to compare two objects. In information processing accounts, the "comparator" usually appears as a little black box whose performance is not open to analysis. The idea that comparison is an atomic operation is partly rooted in the computer metaphor of the mind. Designers of computers have found it useful to implement comparison as one of the fastest and the most elementary of operations. But, is our ability to compare actually a primitive operation of the brain, or can it be broken down into smaller parts?

In the last 20 years, some careful experiments have challenged the atomicity of comparison—in particular, numerical comparison. Obviously, the manipulation of numbers is a cognitive process; yet in comparison tasks, digits (Moyer & Landauer, 1967; Restle, 1970) as well as two-digit numbers (Dehaene, Dupoux, & Mehler, 1989; Hinrichs, Yurko, & Hu, 1981) appear to be represented in an analogical, perceptual-like fashion. However, unlike perceptual processes, the numerical system has the important property of being symbolic, and thus unambiguous: the effects found cannot be attributed to peripheral factors such as noise, masking, thresholds, and so forth. In contrast to many cognitive processes, numerical comparison is formally simple, and it has been unambiguously defined. This is not the case, for example, in the comparison of the size of objects designated by their names (Holyoak, 1977; Kosslyn, Murphy, Bemesderfer, & Feinstein 1977; Moyer, 1973).

Essentially, two tasks have been used to study numerical comparison. They will hereafter be referred to as *selection* and *classification*. In selection, two numbers are presented visually, and the subject has to press one of two keys to indicate which number is the larger (or the smaller). In classification, a single target number is presented, and the subject indicates with one of two response keys whether it is larger or smaller than a standard of reference. The numerical standard is usually stored in memory throughout the experiment.

In both tasks, response latencies decrease as the distance that separates the items to be compared increases. This pattern of response is called the *distance effect*. First discovered in the case of digits by Moyer and Landauer (1967), it has been reproduced by several experimenters (e.g., Banks, Fujii, & Kayra-Stuart, 1976; Buckley & Gillman, 1974; Parkman, 1971; Sekuler & Mierkiewicz, 1977; Sekuler, Rubin, & Armstrong, 1971). In fact, the distance effect appears not only in the comparison of digits, but also with all sorts of materials: two-digit numbers (Dehaene et al., 1989; Hinrichs et al., 1981; see also Restle, 1970), bars of varying length (Johnson, 1939), dot arrays compared for numerosity (Buckley & Gillman, 1974), objects compared for size (Holyoak 1977; Kosslyn et al., 1977; Moyer, 1973), and still others. With numbers, reaction times (RTs) have often been found to be proportional to the logarithm of distance (log D func-

I gratefully acknowledge the constant support of Emmanuel Dupoux, Peter Jusczyk, and Jacques Mehler during this research. Jean-Luc Aucouturier and Susana Frank provided technical help. Two anonymous reviewers made helpful comments. This work was supported by grants from the CNRS (ASP Processus cognitifs en jeu dans la production et la compréhension du langage) and the Fondation pour la Recherche Médicale. Address reprint requests to Stanislas Dehaene, Laboratoire de Sciences Cognitives et Psycholinguistique, 54 Bd Raspail, 75270 Paris Cédex 06, France.

tion). However, there has been an occasional report of a linear distance effect (Sekuler & Mierkiewicz, 1977).

In agreement with the apparent similarity of classification and selection, the distance effect is invariably found in both types of tasks. Surprisingly, however, the analogy between the two tasks appears to stop there. Two other effects have been found in selection and have not been reported in classification. One is the *minimum effect* (Banks et al., 1976; Buckley & Gillman, 1974; Parkman, 1971): for equal distance, comparison times vary with the lesser of the two numbers. The term "minimum effect" is rather misleading, because when distance is kept constant, the minimum and maximum are perfectly correlated, and the minimum effect could thus just as well be termed the "maximum effect." To avoid this difficulty, the term *magnitude effect* is used throughout this paper. For equal distance, the finding is indeed an effect of the magnitude of both operands.

The magnitude effect is often described as a linear relation of RTs and the minimum of the digits. However, some descriptions of comparison times include a logarithmic term for it. This is the case with the Welford (1960) function

$$RT \alpha \log \frac{L}{L-S} \quad (1)$$

(where L and S are respectively the larger and the smaller of the digits to be compared), which has been used to describe data from psychophysical and numerical comparison tasks (Moyer & Landauer, 1967). This function includes, in addition to the logarithmic distance effect, a magnitude effect in the form of a logarithmic relation of RTs to the larger digit.

The other effect found in selection tasks is the *congruity effect* (Banks et al., 1976). In its purest form (the *crossover effect*), the effect is simply that responses are faster when internal and response codes are congruent; thus, large digits (say, 7 and 9) are compared faster in the "choose larger" condition than in the "choose smaller" condition, and the reverse is true for small digits (say, 2 and 4). In general, the congruity effect is superimposed on top of other effects (for example, the fact that the "choose smaller" condition is often slower than the "choose larger" condition). This may preclude a full crossover effect. In this case, the congruity effect may take only a funnel form and appear as an interaction between the instructions used (to choose the larger or to choose the smaller) and the magnitude of the operands. The effect has been reported with a diversity of materials (e.g., by Audley & Wallis, 1964, and Jamieson & Petrusic, 1975).

Very few studies, if any, have contained reports of magnitude or congruity effects in classification. The typical RT curve in a classification task is a fairly symmetrical logarithmic function of distance (Dehaene et al., 1989; Hinrichs et al., 1981). Thus it is not the case, as the magnitude effect seems to imply, that at equal target/standard distances, "smaller" responses are faster than "larger" ones. For instance, in classification with Standard 5, Targets 3 and 7 are responded to at about the same speed; yet in a "choose smaller" selection task, (3, 5) is faster than (5, 7). In other words, the Welford function, with its implicit term for the magnitude effect, is always a worse predictor of classification times than is a simple log D function (Dehaene et al., 1989; Hinrichs et al., 1981).

There is no report of a congruity effect in classification either. However, it is not clear what the congruity effect would look like in a classification task. To test the influence of congruity implies the examination of the same stimuli under two different conditions of response. Yet in classification with a fixed standard, a given number always elicits the same response—either "larger" or "smaller"—depending on its position with respect to the standard. Thus, this paradigm may not offer an opportunity to assess congruity.

The aim in this paper is to find connections between selection and classification tasks, despite their apparent dissimilarity. First, two experiments are described. They represent slight departures from the standard classification paradigm. A few modifications of the standard procedure will permit reproduction of magnitude and congruity effects within classification. I shall then turn to theoretical models of numerical comparison, to see whether any of them predicts or explains the results.

EXPERIMENT 1

The discrepancy between the results of experiments with classification and those of experiments with selection tasks may stem from a difference in the sampling of stimulus space. In selection, all possible couples of stimuli (e.g., digits) are usually proposed for comparison. In contrast, in classification, the standard of comparison is usually fixed throughout the experiment, thus severely restricting which couples of numbers can be tested. Such a design is useful with two-digit numbers for instance, in which case it would be impractical to test all couples of stimuli. Nevertheless, it has one shortcoming: within one set of responses (e.g., "larger" responses), magnitude and distance are perfectly correlated. An intuitive way to express this is to note that in classification, "large" numbers always receive a "larger" response, and "small" numbers a "smaller" response. There is no room for the congruity effect if targets and responses are always congruent. However, were one to separate the two factors, would magnitude and congruity effects emerge? In Experiment 1, a possible way to make targets and responses incongruent in classification is examined. A variable-standard classification procedure is used: a new standard of comparison is presented before each trial. Possible standards span the 35–75 interval. Thus some operands, like

37 with Standard 35, are small even though the correct response is "larger." A congruity effect may be expected in such conditions.

Method

Procedure. A random list of two-digit standards and targets was constituted for each trial. The standards were chosen from the numbers 35, 45, 55, 65, and 75. For each standard, all the numbers at a distance of 24 and under were presented as targets. Thus, each list contained 240 different couples (standard/target) that were presented once, plus an initial training list of 20 couples. Each list was controlled so that the same target did not appear twice in a row. The order of the standards was fully random.

The temporal presentation of the stimuli was as follows: After a 1,500-msec blank, an empty frame appeared at the center of the screen, together with the standard. One group of subjects viewed the standard *above* the frame throughout the experiment, while the other group viewed it *under* the frame. After 400 msec, a target number appeared in the frame. Neither the frame nor the standard was erased. The subject's response was then recorded via two Morse keys with a 1-msec precision, over a period of 1,500 msec. The whole display disappeared 2,000 msec after the target first appeared, and a new testing cycle began. The presentation rate was thus one test couple every 3,900 msec. The whole experiment lasted less than 20 min.

Subjects. Twenty right-handed French students, aged between 18 and 30, were tested individually. All were naive in numerical comparison tasks. Ten subjects viewed the standard above the frame, the other 10 under it.

Instructions. The subjects were told that they would see a frame with a two-digit number (35, 45, 55, 65, or 75) figuring above or below it. After some time, another two-digit number would appear inside the frame. The subjects were told to press the right-hand button if the number in the frame was larger than the one above or below, and the left-hand button if it was smaller. The instructions emphasized the need to answer fast while avoiding errors.

Results

Distance effect. No effect of the position of the standard's being above or below the frame was found. Accordingly, the two groups of subjects were mixed. The data were averaged across subjects for each value of target and standard. The resulting RT curves for each standard are shown in Figure 1. In all 10 cases, RTs decreased significantly ($p < .001$) as the absolute difference between target and standard decreased (distance effect). The overall correlation with log D, the natural logarithm of this difference, reached $r = -.95$ ($p < .001$). Errors followed the same tendency ($r = -.56$, $p < .01$).

Influence of units. The difference between the observed RT and the mean RT of the decade was computed for each target number outside the decade of the standard. Difference scores for targets ending with the same ones digit were then averaged together, separately for "larger" and "smaller" responses. Finally, to reduce variability, the curve for "larger" responses was combined with the curve for "smaller" responses by means of averaging together RTs corresponding to symmetrical ones digits: RT (1, larger) with RT (9, smaller); RT (2, larger) with RT (8, smaller); and so forth. The resulting units curve presented a significant increase with the ones digits ($r = .87$, $p < .005$). Thus there was a substantial contribution of units within decades to the distance effect.

Magnitude and congruity effects. An analysis of variance on the subjects' mean RTs was performed with standard and response type ("larger" or "smaller") as within-subject factors. The ANOVA revealed a marginally significant influence of standard ($F = 2.51$, $p < .05$) and a considerable interaction between standard and response type ($F = 13.0$, $p < .001$), but no influence of response type per se. The means for each condition are plotted in Figure 2. It can be seen that when the subjects responded "smaller," the RTs increased regularly along with the standard. When they responded "larger," the RTs decreased along with the standard. It is also clear that "larger" RTs changed less along with the standard than did "smaller" RTs. This was assessed statistically by computing the absolute values of the slopes of evolution of the mean RTs along with the standard, separately for "larger" and "smaller" responses and for each subject. The slopes for "larger" responses were significantly smaller than those for "smaller" responses on both a t test ($p < .03$) and a Wilcoxon test ($p < .04$).

In order to study whether the magnitude of the distance effect varied with standard and response type in a fashion similar to the mean RTs, slopes of regression with log D were computed for each subject in each of the 10 conditions, and they were submitted to the same analysis of variance as above. No significant results were observed. There was a trend toward an interaction between standard and the response given ($F = 2.09$, $p < .10$). There was also a tendency for "larger" responses to yield lesser slopes (median 57.0) than "smaller" responses (median 77.8). This tendency was marginally significant when separate regression analyses with log D were performed on the raw data for "larger" and "smaller" responses (the slopes were 81.1 for "smaller" and 68.0 for "larger"; $p < .10$, two-tailed).

Finally, to provide results directly comparable with those from selection tasks, the target-standard couples that were tested twice under different response conditions were examined. For a target-standard distance of 10, these are the couples 35-45, 45-55, 55-65, and 65-75. The differences between "smaller" and "larger" response times for each of these couples are, respectively, 0, 41, 57, and 105 msec. For a target-standard distance of 20 (Couples 35-55, 45-65, and 55-75), the corresponding differences are −17, 5, and 10 msec. In both cases, the difference increases with the magnitude of the operands. This is a congruity effect in a funnel form for split 10 couples, and a crossover form for split 20 couples.

Discussion

Experiment 1 enables us to compare the results obtained in selection and classification tasks. First, the distance effect is reproduced in detail. The correlation of RTs with

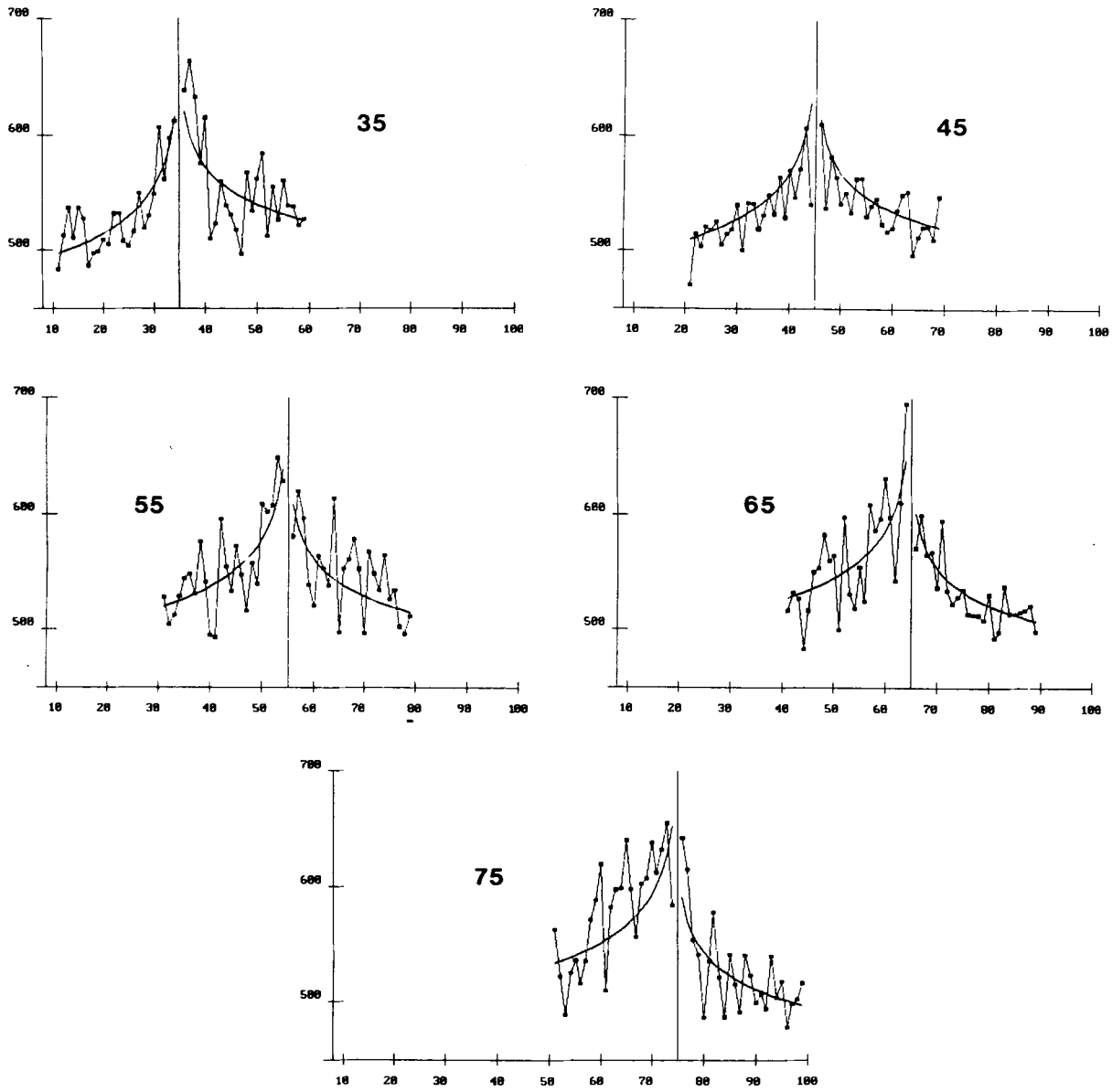

Figure 1. RTs to compare two-digit target numbers in the variable-standard classification tasks of Experiment 1. RTs are plotted as a function of target number, separately for each standard. The model fitted to the data is given by Equations 4a and 4b.

log D is good in all 10 cases. The significant influence on RTs of units within decades confirms the earlier finding (Dehaene et al., 1989; Hinrichs et al., 1981) that two-digit numbers are not compared lexicographically (first by decades, then by units when decades do not suffice to conclude), but rather holistically (the whole quantities that the two numbers represent are compared).

Second, varying the standard enabled us to discover, in a two-digit number classification task, two effects that had previously only been found in selection tasks with single digits. The magnitude effect appears as a significant influence of standard on the time to respond "smaller" or "larger." For "larger" responses, RTs decrease along with the standard; for "smaller" responses, they increase along with the standard. This interaction of response type and magnitude reveals the congruity effect: large numbers receive a "larger" response faster than a "smaller" response; the opposite is true for small numbers. The congruity is also found in a form directly comparable to that for selection tasks: some couples received both a "larger" and a "smaller" response, depending on which number in the couple played the role of the standard. With these couples, and for equal target-standard distances, we have found that the larger the magnitude of the numbers, the faster the "larger" response relative to the "smaller" response—again, a congruity effect.

Before examining models for these findings, we have to answer one question: Why were magnitude and con-

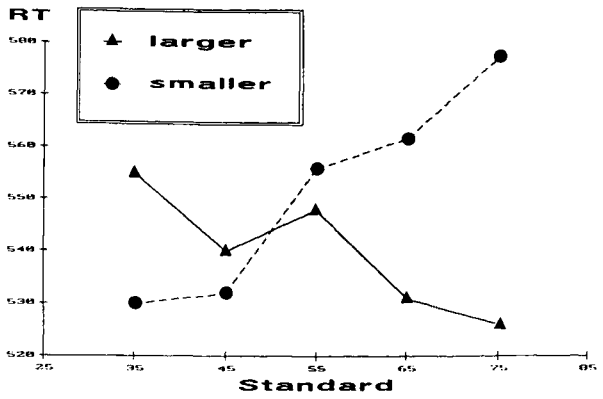

Figure 2. Mean comparison times as a function of standard for larger and smaller responses in Experiment 1.

gruity effects found in a variable-standard classification task, but not in classification with a fixed standard? The simplest reply is that the effects were already present with a fixed standard, but that the particular choice of standard rendered them obscure. Commonly, in classification with a fixed standard, targets are distributed symmetrically around the standard. For instance, Dehaene (1989) and Hinrichs et al. (1981) both chose a standard of 55, presumably because 55 stands precisely at the middle of two-digit numbers. Consider the comparison times with Standard 55 in Figure 1. They look fairly symmetrical when seen in isolation, as compared to the curves with Standard 75, for instance. The symmetry seems to stem from the particular position of 55 at the center of the range of targets tested. Around the center, the congruity effect cancels out, since no bias for "larger" or "smaller" responses is perceptible. This argument leaves open the possibility that with an asymmetrical range of targets, a congruity effect should emerge even in classification with a fixed standard. This possibility was tested in Experiment 2.

EXPERIMENT 2

In Experiment 2, the subjects were asked to compare a list of targets with a fixed standard of 75. "Larger" and "smaller" responses were equiprobable, but target numbers were drawn from the interval (20, 21, ... 99). Thus, the standard was much closer to the larger extreme of the range of targets (99) than to the smaller extreme (20). Accordingly, the subjects were expected to bias their judgments toward the congruent, "larger" response. Alternatively, the variable-standard task of Experiment 1 may have induced qualitatively different processes that are not normally used in classical comparison with a fixed standard. If this is true, then no congruity effect should be found in Experiment 2.

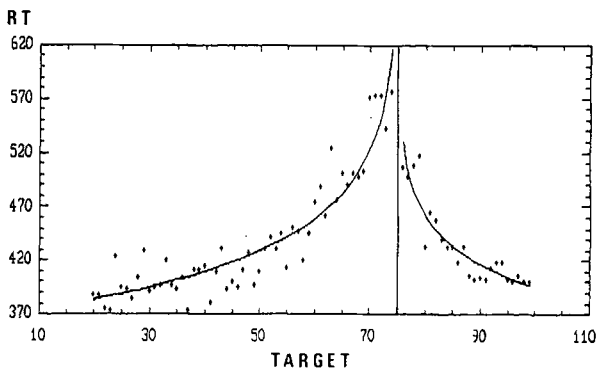

Figure 3. RTs to compare a two-digit target number to 75 in Experiment 2. Two independent regressions with log *D* are plotted.

Method

Subjects and instructions. Five members of the Laboratoire de Sciences Cognitives et Psycholinguistique were tested individually. Their ages ranged from 22 to 34. They were told that they would see a list of two-digit numbers, ranging from 20 to 99, to be compared to 75. They had to respond as fast as possible by pressing the right-hand response key if the target was larger than 75, and the left-hand key if it was smaller than 75.

Procedure. The same apparatus as in Experiment 1 was used. First an empty frame appeared at the center of the screen for 300 msec. Then the target appeared in the frame, and responses were recorded during the next 1,300 msec. Finally, 900 msec of blank screen preceded the next trial.

A random list, starting with 15 training trials, was made up for each subject. In the list, numbers from 20 to 74 appeared three times each, numbers from 76 to 99 seven times each. This ensured the equiprobability of "larger" and "smaller" responses. The order of the targets was random, with the constraint that the same target was never presented twice in a row. The list contained a total of 348 items. The experiment lasted less than 20 min.

Results and Discussion

Correct RTs were averaged for each target across subjects and trials to yield the RT curve of Figure 3. Two regressions of RTs with log D were performed separately for "smaller" and "larger" responses (Table 1). Both regressions were extremely good (the values of r^2 were 86.8% and 83.1%, respectively). The predicted RTs to Points 20 and 99 were not significantly different, but the predicted RT to Point 74 was significantly slower than to 76 ($p < .001$). The direction of the difference is compatible with a congruity effect.[1] The slopes of regression were also significantly different ($p < .002$): as in Experiment 1, the slope for "smaller" responses was steeper than the slope for "larger" responses.

The results of Experiment 2 with a fixed standard thus replicate those obtained with Standard 75 in the variable-standard condition of Experiment 1 (compare Figure 1, Standard 75, with Figure 3). It was demonstrated that congruity effects can be obtained in a classification task with a fixed standard. A unified description of data from selection and classification tasks now seems within reach. This will be the goal of the next section.

EQUATIONS FOR NUMERICAL COMPARISON

To understand the comparison of numbers, several facts should be considered. First, in Experiments 1 and 2 we found a continuous distance effect, with a significant in-

Table 1
Regression Analyses of Reaction Times in Experiment 2

Condition	Slope	r^2	*SEE*	RT(*D*max)	RT(*D*min)
Regressions with Log |Target − Standard|					
Target < 75	−58	86.8%	20	384	616
Target > 75	−42	83.1%	16	397	530
Regressions with Log |Log(Target) − Log(Standard)|					
Target < 75	−50	86.0%	21	375	602
Target > 75	−44	82.9%	16	398	531

Note—*SEE* = Standard Error of Estimate (in msec). RT(*D*max) is the predicted RT in response to 20 when the target < 75, to 99 when the target > 75. RT(*D*min) is the predicted RT in response to 74 when the target < 75, to 76 when the target > 75.

fluence of units within decades. We (Dehaene et al., 1989) and others (Hinrichs et al., 1981) have already argued that such data naturally point towards an analogical encoding. In comparison tasks, the magnitude of numbers appears to be represented on a continuum that conserves neighborhood relationships, a mental map called the *number line* (Restle, 1970). Comparison operations reveal an internal psychophysics (Moyer, 1973) that bears some similarity to the psychophysics of other, more perceptual comparison tasks.

Second, we have already noted that the position of the comparison standard with respect to the extremities of the range of targets has a considerable effect on RTs. When the standard comes closer to the "larger" extreme, for instance, "larger" responses become faster and faster. In our results, the congruity effect seems to reduce to an effect of the distance separating the standard from the extremes of the continuum.

Analogical encoding and the influence of distance from extreme points are the two characteristic features of the comparison model proposed by Jamieson and Petrusic (1975). This model, which we will now examine in detail, was proposed to account for data from selection paradigms. It will be our task to adapt it to fit data from the classification paradigms as well.

The Discriminability Model

In selection paradigms, subjects have to choose the larger (or smaller) of two digits. Jamieson and Petrusic (1975) suppose that instructions such as "choose larger" induce the choice of a reference point at the corresponding extreme of the continuum. In their model, the two operands are never compared directly. Rather, the relative distances of their representations from the reference point are compared. The model is named a *discriminability model* because it supposes that the same internal distance between two objects is more discriminable when the objects are close to a reference point than when they are far from it.

Consider, for example, the comparison of 5 to 6 in a "choose larger" selection task. According to the discriminability model, the two operands are first encoded on the number line at locations $s(5)$ and $s(6)$. The subject then sets a reference point at the "larger" extreme of the continuum. If only digits are compared, this "anchor" may be set around $s(9)$. The subject then responds by comparing $s(9)-s(5)$ to $s(9)-s(6)$, in other words by choosing the number closest to the reference point. Jamieson and Petrusic (1975) assume that reaction time is a function of the ratio of the two distances. Their model may be summarized by the following equation:

$$\mathrm{RT}_{\mathrm{selection}}(x,y) = F\left(\left|\frac{s_{\mathrm{ref}}-s(x)}{s_{\mathrm{ref}}-s(y)}\right|\right) \qquad (2)$$

where F is some function increasing on (0,1) and verifying $F(x) = F(1/x)$ for logical consistency, and s_{ref} is the location of the reference point, which varies according to whether the instructions specify "choose larger" or "choose smaller."

Extending the Model to Classification Tasks

The discriminability model applies correctly to selection tasks, but also to tasks that involve choosing which of two points is closer to a third (Holyoak, 1978; Holyoak & Mah, 1982). How can this model be extended to account for classification data? In classification, the two types of responses ("larger" and "smaller") are present. But Jamieson and Petrusic assume that each type of response stems from a comparison with a different reference point. Thus, the only way that their model may be extended to classification with a fixed standard is by assuming the simultaneous use of two reference points. Subjects should simultaneously compare the relative distances of the standard and the target both from the "larger" and the "smaller" extremes of the continuum. They should respond according to the reference point closer to the target than to the standard. If this view is correct, reaction times can be predicted by two equations:

$$RT\ (x > \text{standard}) = F\left(\left|\frac{s_{\max}-s(x)}{s_{\max}-s(\text{standard})}\right|\right) \qquad (3a)$$

and

$$RT\ (x < \text{standard}) = F\left(\left|\frac{s_{\min}-s(x)}{s_{\min}-s(\text{standard})}\right|\right) \qquad (3b)$$

where $s_{\max}$ and $s_{\min}$ are the two reference points on the continuum.

Fitness of the Model

It seems very unlikely that in a task as simple as numerical comparison with a fixed standard, two reference points are used. The discriminability model implies that four analogical distances are computed. This is a very uneconomical assumption indeed. Nevertheless, the match between the model and the data is excellent. Equations 3a and 3b successfully predict that if the standard varies in a classification task, then effects similar to those found with selection should be observed. This was verified in Experiment 1. Equation 3a further predicts that for equal target-standard distances, the time taken to respond "larger" should decrease with the standard, since the same distance becomes more and more discriminable with respect to the "larger" reference point. Similarly, the time taken to respond "smaller" should increase with the standard (Equation 3b). These predictions are nicely upheld in Experiment 1. The model presupposes that there are no fundamental differences in comparison tasks with a fixed or variable standard, as Experiment 2 demonstrated. Finally, the equations imply that when the target coincides with a reference point, RT should be minimal and independent of the standard. Indeed, RTs to 21 (respectively 20) and to 99 do not differ either in Experiment 1 or in Experiment 2.[2]

The data in Experiment 1 further constrain Equations 3a and 3b. In Experiment 1, RTs decreased logarithmically along with the distance between target and standard, independently of the standard: the distance effect and the effects of standard and response were essentially additive. It can be shown that these constraints suffice to uniquely determine the function F in the discriminability model. The following equations for classification times then replace Equations 3a and 3b:

$$RT(x > \text{standard}) = a \log \left| \frac{s(x) - s(\text{standard})}{s_{\max} - s(\text{standard})} \right| + b \quad (4a)$$

and

$$RT(x < \text{standard}) = a \log \left| \frac{s(x) - s(\text{standard})}{s_{\min} - s(\text{standard})} \right| + b \quad (4b)$$

Similarly, for selection times, the equation corresponding to Equation 2 is

$$RT_{\text{selection}}(x,y) = a \log \left| \frac{s(x) - s(y)}{\text{Max}\,(|s_{\text{ref}} - s(x)|, |s_{\text{ref}} - s(y)|)} \right| + b \quad (5)$$

where s_{ref} is the location of the reference point, which will coincide with $s_{\max}$ if the task is "choose larger," and with $s_{\min}$ if it is "choose smaller." Note that this equation reduces to the Welford (1960) function when the reference point is zero. Thus, it is appropriate to call it a generalized Welford function.

Equations 4a and 4b were fitted to the data of Experiment 1, using a least-square algorithm. Measures of goodness of fit did not vary greatly with the values of $s_{\min}$ and $s_{\max}$, as long as they stayed sufficiently close to $s(11)$ and $s(99)$; so the latter values were used throughout. A first regression was performed, assuming a linear encoding of numbers on the continuum (s = identity function). A two-parameter regression imposing equal slopes and intercepts in Equations 4a and 4b accounted for 53% of the variance of the 240 data points, as compared to 42% for a simple log D model. A slightly better fit was obtained by allowing for different slopes and intercepts in Equations 4a and 4b. The intercepts did not differ, but the slopes were significantly different for "larger" and "smaller" responses, thus agreeing with the experimental observation that mean RTs vary less with the standard for "larger" than for "smaller" response (see Figure 2). The model with three adjustable parameters (two slopes and one intercept) accounted for 56% of the variance, a considerable figure given that 240 data points were fitted and that each point is an average of only 20 measures (1 per subject). The corresponding best-fitting curves are plotted on Figure 1.

Reintroducing Fechner's Law

The asymmetry in the slopes found in Experiments 1 and 2 suggests that at an equal numerical distance from the extremes, a number is internally less distant from the "larger" extreme than from the "smaller" one. This is reminiscent of Fechner's law, which states that distances are internally compressed. Indeed, hypothesizing a logarithmic internal encoding (s = log function) slightly improves the proportion of the variance accounted for in Experiment 1 (54.2% with only two free parameters). It also avoids the unjustified hypothesis of different slopes for Equations 4a and 4b. The improvement is weak, probably because the correct encoding is halfway between linear and logarithmic. Nonlinear regression with a power function for s gives a slightly better fit ($r^2 = 56.3$), but with three free parameters; the best-fitting exponent for the power function is 0.56.

Similarly, regressions with $\log |\log(\textit{target}) - \log(\textit{standard})|$, an equation of the distance effect that takes into account a Fechnerian compression of internal distances, were performed on the data from Experiment 2 (see Table 1). This corresponds to taking a logarithmic function instead of a linear function for function s in Equations 5a and 5b; the first log models the nonlinearity of the distance effect, whereas the second log models Fechner's law. The inclusion of such Fechnerian encoding does not improve the proportion of variance accounted for, nor does it change the pattern of asymmetries in RTs: predicted RTs to 20 and 99 still do not differ, whereas predicted RTs to 74 and 76 differ significantly at $p < .01$. However, the slopes of regression no longer differ

($p > .20$). Thus, when Fechner's law is taken into account, two parameters—one slope and one intercept—suffice to account for the data using the discriminability model.

The inclusion of Fechnerian encoding also sheds light on two other previous findings. First, it agrees with the old finding that in selection paradigms, the "choose smaller" task takes longer than the "choose larger" task: because of the internal compression of distances, the "smaller" reference point is on the average more distant from the operands than the "larger" one is. Second, Fechnerian encoding predicts that in classification with a fixed standard, even when the standard is at the numerical center of the range of presented targets, slight asymmetries should be present in the RT curve. RTs should be equal at the extremes; close to the standard, however, "larger" responses should be slightly faster than "smaller" responses, because although numerically centered, the standard is internally closer to the upper extreme of the continuum. This pattern of asymmetries was observed in two-digit numbers comparison by Dehaene et al. (1989) and by Hinrichs et al. (1981), but it has remained unexplained.

OTHER PREVIOUS MODELS OF COMPARISON

Jamieson and Petrusic's (1975) discriminability model predicts in great detail the RTs for both classification and selection tasks. However, several other models of comparison have been proposed, and they must also be confronted with our data. As I will try to make clear below, most of them can only predict a limited portion of the data set.

Any model with analogical encoding, for instance, can easily account for the distance effect: two numbers that are close on the number line will be compared more slowly, either because their representations are more easily confused or because information retrieval has to be more accurate. Buckley and Gillman (1974) proposed a model based on the latter idea. In their model, evidence must be accumulated about the sign of the difference between the digits. When this difference is small, given internal noise, much more evidence has to be accumulated before a decision criterion is reached.

It is hard to see how such a model could predict magnitude and congruity effects without suffering ad hoc distortions. The magnitude effect could be explained by positing a representation of numbers that would obey Fechner's law: for equal numerical distance, the internal distances would decrease along with the magnitude of the numbers. This would slow down comparison—hence the magnitude effect. Yet attributing the magnitude effect to a fixed bias in the representation itself is problematic. It cannot explain the lability of the magnitude effect, which is modulated by the congruity effect. It would also predict that in classification, at equal numerical distances, "larger" responses should be faster than "smaller" ones, an asymmetry that has never been observed.[3]

Another model with important predictive power was proposed by Banks et al. (1976): the *semantic coding model*. This model assumes that after analogical encoding, numbers are labeled L+ or S+ according to whether they are larger or smaller than a boundary point. If the two numbers receive the different labels L+/S+, a response can be given. If the two labels are identical (L+/L+ or S+/S+), further processing is needed before the relative magnitudes of the numbers can be determined, the final labeling being L/L+ or S/S+. Because the labeling process is probabilistic, more distant numbers have a greater chance of being labeled L+/S+, and thus receive fast responses; this accounts for the distance effect. The magnitude and congruity effects arise from the congruency of internal and response codes. If the task is to select the larger number ("choose larger"), pairs of numbers labeled L/L+, where the internal code is congruent with the instructions, receive faster responses than pairs labeled S/S+, where a conversion of labels is necessary. Since the larger the numbers, the more likely the L/L+ labeling, the model predicts a magnitude effect in the "choose larger" condition. The direction of this effect will reverse in the "choose smaller" condition, where the congruent code is S/S+; hence the congruity effect.

Banks's model is definitely not refuted by the present data. On the contrary, it is fairly able to account for the gross findings. Nevertheless, its details are not fully specified (for example, what is the process by which an ambiguous code L+/L+ can be disambiguated into L/L+?). As a consequence, there are some experimental details with respect to which Banks's model is indifferent. Several features of Experiments 1 and 2 are predicted by the discriminability model, but they would be a matter of chance for Banks's semantic coding model. This is the case for the equality of RTs in response to extreme points 21 and 99, which has been observed in other papers as well (Dehaene et al., 1989; Hinrichs et al., 1981). It is also true of the subtle asymmetries observed close to the standard in fixed-standard classification tasks with a symmetrical range of targets (Dehaene et al., 1989; Hinrichs et al., 1981).

In addition, there has been one empirical objection to Banks's model, although not in the field of numerical comparison. Jamieson and Petrusic (1975) questioned Banks's hypothesis that the initial labeling (L+ or S+) is probabilistic. In the case of comparison of animal sizes, they have shown that individual animals are consistently classified as small or large. Nevertheless, the magnitude of the congruity effect is graded, not discrete as the model of Banks et al. (1976) would predict.

The finding of a graded congruity effect naturally points to an analogical explanation. This is exactly what Jamieson and Petrusic's (1975) discriminability model proposes. However, earlier work by Marks (1972) has also emphasized the same idea. Both models rely on the notion of a reference point or "anchor." However, according to Marks' (1972) *discriminal dispersion model*, the variance in the representation increases with distance from the reference point. Assuming that for equal numerical dis-

tance, a higher variance slows down the reaction time, the model predicts that when the numbers are close to the reference point (i.e., congruent with the instructions), they will be compared faster than when they are on the side of the continuum opposite the reference point—hence, a (graded) congruity effect.

The data from Experiments 1 and 2 fit Marks' (1972) model quite well. However, the model requires that the representation of objects vary with the instructions ("choose larger" or "choose smaller"). This feature was indirectly rejected by Banks and Root (1979) for the case of the loudness of sounds. Banks and Root used a production task, making the hypothesis that the representation of loudness tackled would be the same as the one used in comparison tasks. They showed that variability in the loudnesses of produced sounds was not affected by the form of the instructions, which specified the sound to be produced either in terms of loudness or in terms of softness. In the case of numbers, because the visual input is symbolic, it is equally unlikely that the variance of internal representations will depend either on the instructions or on which particular number is represented. Rather, this internal variance is likely to be constant, although not necessarily equal to zero.

CONCLUSION

Experiments 1 and 2 have shown that the magnitude and congruity effects found in selection tasks with single digits also appear in a predictable way in classification tasks with two-digit numbers. In Experiment 1, the subjects had to compare a two-digit target to a standard number that varied from trial to trial. "Larger" response times gradually decreased along with the standard, while "smaller" response times increased along with it. Thus congruent responses—for example, a "smaller" response to a small target— were indeed faster than incongruent ones. Experiment 2 showed that the congruity effect could be obtained in a fixed-standard classification task. Subjects had to compare targets ranging from 20 to 99 to a standard number 75, which was thus "large" in the range tested. Close to the standard, congruent responses ("larger") were faster than incongruent ones ("smaller").

The data of both experiments, as well as data from other experiments, are compatible with the discriminability model first proposed by Jamieson and Petrusic (1975) and later confirmed by Holyoak (1978) and Holyoak and Mah (1982). The model must be slightly extended to account for classification data. One has to suppose that two reference points are simultaneously used: the "larger" extreme of the continuum is used to respond "larger," and the "smaller" extreme to respond "smaller." This hypothesis yields Equations 4a and 4b, which predict classification times in full detail. The inclusion of Fechnerian encoding in these equations also accounts for subtle asymmetries between "smaller" and "larger" RTs. Other models, like Banks's semantic coding model (Banks et al., 1976), are not refuted by the data, but they do not seem to provide a complete account of the details.

Several questions remain open concerning numerical comparison. I have merely proposed equations that correctly describe RTs, not an algorithm that could produce such RTs. Similarly, Jamieson and Petrusic's (1975) model does not specify the mechanism by which the subjects perform comparison with respect to the reference points. Merely to find satisfactory equations is not sufficient to infer this mechanism. In fact, it is even possible that our equations confuse effects belonging to several distinct stages of processing. For example, Duncan and McFarland (1980) provide evidence suggesting that the distance effect is central to comparison per se, while the congruity effect stems from an initial encoding stage. Yet in the equations above, the two effects are given a homogeneous treatment, and it is tempting, but logically unfounded, to attribute them to common sources.

In trying to infer the comparison algorithm, a logical problem appears. One must explain why "larger" responses are always performed in reference to the "larger" extreme of the continuum, and "smaller" responses to the "smaller" extreme. This strategy appears inefficient in some cases. Other simpler strategies should be possible. In fact, the use of one reference point is always sufficient to determine the relative magnitudes of the standard and the target. Such a single-reference strategy is indeed used in "choose larger" or "choose smaller" tasks. Why can't it be utilized in classification?

Another possible strategy with two reference points would be to respond using the anchor from which the target/standard distance is optimally discriminable. This would correspond to a race between the two anchors to provide the correct response. But Experiment 2 clearly demonstrates that "smaller" responses are always reached in reference to the "smaller" anchor, even though close to the standard, a much faster response could be reached by referring to the "larger" anchor. Why is the time to compare 74 to 75 so different from the time to compare 76 to 75? This finding is incompatible with a race between anchors.

In brief, numerical comparison is a highly constrained, inflexible process. In particular, "smaller" responses can only be reached using the "smaller" anchor, and "larger" responses using the "larger" one. This feature has several drawbacks, in particular the necessity to use two simultaneous reference points even in a classification task with a fixed standard. The attempt to determine which algorithm is responsible for such functional properties will be examined in a coming publication.

REFERENCES

AUDLEY, R. J., & WALLIS, C. P. (1964). Response instructions and the speed of relative judgments: I. Some experiments on brightness discrimination. *British Journal of Psychology*, **55**, 59-73.

BANKS, W. P., FUJII, M., & KAYRA-STUART, F. (1976). Semantic congruity effects in comparative judgments of magnitudes of digits. *Jour-*

nal of Experimental Psychology: Human Perception & Performance, **2**, 435-447.

BANKS, W. P., & ROOT, M. (1979). Semantic congruity effects in judgments of loudness. *Perception & Psychophysics*, **26**, 133-142.

BUCKLEY, P. B., & GILLMAN, C. B. (1974). Comparison of digits and dot patterns. *Journal of Experimental Psychology*, **103**, 1131-1136.

DEHAENE, S., DUPOUX, E., & MEHLER, J. (1989). *Is numerical comparison digital? Analogical and symbolic effects in two-digit number comparison*. Manuscript submitted for publication.

DUNCAN, E. M., & MCFARLAND, C. E., JR. (1980). Isolating the effects of symbolic distance and semantic congruity in comparative judgments: An additive-factors analysis. *Memory & Cognition*, **8**, 612-622.

HINRICHS, J. V., YURKO, D. S., & HU, J. M. (1981). Two-digit number comparison: Use of place information. *Journal of Experimental Psychology: Human Perception & Performance*, **7**, 890-901.

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

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

HOLYOAK, K. J., & MAH, W. A. (1982). Cognitive reference points in judgments of symbolic magnitude. *Cognitive Psychology*, **14**, 328-352.

JAMIESON, D. G., & PETRUSIC, W. M. (1975). Relational judgments with remembered stimuli. *Perception & Psychophysics*, **18**, 373-378.

JOHNSON, D. M. (1939). Confidence and speed in the two-category judgment. *Archives of Psychology*, **241**, 1-52.

KOSSLYN, S. M., MURPHY, G. L. BEMESDERFER, M. E., & FEINSTEIN, K. J. (1977). Category and continuum in mental comparisons. *Journal of Experimental Psychology: General*, **106**, 341-375.

MARKS, D. F. (1972). Relative judgment: A phenomenon and a theory. *Perception & Psychophysics*, **11**, 156-160.

MOYER, R. S. (1973). Comparing objects in memory: Evidence suggesting an internal psychophysics. *Perception & Psychophysics*, **13**, 180-184.

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

PARKMAN, J. M. (1971). Temporal aspects of digit and letter inequality judgments. *Journal of Experimental Psychology*, **91**, 191-205.

RESTLE, F. (1970). Speed of adding and comparing numbers. *Journal of Experimental Psychology*, **83**, 274-278.

SEKULER, R., & MIERKIEWICZ, D. (1977). Children's judgments of numerical inequality. *Child Development*, **48**, 630-633.

SEKULER, R., RUBIN, E., & ARMSTRONG, R. (1971). Processing numerical information: A choice time analysis. *Journal of Experimental Psychology*, **90**, 75-80.

WELFORD, A. T. (1960). The measurement of sensory-motor performance: Survey and reappraisal of twelve years' progress. *Ergonomics*, **3**, 189-230.

NOTES

1. It may be argued that the "larger" RTs were faster in Experiment 2 simply because of the differing amount of practice received: numbers larger than 75 were presented 7 times each, whereas numbers smaller than 75 were presented only 3 times each. This hypothesis is invalidated by the fact that the data from the first 50 trials of each subject, in which the presence of practice effects is unlikely, show the same asymmetry in RTs. Also note that in Experiment 1, the ranges of numbers that get the least repetition (11-30 and 80-99) in fact yield the fastest responses, not the slowest.

2. I have implicitly assumed that the subjects chose reference points adapted to the range of targets tested. Another possibility is that reference points assume fixed locations on the number scale. The two hypotheses are not distinguishable on the basis of the present experiments, but they can be separated, for example, in classification with Standard 65 and numbers ranging from 31 to 99: an asymmetrical RT curve like the one observed in Experiment 2 would show that 11 or so is still used as a reference point, whereas a symmetrical curve would demonstrate that the reference point has shifted to about 31. Dehaene et al. (1989) did precisely this experiment and found an almost symmetrical curve, with equal RTs in response to 31 and 99. Even without this experiment, it is obvious that reference points are shifted to adapt to one-digit or two-digit number comparison.

3. Although I have had to reintroduce Fechner's law in the discriminability model, the reader should persuade himself that it does not encounter the same objection as above. In the discriminability model, the magnitude effect is described as an effect of distance from the reference points. Fechner's law plays only a minor role in accounting for subtle asymmetries between "smaller" and "larger" responses.

(Manuscript received August 3, 1988;
revision accepted for publication November 23, 1988.)