

Prototypicality and Deductive Reasoning

CHRISTOPHER CHERNIAK

Department of Philosophy, University of Pennsylvania

It is hypothesized that people use formally incorrect deductive procedures, and this is sometimes advisable. The particular "prototypicality heuristic" investigated is: to determine validity of a proof, (a) work out an example, and (b) pick a "good" rather than arbitrary example. An interaction was predicted between validity of inference and prototypicality of example. Experiment 1, although quite sensitive to "calibration" variables, does not reveal the interaction in reaction times. However, Experiment 2, in which subjects' time was limited, seems to elicit the interaction in subjects' first-trial confidence judgments; the explanation proposed is that subjects then decide they must shift to a "quick and dirty" heuristic. Experiment 3 provides some preliminary evidence that subjects' "metaheuristic" decision to use the heuristic does lower error rates. © 1984 Academic Press, Inc.

The hypothesis of this paper is that people use a wide variety of formally incorrect deductive procedures, and that this strategy is sometimes rationally advisable. If people in fact use quick and dirty deductive reasoning heuristics, one of the most important candidates for such a heuristic would extend a familiar observation (Hadamard, 1945) and recommendation (Polya, 1954) in the field of mathematical problem solving: Do not just "turn the crank" algorithmically. Instead, when trying to discover a proof of a theorem, it is often indispensable to work out first a specific example that satisfies the initial conditions.

The rationale would be that empirical studies indicate that more abstractly formulated deductive tasks are more difficult; for example, the same form of syllogism with variables ("A") as opposed to concrete terms ("freshman") is harder (Wilkins, 1928; cited in Wason & Johnson-Laird, 1972). In addition, Rosch, Mervis,

Gray, Johnson, and Boyes-Braem (1976) have obtained evidence that basic-level categories (e.g., "eagle") are more processable than either more general superordinate categories (e.g., "bird") or more specific subordinate categories (e.g., "bald eagle") (this also suggests limits for this level-descending strategy). These results fit with Paivio's (1971) findings that concrete and more imageable words are remembered better than abstract words, and pictures better than words. Abstraction, therefore, can be costly; thinking out a problem in terms of an example would be a way of descending to a lower, and thus more manageable, level of abstraction—a kind of local and temporary Piagetian regression from abstract to concrete.

Furthermore, we can add a second heuristic truism, so obvious as to be commonly overlooked: The example chosen should be a "good" one—that is, representative or typical of the problem, rather than an arbitrary instance. Such a procedure will minimize the cost of the overgeneralization involved in the level-descending strategy. This is important given the hypothesis of this paper that at least in everyday life we seem to extend use of these procedures, not necessarily inadvisably, from the heuristic "context of discovery" of ideas for proofs

William Gutowski provided indispensable assistance with data analysis. I am also indebted to H. Clark, K. Hemenway, W. Onifer, L. Rips, E. Rosch, J. Rosenberg, and D. Swinney for their suggestions and generous assistance. Requests for reprints should be sent to Christopher Cherniak, Department of Philosophy, University of Pennsylvania, Philadelphia, Penn. 19104.

into the context of our actual rough intuitive proofs. If we then suppose that the more representative instances of a category are the more prototypical ones, as Rosch and Mervis (1975) argue, then the hypothesis here is that thinking in terms of prototypes simplifies reasoning.

What are the implications of a prototype theory of mental representation for human deductive reasoning? The key element of a prototype theory, in particular as articulated by Rosch (1977), is rejection of the classical assertion that in all situations every category is defined in terms of a set of features that are necessary and sufficient for membership in the category. Instead, members of the category sometimes may not be equally and fully members of the category (e.g., subjects reliably report a robin is a better example of the category "bird" than a turkey); membership in the category is evaluated in terms of centrality—resemblance of an instance to a prototype or best example of the category.

The role of prototypicality in a sentence verification task has been studied extensively: On the one hand, subjects determine more quickly that a prototypical instance belongs to a category than does a relatively atypical one. (Rosch, 1973; Rips, Shoben, & Smith, 1973; Smith, Shoben, & Rips, 1974). On the other, reaction time is longer for correctly rejecting a false sentence (e.g., "A bat is a bird") if the instance is very similar to the category prototype (Rips et al., 1973). At the opposite extreme from a simple task like sentence verification is a strategy Kahneman and Tversky (1972, 1973) have extensively investigated, one very like prototype matching for everyday intuitive reasoning about the future, the "representativeness heuristic." In estimating the likelihood of an outcome, subjects tend to ignore prior odds and instead ("fallaciously") base their judgment on how well the outcome represents the essential features or "stereotype" of the prior evidence. In addition, Rips (1975) has found that subjects inductively generalize

more easily from prototypical or representative instances of a category to other instances, than from less prototypical ones.

It would be natural to wonder about the role of a phenomenon as important and widespread as prototypicality in tasks that fall between simple sentence verification and informal inductive reasoning—namely, deductive reasoning. Such experiments may help to flesh out models to explain the remarkably ubiquitous effects for the more complex cases of nondeductive reasoning (indeed, it is curious that Kahneman and Tversky-type effects were not first sought for deduction). However, current accounts of deductive reasoning ignore the role of prototypicality. As a recent example, according to the "analogical theory of reasoning" of Johnson-Laird and Steedman (1978), the first stage in the process of making a syllogistic inference is the semantic representation of the quantified assertions involved: "A class is represented simply by thinking of an arbitrary number of its exemplars." But are all exemplars equal? It seems plausible that prototype structuring of the class would (and, as explained above, should) affect the choice of exemplars.

Do human beings employ a prototypicality heuristic in their deductive reasoning, that is, in such tasks do they exploit the prototype structuring of categories? The prototype theory of category structure is stronger than a classical theory, in that it suggests that making some valid inferences will be more difficult than making others, and that refraining from making some invalid inferences will be more difficult than not making others. The two simplest implications of a prototype theory involve how information about category-instance relations will be used in elementary deductive inferences. Prototypicality would affect performance of the tasks through an implicit premise. The first prediction is of a differential effect of instance prototypicality on making valid inferences. For example, the inference

All A's are robins.
All A's are birds.

when it is interpreted as including the definitionally true tacit premise "All robins are birds" constitutes the standard three-term syllogism

All A's are robins.
[All robins are birds.]
All A's are birds.

The original "tacit syllogism" is valid in the sense that it is not possible for both the one explicit premise to be true and the conclusion to be false. Given the above evidence suggesting that true statements of the form "All [instances] are [categories]" are more easily verified when the instance is central as opposed to peripheral, one would expect an inference which will be valid with such a statement as a tacit premise correspondingly to be easier to perform for central than for peripheral instances. For example, the first inference above and the following similarly definitionally valid inference

All A's are turkeys.
All A's are birds.

should be affected this way, since they respectively require implicit use of the premises "All robins are birds" and "All turkeys are birds."

The second prediction of the prototype theory is of a differential influence of prototypicality on making invalid inferences, but in the opposite direction from the effect on valid inferences. The false-sentence reaction time evidence cited above suggests the tendency to infer, or "illicitly convert," a definitionally false statement of the form "All [category] are [instances]" from a true statement of the form "All [instances] are [category]" will be greater for central instances of the category. Besides conventional feature-analytic accounts of this differential confusability, a pretheoretic explanation would just be that subjects reason: "On the one hand, to a rough first approximation, a bird just *is* a robin; on the other,

a bird isn't a turkey, since a turkey is hardly a bird at all." Consequently, one would expect a greater tendency of subjects in effect to accept, and so to use as a tacit premise in their reasoning, a statement of the form "All [category] are [instances]" when the instance involved is more central. The second prediction regarding deductive reasoning, then, is that subjects should be more likely to make certain invalid inferences, namely, those which would be valid with an additional premise of the form "All [category] are [better instances]."

One of the deductive tasks in which such an implicit premise would be most easily used is again a "tacit syllogism" which, with this premise, constitutes a three-term syllogism. For example, the inference

x Some A's are birds.
Some A's are robins.

is not valid. But with the false premise "All birds are robins," the inference becomes a valid syllogism:

Some A's are birds.
[All birds are robins.]
Some A's are robins.

The "differential confusability" prediction here is that such invalid tacit syllogisms will be more difficult to judge correctly to be invalid when a central instance is involved than when a peripheral one is. In this way the above invalid tacit syllogism should be harder than

x Some A's are birds.
Some A's are turkeys.

Therefore, the basic overall prediction of the prototype theory is of an interaction between whether or not an inference is valid and whether or not it involves in the above way a central instance. The interaction should be of the strongest kind, in that, depending on whether an inference is valid, the effect of centrality should entirely reverse, rather than just change strength.

Of course, people—readers of this

paper, for example—are able to avoid “thinking in stereotypes” on the above simple deductive tasks, given enough time. Any account of the role of prototypicality in reasoning must allow for the fact that people can reason without prototypicality effects in some situations. That is, such a theory must postulate at least two distinct kinds of reasoning processes, each with its own benefits and costs: (1) “Correct but slow” processes which employ classical category definitions in terms of necessary and sufficient truth conditions. (2) “Prototypicality heuristics,” “quick but dirty” processes which employ the prototype representation of the category. It should be emphasized that prototypicality heuristics need not be the only means of making a speed–reliability trade-off; various procedures may be scattered throughout the space of possible speed–accuracy trade-offs.

Under what conditions will each process occur? The most interesting conjecture to begin with is: Quick and dirty heuristics in general, and prototypicality heuristics in particular, would more likely be used—and ought to be used—when there is pressure to sacrifice formal correctness for speed. The less time perceived to be available for a given task, the more likely prototypicality effects should be. (In addition, subjects may sometimes employ the “simultaneous strategy” of undertaking both types of procedure in parallel, waiting for the slow but correct results unless they decide there is not time, in which case any available earlier quick but dirty results would first be used.) Such decisions may be involved in rapid, automatized processes as well as deliberate, conscious problem-solving strategy.

It is therefore clear that “context matters:” performance of the inferences here can be expected to be sensitive to experimental conditions. Besides the speed–accuracy constraints of a given experimental task, there is the question of when the subject is in fact interpreting his job as a “pure logic” task. When subjects naive to formal

logic are asked to determine validity of an inference, they tend instead to evaluate truth or probability of the premises and/or conclusion (Henle, 1962). Such probability estimates can be expected to vary significantly (and in either direction) for a given sentence context, depending on whether a central or peripheral instance occurs (e.g., compare “John ate a turkey” and “John ate a robin”). Therefore, prototypicality could affect performance of inferences through an entirely different process, content–validity confusions, from the one we are investigating. The simplest approach for studying deductive reasoning here is to prevent subjects from using their real-world knowledge, besides knowledge of the structure of the relevant category, to make unwarranted interpretations. Consequently, in the syllogisms for the present experiments all content except the category and its instance has been eliminated by using the variable “A” in the explicit premise and conclusion.

EXPERIMENT 1: REACTION TIME

This experiment tested for the predicted validity–centrality interaction in subjects’ reaction-time performance on tacit syllogisms. According to the “quick but dirty” rationale for the prototype heuristic, subjects should shift most to the heuristic under time pressure. Nonetheless, the natural first assumption to try is simply that the predicted difficulty of the task, depending on whether the syllogism was valid or not, and on whether a central or peripheral instance is involved, would be reflected in the time required for subjects to accomplish it. Therefore, subjects’ reaction times should be longer for correctly judging syllogisms to be valid with peripheral instances than with central ones, but the reaction times should be shorter for invalid syllogisms with peripheral instances than with central ones. Materials were chosen so that (1) the variables of whether or not an inference is valid, or affirmative, or universal can themselves be used as a calibra-

tion of the sensitivity of the experiment; and (2) the predicted validity–centrality interaction can be checked under two basic conditions, whether or not an inference involves affirmative or universal propositions. (In addition, the experiment was designed to try to test whether training in logic would affect the predicted interaction, that is, to determine whether the prototypicality heuristic was stable enough not to be reduced under this condition.)

Method

Materials. Arguments presented to the subjects were tacit syllogisms in the sense explained above, that is, three-term syllogisms with one explicit and one unstated premise and a conclusion. The eight premise–conclusion pairs that are the explicit portions of the syllogisms constitute four forms of the syllogism which are valid for the true tacit premise of the form “All [instances] are [category],” and four forms of the syllogism which are invalid for such a premise, but are valid for the false tacit premise “all [category] are [instances]” (see Table 1). Equal numbers of the valid and of the invalid inferences are composed of affirmative and of negative propositions, and the sets of affirmative and of negative inferences in turn are equally composed of universal and of particular propositions; completely crossing these three variables in this way satisfies the above desiderata.

The seven categories used, with one central and one peripheral instance of each (see Table 2), were chosen from the results in Rosch’s (1975) normative study. Each pair of instances of each category satisfies the following criteria: (1) One of the instances is rated a good example of the category, that is, ranked higher than 1.30 on Rosch’s 7-point scale; (2) the other is rated a poor example, that is, lower than 2.75; (3) the mean word frequency of the good instances does not exceed that of the poor instances according to the Kucera and Francis (1967) sample of written English; and (4) vice versa for the mean number of syllables.

Each of the eight forms of the syllogism was completed in two versions for each of the seven categories: The “[category]” position was filled by the category name; in one version the “[instance]” position was filled by the central instance of that category, and in the other by the peripheral instance. For example, the first tacit syllogism displayed in this paper is an instance of inference No. 1 with the “bird” category and the “robin” instance substituted. Each of the resulting 112 items was typed in the above format: the explicit premise above the conclusion, with a line separating the two. An additional six practice syllogisms were constructed from forms and categories not previously used.

The two versions of each syllogism for a given category were randomly assigned to

TABLE 1
FORMS OF THE SYLLOGISM USED IN EXPERIMENTS 1 AND 2

Valid (with true tacit premise “All [instance] are [category]”)	
(1) <u>All A’s are [instance].</u> All A’s are [category].	(2) <u>Some A’s are [instance].</u> Some A’s are [category].
(3) <u>No A’s are [category].</u> No A’s are [instance].	(4) <u>Some A’s are not [category].</u> Some A’s are not [instance].
Invalid (with tacit premise “All [instance] are [category]”)	
(5) <u>All [instance] are A’s.</u> All [category] are A’s.	(6) <u>Some A’s are [category].</u> Some A’s are [instance].
(7) <u>No A’s are [instance].</u> No A’s are [category].	(8) <u>Some [category] are not A’s.</u> Some [instance] are not A’s.

TABLE 2
CATEGORIES AND INSTANCES USED IN
EXPERIMENTS 1 AND 2

Category	Instance	
	Central	Peripheral
Piece of furniture	Sofa	Bench
Fruit	Banana	Fig
Vehicle	Bus	Tractor
Vegetable	Carrot	Potato
Bird	Robin	Turkey
Item of clothing	Shirt	Hat
Weapon	Machine Gun	Whip

two different sets, with the restrictions (1) that each set contain an equal number of central and peripheral versions of each category, and (2) that in each set the number of central and of peripheral versions for each syllogism differ by only one. Each resulting set of 56 items was separately randomly ordered, with the constraints that no two consecutive items be of the same syllogism form or category, nor should four consecutive items be all valid or invalid.

Procedure. The logic-trained and untrained subject groups were each randomly divided into two equal subgroups, each of which received one of the sets of 56 items; each subgroup in turn was randomly divided into two groups, one of which received the item set in reverse order from the other. Thus, each subject judged the validity of 56 items: each of the eight syllogism forms for each of the seven categories with either the central or peripheral instance.

The items were presented to one subject at a time in a Harvard-type tachistoscope. First, the concept of validity was briefly explained. The subject was instructed to press one of two keys to indicate whether or not he believed the displayed syllogism was valid (dominant hand was randomly switched), and to respond as quickly as possible while keeping errors to a minimum. In a trial, a fixation cross appeared in the center of the screen; after 1½ seconds, one of the stimuli appeared, and the subject responded. The field then darkened

for 10 seconds, until the next trial began. Before the test trials were six practice trials using items not in the test series, with feedback and the opportunity for questions after each response; subjects were reassured that peripheral practice items were in fact members of the corresponding categories.

Subjects. Subjects were 50 University of California, Berkeley, undergraduates, required to be native English speakers. Twenty-eight of these were also required to have taken no college logic course, and were participating in the experiment for psychology course credit. The remaining 22 subjects were paid volunteers recruited from the second-quarter course in the Berkeley Philosophy Department's introductory logic sequence.

Results and Discussion

Evidence of a validity–centrality interaction occurs in almost no portions of the reaction-time data, although the experiment is quite sensitive to the “calibration” variables.

Subjects with less than 67% correct scores (two logic trained, four untrained) were eliminated from the analysis. In each of the training groups, pairs of subjects, where one had received the 56 items which were the “mirror image” with respect to category of the series the other received, were yoked by matching error scores as closely as possible. There were then 10 such logic-trained subject pairs and 12 untrained subject pairs, each of which had received all 112 items. An analysis of variance was performed on correct reaction times, with logic training as a between-group variable, and whether or not an inference was valid, or affirmative, or universal, or involved a central category instance, as within-subject variables.

An *F* test of the variable of logic training of subjects did not approach significance, $F(1,21) = .24, p > .62$; nor was any interaction of this variable with the other variables significant. All analyses below there-

fore are of trained and untrained subjects combined. The principal *prima facie* prediction, of an interaction between validity of inference and centrality of instance, was not confirmed; the *F* test for this interaction did not approach significance, $F(1,21) = 1.55, p > .23$. The variable of centrality alone also did not have a significant effect, $F(1,21) = .01, p > .91$. However, the "calibration" variables of validity, affirmativeness, and universality of inference all showed differences with high levels of significance (see Table 3a). (Analysis of logarithms of the reaction times, and of binomially transformed error scores, showed the same pattern.) Tests of the mean central-peripheral difference for each inference in each group revealed that only inference #6, for logic-trained subjects, showed a significant difference, with central 5.58 seconds and peripheral 4.44 seconds, $F(1,9) = 18.00, p < .002$; this was in the predicted direction, of centrality hindering performance for an invalid inference.

The earlier discussion of the likelihood that the experimental task is highly sensitive to experimental setting suggests that the first occurrence of a central and of a peripheral version of each inference for each subject should be examined, since doing 56 consecutive tasks of identical format is so repetitive that it may blunt prototypicality effects. These first-occurrence

data follow the same pattern as the above results for subjects' overall performance on all 56 items: Whether or not subjects were logic trained by itself was not a significant variable, $F(1,42) = .01, p > .92$, hence the following analyses are of trained and untrained groups combined. The main prediction of a validity-centrality interaction is again not confirmed, $F(1,42) = 1.78, p > .19$; nor is the variable of centrality alone significant, $F(1,42) = 1.02, p > .32$. The "calibration" variables all show very significant differences in the same directions as for the overall data (see Table 3b). (An analysis of variance of the logarithms of these reaction times, and an analysis of the reciprocal of the reaction times show the same pattern; so also do sign tests of error scores.)

The major feature of the reaction time performance is a dominance of form over content, in the sense that formal variables involving the logical constants ("not," "all," etc.) show strong effects, while the content variable of instance centrality has no effect. These "formal effects" for inferences each correspond to effects reported in sentence-verification tasks. The grand mean time for overall performance of 5.4 seconds and the 13.3% error rate are rather high for reaction time experiments. Of course, the lack of a validity-centrality interaction may just be the result of some pe-

TABLE 3
EFFECTS OF INFERENCE FORM ON RESPONSE

(a) Experiment 1: Overall reaction time (s)				<i>F</i> (1,21)	<i>p</i> <
Valid	5.020	Invalid	5.698	26.58	.0001
Affirmative	4.689	Negative	6.030	77.17	.0001
Universal	4.912	Particular	5.807	21.47	.0001
(b) Experiment 1: First response reaction time (s)				<i>F</i> (1,42)	<i>p</i> <
Valid	5.448	Invalid	5.960	11.53	.005
Affirmative	5.070	Negative	6.338	36.31	.0001
Universal	5.320	Particular	6.088	8.25	.007
(c) Experiment 2: First response 10-point ratings				<i>F</i> (1,45)	<i>p</i> <
Affirmative	8.565	Negative	7.895	9.79	.004
Universal	8.950	Particular	7.510	33.07	.0001

culiarity of this experiment (one which still allows the significant "formal" effects), or because centrality effects are so slight that they are swamped in this type of experiment.

The earlier discussion of the likelihood that we use prototypicality heuristics when we perceive pressure for a tradeoff of reasoning reliability for speed also suggests another explanation: In the reaction time experiment subjects may be interpreting the instruction "Respond as quickly as possible while keeping errors to a minimum" so strictly that they choose to employ slow but correct deductive procedures rather than quick but dirty heuristics which involve thinking in terms of prototypes. The experiment seems to provide the latitude for such a cautious speed-reliability tradeoff. While there would then be no centrality effects, the formal effects would still remain. Indeed, according to the above theory of prototypicality heuristics, only the centrality effect should be sensitive to speed-accuracy trade-offs, since that theory does not imply that when subjects use prototypicality heuristics they do, or can, ignore the formal structure of arguments.

EXPERIMENT 2: FIXED INTERVAL

The two-process model of reasoning therefore suggests an experiment in which the possibility of a cautious speed-accuracy tradeoff is restricted: "Clamp" the time available to perform each inference, so that subjects are more likely to decide they have to shift to a quick and dirty heuristic involving prototypes. In addition, to explore the stability of the effect of a fixed time constraint, one subject group had a 4-second interval, somewhat smaller than the 4.2-second reaction time for the fastest inference (No. 2), and another had a 10-second interval, somewhat larger than the 7.3-second reaction time for the slowest inference (No. 8). Since error scores by themselves are rather coarse indicators of performance, sensitivity of the experiment was boosted by obtaining subjects' estimates of confidence in their answers.

Method

Materials. The two sets of 56 cards from Experiment 1 were used; the order within each set was unchanged, and the sets were presented as two consecutive uninterrupted balanced blocks.

Procedure. Every subject received all 112 items. Thirty subjects had 10 seconds/item, and 19 subjects had 4 seconds/item; each of these groups was randomly divided into two subgroups, one of which was shown the item set in reverse order from the other. For each subgroup, the stimulus cards were displayed on a motion-picture screen using an opaque projector, with a clock-actuated electric shutter controlling stimulus duration and also the between-stimulus interval of 9 seconds. Subjects responded on machine-readable answer sheets: For each syllogism, there were two 5-point confidence scales (where "1" meant "completely sure," "5" meant "barely sure," etc.), one in the "Valid" column and one in the "Invalid" column. The concept of validity was explained to the subjects exactly as in Experiment 1. Subjects were instructed that as soon as, but not before, the screen darkened after an item was projected, they were first to decide whether or not the displayed syllogism was valid, and then to decide and mark how confident they were of their answer on the corresponding 5-point scale. The test trials were preceded by six practice trials using the same items in the same order as in Experiment 1, with feedback and opportunity for questions (peripheral practice instances were again identified as members of their categories).

Subjects. Subjects were 49 Tufts University undergraduates participating for psychology course credit. They were required to be native English speakers and to have taken no college logic course.

Results and Discussion

The experiment provides some evidence suggesting that "clamping" subjects' time elicits the predicted validity-centrality in-

teraction; in particular, this seems to arise for first-occurrence, but not overall, data.

One subject each from the 10- and 4-second groups had to be eliminated from the analysis because they lost their places on the answer sheets. The error score and confidence rating of each response were combined into a single 10-point scale rating, where "1" meant the subject was wrong and completely sure, "10" meant correct and completely sure, and so on. An analysis of variance was performed on all of these 10-point scores, with stimulus duration as a between-group variable, and with validity, affirmativeness, and universality of inference, and centrality of instance as within-subject variables. The results paralleled those for overall reaction time performance; in particular, the interaction between inference validity and instance centrality did not approach significance, $F(1,45) = 1.43, p > .24$. (The same pattern held for an analysis of performance on the first block of 56 items each subject received. Furthermore, the interaction between block number, validity, and centrality did not reach significance.) The variable of centrality alone again was not significant, $F(1,45) = .41, p > .53$. The effect of validity was no longer significant; however, there were still significant effects, in the same directions as in all earlier analyses, for affirmativeness and universality.

On the other hand, performance on the first occurrences of a central and of a peripheral version of each inference for each subject seems to show a different pattern. Whether stimulus duration was 10 or 4 seconds had no significant main effect, $F(1,45) = 2.70, p > .11$, nor were any relevant interactions between this and other variables significant (in particular, the duration–validity–centrality interaction was not, $F(1,45) = .18, p > .67$). Therefore, all results below are for both groups of subjects combined. The predicted validity–centrality interaction (see Table 4) is now significant, $F(1,45) = 7.50, p < .008$.

The effect of instance centrality on per-

TABLE 4
EXPERIMENT 2: VALIDITY \times CENTRALITY, FIRST
RESPONSE 10-POINT RATINGS

	Central	Peripheral
Valid	8.485	7.847
Invalid	8.083	8.505

formance for each inference separately was analyzed for the first-occurrence results. Two of the eight inferences showed significant central–peripheral differences in the predicted directions (inference #1, $F(1,45) = 13.23, p < .01$; inference #6, $F(1,45) = 4.49, p < .05$). In addition, the effect of centrality on all valid inferences alone is significant, $F(1,45) = 6.01, p < .05$, but there is not a significant centrality effect for all the invalid inferences alone, $F(1,45) = 2.63, p > .10$. However, the mean central–peripheral differences for 7 of the 8 inferences are in the predicted directions; No. 7 is in the wrong direction, but its central–peripheral difference is the smallest of all eight inferences. Therefore, the inference validity–instance–centrality interaction is also significant by a Wilcoxon matched-pair sign test, $.005 < p < .01$. The first-occurrence error scores show a pattern similar to that of the confidence ratings. In particular, a Wilcoxon sign test of differences between mean central and peripheral error scores of each inference was as significant as for the confidence-scale data, confirming the validity–centrality interaction also for the error data by itself.

For a given inference type, the first occurrence with a central instance and with a peripheral instance in each random series could not be at the same trial number. Therefore, the apparent effect of instance centrality might actually have been an artifact of the presentation order of these first occurrences. However, performance for each inference was better on later trials than earlier trials in 48.3% of all cases, closely approaching chance; furthermore, a χ^2 test of this measure of differential performance for the central versus the peripheral instances also did not approach signif-

icance ($\chi^2_1 = .85, p > .25$). Thus, trial number here does not seem to be a confounding variable.

The first-occurrence data seem to confirm a validity–centrality interaction of the predicted strong kind, where the central–peripheral difference does not just change magnitude depending on validity/invalidity, but entirely changes direction. Furthermore, when the interaction does occur, it appears to be robust; comparison of it with the “calibrator” affirmativeness and universality effects (see Table 3c) indicates they are all of about the same magnitude, roughly one confidence-scale point. Form no longer dominates over content. The emergence of a significant validity–centrality interaction in the first-occurrence data is particularly salient, given that the number of measurements per subject in first occurrence as opposed to overall data drops from 112 to 16. The fact that varying stimulus duration from 4 to 10 seconds does not significantly affect the interaction suggests that use of prototypicality heuristics is rather stable at least with respect to this variable, and does not require an exotic time regime.

On the evolution of processing strategies from first occurrence to later tasks, the above analyses suggest that in this experimental setting subjects were using prototypicality heuristics for the first-occurrence tasks (which were as late as #30 in the randomly ordered trial series) but had already ceased to do so detectably by the end of the first 56-trial block. Grand mean first-occurrence error rate was 18.8%, while the overall rate was lower, 15.3%, as would be expected from practice. On the nature of the difference between the conjectured two kinds of processing, although both Experiment 2 error rates are higher than the 15.2% first occurrence and 13.3% overall reaction time rates, subjects were not guessing here even on the first-occurrence tasks. Subjects could not achieve such good first-occurrence or overall performance if in either case they ignored the form of an infer-

ence—the logical constants involved—or its content—the category–instance relationship.

One cannot expect that decreasing available time will just directly shift subjects' processing toward more use of faster heuristics. For, subjects at least at first are unlikely to be able to make individual inference by inference speed–accuracy trade-off decisions, because they would initially have a limited inductive basis for these decisions. In fact, subjects required a mean 5.7 seconds with 15.2% errors for their first-occurrence reaction times where they were taking as much time as they thought they needed, and exhibited no prototype effect; however, with available time nearly doubled in the first interval experiment to 10 seconds, there was a higher first-occurrence error rate of 16.4% and some indication of a prototype effect, $F(1,28) = 4.08, p < .06$. A more adequate explanation of the observed time course of the prototype effects therefore would be that subjects initially make the longer term decision that, for this new task with a fixed time limit, prudence requires use of quick but dirty prototypicality heuristics. Then, as they become more familiar with the task in the course of the fixed interval experiment, practice makes them more proficient (if not perfect) and so they tend to decide that they can afford more use of slower but correct procedures and readjust their overall strategy in that direction.

However, practice here might have a quite different effect. Once subjects have determined an instance–category relationship at the beginning of the experiment, they might store this information temporarily, so that the next time they encounter that instance–category pair they do not have to recompute the relationship again; the effects of instance prototypicality would therefore subsequently disappear. (The two phenomena, of a shift to correct inference processing strategies and of using temporarily stored category–instance relationships, need not be incompatible.) How-

ever, since even the first-occurrence items in the quite sensitive tachistoscope experiment showed no prototypicality effect at all, the disappearance of the effect in Experiment 2 seems unlikely to be entirely due to subjects' determining category-instance relationships. Nonetheless, more information about subjects' "metaheuristic" decisions between different deductive procedures would be valuable.

EXPERIMENT 3: EARLY USE OF THE HEURISTIC

When and how do people decide to use the prototypicality heuristic? The primary objective of this experiment was to obtain finer grained information about the item-to-item time course of use of the heuristic, in particular, about subjects' earliest use of it when presented with a series of logic problems. Another goal was to find out more about how widespread use of the heuristic is: to study subjects' use of it on types of material different from Experiments 1 and 2, and under somewhat different experimental conditions—with an exposure duration of 8 seconds instead of 4 or 10, and where subjects had a page-turning task rather than viewing items on a screen.

In addition, the new types of inference were chosen to provide a chance to check a further conjecture: The more complex the deductive task, the more pressure for speed-reliability trade-off, and so the more likely subjects will decide to use quick and dirty procedures—in particular, the more likely their use of the prototypicality heuristic. There should be a three-way interaction between complexity of inference, validity, and instance centrality; the more complex the inference, the more prototype effect there might be. People seem to possess a common-sense theory of the feasibility of deductions which enables them to identify intuitively a rough ordering of deductive tasks by difficulty (Cherniak, 1981b); the one-premise tacit syllogisms of Experiments 1 and 2 appear to fall toward the simpler end of such a scale. An ob-

viously more complex, but closely related, task would be just a two-premise tacit syllogism (see Table 5).

Method

Materials. Half of the arguments presented were drawn from the one-premise "simple" syllogisms of Experiments 1 and 2, namely Nos. 1, 2, 5, and 6 of Table 1; the other half were the "complex" two-premise syllogisms of Table 5. Each complex inference was chosen to be as close as possible in logical form to the corresponding simple inference, with an additional "All A's are B's" transitive-chain premise. Thus, equal numbers of the simple and of the complex inferences are valid and are invalid, and in turn half of the valid and of the invalid inferences each involve either universal or particular propositions. For a balanced block design, eight categories were needed. The seven categories of Experiments 1 and 2 were used; in addition, since this is not a tachistoscope experiment, the category of "toy," with "yo-yo" as central and "seesaw" as peripheral instances, was also drawn from Rosch (1975) (it satisfies the criteria from categories in Experiment 1, except that the central instance is rated only 1.62).

Each of the eight forms of the syllogism was completed with six different category/central-instance pairs, and with six different category/peripheral-instance pairs. Each of these 96 items was typed in the format of the earlier experiments and photocopied on 34 sheets of paper. On the back of each sheet were two 7-point confidence scales, one titled "VALID," the other "INVALID," with "1" also labeled "not sure at all," "4" "fairly sure," and "7" "absolutely sure." An additional four practice items were constructed from syllogisms and categories not already used.

Procedure. Each subject received 48 uninterrupted items in three 16-item blocks. In each block all eight syllogisms occurred, each in one peripheral and one central version (never from the same category); the

TABLE 5
COMPLEX FORMS OF THE SYLLOGISM USED IN EXPERIMENT 3

Valid (with tacit premise "All [instance] are [category]")	
(9) All A's are B's. <u>All B's are [instance].</u> All A's are [category].	(10) Some [instance] are A's. <u>All A's are B's.</u> Some [category] are B's.
Invalid (with tacit premise "All [instance] are [category]")	
(11) All [instance] are A's. <u>All A's are B's.</u> All [category] are B's.	(12) Some [category] are A's. <u>All A's are B's.</u> Some [instance] are B's.

first 16 subjects had one such inference-category assignment, the next 18 had an entirely different one. Each block was therefore balanced with respect to all variables: inference complexity, validity, universality, and instance centrality. The first 8 subjects each received the 48 items in a different random order, with the constraints that no two consecutive items could be of the same form of the syllogism or involve the same category, and no four consecutive items could all be the same with respect to either complexity, validity, universality, or centrality. The constraints were satisfied for the first 8 subjects in a manner that permitted them also to be satisfied for the next 8 subjects, each of whom received the items of each block in the reverse order from a corresponding one of the first 8 subjects. This design has the property that within each block, the set of all items received by all 16 subjects in the first eight trials is identical with the corresponding set for the last eight trials; each block can be split in half in this way. The next 16 subjects each exactly corresponded to one of the first 16, except that each syllogism in each block had a category assigned to it different from that for the first 16 subjects. Two additional subjects received copies of the series for two of the second 16 subjects.

Each 48 item set, preceded by the same four practice items, was presented as a booklet, with the argument side of each page facing upward, and blank spacer pages covering each item. A clock-actuated tone generator operated on a repeated cycle of

8 seconds of silence, during which subjects were to turn a cover page and look at an argument, followed by a 6-second audible tone, during which subjects were immediately to turn over the page and circle on the back the digit indicating their confidence either that the argument was valid or invalid. The concept of validity was explained to the subjects as in the earlier experiments.

Subjects. Thirty-four Tufts University undergraduates participated in the experiment for psychology course credit; they were required to be native English speakers who had taken no college logic course.

Results and Discussion

The predicted validity–centrality interaction again seems to occur; its time course agrees with Experiment 2. In addition, some evidence suggests that subjects do not decide to make the fullest use of the prototype heuristic at the very beginning of the problem series.

Two additional subjects' data had to be discarded because they reported during debriefing that they were unable to follow the page-turning instructions. The error and confidence scores of each response were combined as a single 14-point scale rating: "1" if the subject was wrong and absolutely sure, "14" if he was right and absolutely sure, etc. An analysis of variance for all three blocks combined was performed, with complexity, validity, and universality of inference and instance centrality as within-subject variables (see Table 6a). The validity–centrality interaction for

TABLE 6
EXPERIMENT 3: EFFECTS OF INFERENCE FORM ON RESPONSE

14-Point confidence rating				<i>F</i>	<i>p</i> <
(a) Overall				<i>df</i> = 1,33	
Simple	11.479	Complex	10.375	24.78	.0001
Universal	11.430	Particular	10.424	6.87	.013
(b) Block 1				<i>df</i> = 1,33	
Simple	11.040	Complex	9.820	19.39	.0001
Universal	11.125	Particular	9.735	11.93	.002
(c) Second half, Block 1				<i>df</i> = 1,16	
Simple	11.191	Complex	10.500	3.75	.071
Universal	11.382	Particular	10.309	5.30	.035

all 48 items (see Table 7a) is now significant ($F(1,33) = 7.79, p < .009$), in contrast with its nonsignificance for 112, or even just 56, of the items in Experiment 2. The conjectured three-way interaction of simplicity, validity, and centrality did not approach significance, $F(1,33) = 0.32, p > .57$. Block 1 performance shows the same pattern (see Table 6b). The validity–centrality interaction (see Table 7b) for these first 16 items is equally significant ($F(1,33) = 7.98, p < .008$), even though only one-third of the original data is in the analysis. Also, the magnitude of the interaction is now nearly double what it is for all three blocks. And once more, the simplicity–validity–centrality interaction did not reach significance, $F(1,33) = 0.11, p > .73$. The validity–centrality interaction is not significant in Block 2 alone, $F(1,33) = 0.55, p > .46$, nor in Block 3 alone, $F(1,33) = 2.42, p > .12$.

As noted above, each block was designed so that, for 32 subjects, its first and second half contained equal numbers of simple and complex, valid and invalid, universal and particular, and central and peripheral items. A sign test of each of these six eight-trial subblocks was performed to see whether mean central-peripheral differences for each of the eight inferences were in the right directions for the predicted validity–centrality interaction. In the second half of Block 1, all eight of these differences are in the predicted directions, confirming the va-

lidity–centrality interaction, $p < .004$, even though the data have been reduced yet again. In each of the other half-blocks, either three or four of the central–peripheral differences are in the wrong direction, approaching chance.

Although in a given half-block, each subject received only 8 of the 16 inference–instance combinations, design of the experiment permitted yoking pairs of subjects so that 16 subject pairs had each received all 16 items. An analysis of variance was then performed for these 16 yoked subjects' (and the remaining two subjects') responses in the second half of Block 1. The pattern resembles that of all three blocks and of Block 1 alone: The “calibration” variables of simplicity and universality reach or approach significance (see Table 6c). The validity–centrality interaction (see Table 7c) is now more significant ($F(1,16) = 8.39, p$

TABLE 7
EXPERIMENT 3: VALIDITY × CENTRALITY,
14-POINT RATINGS

	Central	Peripheral
(a) Overall		
Valid	11.556	10.865
Invalid	10.627	10.659
(b) Block 1		
Valid	11.324	10.044
Invalid	10.184	10.169
(c) Second half, Block 1		
Valid	11.912	10.559
Invalid	9.971	10.941

< .011) and of greater magnitude than the simplicity or universality effects; the magnitude of the validity–centrality interaction in fact is yet again nearly double what it was for all of Block 1. The effect of instance centrality on valid inference performance by itself is significant, $F(1,16) = 5.96$, $p < .027$, and now the effect of centrality on the invalid inferences alone also approaches significance, $F(1,16) = 3.30$, $p < .088$. The simplicity–validity–centrality interaction still does not approach significance, $F(1,16) = 0.001$, $p > .98$. The other side of the nonsignificance of this interaction is that it does suggest generality of prototypicality effects—they are certainly no weaker for complex than for simple inferences.

What seems to be the early time course of subjects' use of the prototypicality heuristic? First, the data suggest that prototypicality effects do not solely occur in trials 9 through 16: Not only is the validity–centrality interaction significant for all three blocks combined, but also there is no significant three-way interaction between the validity–centrality effect in Block 1 vs 2, $F(1,33) = 1.54$, $p > .22$, nor is the corresponding three-way interaction significant for Block 1 vs 3 performance, $F(1,33) = 1.50$, $p > .23$. Nonetheless, there seems to be some concentration of use of the prototypicality heuristic in the second of the six half-blocks. And this turns out to agree rather well with the results of Experiment 2: the “first-occurrence” items of Experiment 2, which showed the validity–centrality interaction, had a mean trial number of 12, while the second half-block in fact has the most closely matching mean trial number of 12.5 (this match also holds when the two more practice items in Experiment 2 than in Experiment 3 are taken into account).

Furthermore, there is an interesting correspondence between the use of the prototypicality heuristic and the evolution of error (and mean confidence) scores (see Table 8). The most salient feature of this data is the drop in error rate, of 9.4%, that

TABLE 8
MEAN ERRORS AND CONFIDENCE RATINGS BY
HALF-BLOCKS FOR 32 SUBJECTS

Block	1		2		3		Grand mean
	(a)	(b)	(a)	(b)	(a)	(b)	
Percent error	29.7	20.3	21.1	18.4	17.2	19.1	21.0
14-Point scale	9.9	10.8	10.8	11.1	11.2	11.2	10.8

occurs from the first to the second half-block; error rates for the second through sixth consecutive half-blocks each change by less than 2%, and in fact stop dropping and begin to rise at points. The drop in this rough step function thus occurs when subjects seem to be first manifesting the most use of the prototypicality heuristic. The apparent shift toward more use of the heuristic certainly does not lead to more errors here. Furthermore, proceeding with some caution, we can say that, given the hypothesis that the prototypicality heuristic is “dirty but quick”—that is, it is significantly faster, if less reliable, than formally correct procedures—a natural interpretation of this possible connection between use of the prototypicality heuristic and the drop in error rate would be that in the first and second half-blocks, use of the prototypicality heuristic results in considerably fewer errors. The data at least suggest that, in this sense, the prototypicality heuristic “works” at the beginning of the experiment: it is fast and yet better than just guessing answers, as one might otherwise have to do if there were not time to complete the formally correct procedure. Subjects' early performance in the experiment provides some preliminary evidence that the benefits of using the heuristic outweigh the costs at that point; it is not irrational sloppiness to use it.

However, there also seems to be less use of the heuristic in the third through sixth half-blocks than in the second half-block, yet the error rate changes little. The best interpretation of this seems to be the earlier one for the 56 and 112 trial results of Ex-

periment 2: By then subjects have had enough practice with the experimental task, and are sufficiently proficient with the formally correct procedures, so that these procedures are now fast enough to be used successfully within the time limitations of the experiment. The subjects therefore decide to shift toward more use of them.

A final feature of the finer grained information of Experiment 3 also concerns how subjects identify appropriate situations for using the prototypicality heuristic: as noted, subjects do not appear to make the fullest use of that heuristic at the very beginning of the experiment—in the first half of Block 1. The fact that the validity–centrality interaction becomes significant from the first to the second half of Block 1 seems to exclude an explanation of the time course of the interaction in terms of subjects just (over)learning and storing the category–instance relationships. The evidence seems to suggest that instead, subjects are taking about eight experimental trials, in addition to the four practice trials, to come most fully to the decision to use the prototypicality heuristic. Thus, an experimental task that is putatively concerned only with “pure” deductive reasoning in fact, like many others in cognitive psychology, also involves an inductive task: subjects may be using the items they have already encountered as a basis for reaching a “metaheuristic” conclusion about what is an appropriate deductive strategy.

GENERAL DISCUSSION

The above experiments provide some support for the hypothesis that people use a formally incorrect deductive reasoning heuristic that exploits the prototypicality structure of their mental representations of categories. Subjects’ use of this prototypicality heuristic should yield a validity–centrality interaction, and this predicted interaction does in fact occur under interesting conditions. Furthermore, there is some preliminary evidence that using the heuristic is sometimes rational: it seems to pay off with

lower error rates. The experiments suggest that it would be interesting to examine the time course, and effects of limiting subjects’ time, for typicality phenomena in other domains—notably, sentence verification and nondeductive reasoning.

Furthermore, given only an argument such as “Some A’s are birds. / Therefore, some A’s are robins.” the subject can interpret the task in at least three different ways, as (i) inductive, (ii) pragmatic, or (iii) deductive inference. First, the subject might regard himself as being asked: Given the premise, what is the probability of the conclusion (e.g., is it more likely than not)? The task then starts to approximate one used by Rips (1975) in the experiment on typicality effects in inductive reasoning mentioned earlier, which was based on Kahneman and Tversky’s studies of the representativeness heuristic, and which indicates that subjects inductively generalize more easily from more “representative” (i.e., here, typical) instances. Minimizing content by use of variables like “A” does not by itself exclude probabilistic reasoning.

But more interestingly, the subject might instead behave as if he were being asked: If a speaker utters the premise by itself, would this be *misleading* if he knows the conclusion is false? Prototypicality probably has important roles in the pragmatics of language understanding. Grice (1975) proposed some pragmatic maxims we seem to use in interpreting conversation, and Clark has developed a theory of the inferences we make by means of them (e.g., Clark & Haviland, 1977). Subjects’ use of these maxims may lead them to make inferences which are logically invalid but are pragmatically well-founded implicatures in natural discourse. We can hypothesize that one conversational maxim involving prototypicality that is relevant here would be, “Given an indefinite assertion about a category (for example, of the form, ‘Some [category] are . . .’), assume the assertion applies to the more central instances of the

category, other things being equal; that is, infer 'Some [central instances] are. . . .'' One "psychological" rationale for this principle would be based on a common-sense cognitive psychology hypothesized to be accepted by speaker and hearer: It would be misleading for the speaker to make the indefinite assertion if he knew it did not apply to central members, because the speaker can expect the hearer to be most likely to think (first) of central members and least likely to think of more peripheral members. (For a review of evidence that subjects in fact do think first of central members, see Rosch, 1977.) A second "semantic" rationale would be that peripheral members of the category are members of it to a lesser degree than central members.

Since the present study was concerned with the basic process of deductive reasoning—which should constitute an element of the more complex processes of inductive and pragmatic inference—the initial goal was to isolate the former process by trying to reduce the latter two. The above tacit syllogisms seemed less likely to elicit pragmatic effects than more natural passages of connected discourse. However, a future question concerns the role of prototypicality in pragmatics.

It is worth concluding with a brief survey of the wider import of studies of formally incorrect deductive heuristics. The rise of modern logic can be characterized to a considerable extent as a response to the discovery, and recognition of the significance, of the classical antinomies of semantics and set theory (see, e.g., Quine, 1966). This modern era began with the realization of the importance for semantics of the ancient "Liar Paradox" (a simplified form is "This sentence is false"), and the discovery of Russell's Paradox for set theory. It became apparent that a broad range of such inconsistencies was embedded in our conceptual scheme, lurking in the foundations of logic and mathematics, even in ordinary discourse. The classical response to the par-

adoxes, which conditions metamathematics to the present day, is to view them as the most severe sort of threat to the deductive sciences: As the logician Alfred Tarski says, "The appearance of an antinomy is for me a symptom of disease" (Tarski, 1969). Tarski's remark in fact recalls a minority opinion of Ludwig Wittgenstein, expressed 30 years earlier, rejecting the view that "finding a contradiction in a system . . . shows that the whole system is diseased" (Wittgenstein, 1976). Instead, Wittgenstein proposed that such formal incorrectness need not vitiate a deductive system.

The experiments reported here are part of an exploration of a more extreme hypothesis than Wittgenstein's: that, in important ranges of conditions, formal incorrectness may be healthy, a positive virtue (Cherniak, 1984). The remarkably widespread occurrence of such simple paradoxes in our conceptual scheme may not be because of regrettable sloppiness, but because they are symptoms of formally incorrect deductive reasoning heuristics we use which work well in the everyday non-technical situations we usually encounter. Furthermore, use of formally correct procedures here might be paralyzingly costly.

Recent results in the field of computational complexity theory seem to support the conjecture of an extreme antagonism between formal correctness and practical adequacy: Very briefly, a perfect decision procedure exists for validity of inferences in, for example, the monadic predicate calculus (which includes syllogistic inference)—that is, an algorithm that always correctly decides for any such inference whether or not it is valid. But this procedure has been shown to require "nondeterministic exponential time" (Lewis, 1978). That is, the computation time required on even a machine with a kind of unbounded parallelism grows in the worst cases as an exponential function of the size of the given problem. This suggests that, as a practical matter, such an algorithmic procedure

sometimes may not be tractable. Even decision procedures for just tautological inference are known to be "nondeterministic polynomial complete," and are therefore strongly conjectured to require exponential time (Garey & Johnson, 1979).

In cognitive psychology, a postulated process is expected not to require infinite resources; in a similar way, computational complexity constraints may sometimes be relevant to psychological theory. The applicability of complexity theory to actual human deductive reasoning is an interesting area for study, since the population distribution of "hard" cases of the intractable problems is not yet well understood. However, acknowledged difficulties of exponential explosion for algorithmic approaches to computer theorem proving (and verification of program correctness) (Newell & Simon, 1976; Rabin, 1974) already suggest that complexity constraints may in fact sometimes arise early and often enough to constitute serious practical limitations for algorithmic deduction procedures. And although there are no complexity estimates for nondeductive reasoning, the many recent experiments suggesting surprisingly widespread formal incorrectness in everyday nondeductive reasoning (Nisbett & Ross, 1980) convergently support such a working hypothesis; indeed, it would provide a basic element of a unifying framework for these results. Departures from deductive correctness therefore might constitute not regrettable irrationality, but the most fundamental speed-reliability trade-off: The compromise with perfect algorithmhood of quick but dirty heuristics may in important cases be indispensable to evade computational paralysis. In this way there is an interesting connection between empirical studies of putative "irrationality," complexity theory, and the development of less idealized theories of rationality.

REFERENCES

- CHERNIAK, C. (1981). Minimal rationality. *Mind*, **90**, 161-183. (a)
- CHERNIAK, C. (1981). Feasible inferences. *Philosophy of Science*, **48**, 248-268. (b)
- CHERNIAK, C. (1984). Computational complexity and the universal acceptance of logic. *Journal of Philosophy*, in press.
- CLARK, H., & HAVILAND, S. (1977). Comprehension and the given-new contract. In R. Freedle (Ed.), *Discourse production and comprehension*. Norwood, N.J.: Ablex.
- GAREY, M., & JOHNSON, D. (1979). *Computers and intractability: A guide to the theory of NP-completeness*. San Francisco: Freeman.
- GRICE, H. (1975). Logic and conversation. In D. Davidson & G. Harman (Eds.), *The logic of grammar*. Encino, Calif.: Dickenson.
- HADAMARD, J. (1945). *The psychology of invention in the mathematical field*. Princeton: Princeton Univ. Press.
- HENLE, M. (1962). On the relation between logic and thinking. *Psychological Review*, **69**, 366-378.
- JOHNSON-LAIRD, P., & STEEDMAN, M. (1978). The psychology of syllogisms. *Cognitive Psychology*, **10**, 64-99.
- KAHNEMAN, D., & TVERSKY, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive psychology*, **3**, 430-454.
- KAHNEMAN, D., & TVERSKY, A. (1973). On the psychology of prediction. *Psychological Review*, **80**, 237-251.
- KUCERA, H., & FRANCIS, W. (1967). *Computational analysis of present-day American English*. Providence R.I.: Brown Univ. Press.
- LEWIS, H. (1978). Complexity of solvable cases of the decision problem for the predicate calculus. *Proceedings of the 19th IEEE Symposium on Foundations of Computer Science*, 35-47.
- NEWELL, A., & SIMON, H. (1976). Computer science as empirical inquiry. *Communications of the Association for Computing Machinery*, **19**, 113-126.
- NISBETT, R., & ROSS, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, N.J.: Prentice-Hall.
- PAIVIO, A. (1971). *Imagery and verbal processes*. New York: Holt, Rinehart & Winston.
- POLYA, G. (1954). *Mathematics and plausible reasoning* (Vols. 1 & 2). Princeton: Princeton Univ. Press.
- QUINE, W. (1966). *The ways of paradox*. New York: Random House.
- RABIN, M. (1974). *Theoretical impediments to artificial intelligence*. In J. Rosenfeld (Ed.), *Information processing 74*. Amsterdam: North-Holland.
- RIPS, L. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior*, **14**, 665-681.
- RIPS, L., SHOBE, E., & SMITH, E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*, **12**, 1-20.
- ROSCH, E. (1973). On the internal structure of perceptual and semantic categories. In T. Moore (Ed.),

- Cognitive development and the acquisition of language*. New York: Academic Press.
- ROSCH, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, **104**, 192-233.
- ROSCH, E. (1977). Human categorization. In N. Warren (Ed.), *Studies in cross-cultural psychology* (Vol. 1). New York: Academic Press.
- ROSCH, E., & MERVIS, C. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, **7**, 573-605.
- ROSCH, E., MERVIS, C., GRAY, W., JOHNSON, D., & BOYES-BRAEM, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, **8**, 382-439.
- SMITH, E., SHOEN, E., & RIPS, L. (1974). Structure and process in semantic memory: A featural model for semantic decision. *Psychological Review*, **81**, 214-241.
- TARSKI, A. (1969). Truth and proof. *Scientific American*, **220**, 63-77.
- WASON, P., & JOHNSON-LAIRD, P. (1972). *Psychology of reasoning: Structure and content*. Cambridge, Mass.: Harvard Univ. Press.
- WILKINS, M. (1928). The effect of changed material on the ability to do formal syllogistic reasoning. *Archives of Psychology*, **16**, (102).
- WITTGENSTEIN, L. (1976). *Lectures on the foundations of mathematics*. Ithaca, N.Y.: Cornell Univ. Press.

(Received July 15, 1983)

(Revision received October 27, 1983)