Bulletin of the Psychonomic Society
1974, Vol. 4 (3), 188-190

The effect of problem difficulty on hypothesis testing and an extension of Levine's theory*

J. DAVID McKEE†
University of Vermont, Burlington, Vermont 05401

Working hypothesis (Hs) and majority rules are examined under four conditions of problem difficulty achieved through combination of exposure or nonexposure of problem cards and informative feedback with two or three values per stimulus dimension. Memory aids facilitate solutions; intradimensional variability has no effect. Working Hs occur more often in exposed conditions; majority rules are equally distributed throughout conditions. Ss demonstrate H behavior on 91.6% of all opportunities; working Hs are more prevalent than majority rules; and the size of the H pool supports Levine's subset sampling assumption. However, some Ss change Hs before and after "right" feedback. Analysis of such discrepant findings suggests an extension of Levine's model.

At the beginning of simple discrimination problems, Ss select a subset of all possible hypotheses (Hs), according to Levine's (1970) subset sampling assumption. In most cases, Ss adopt one subset H as a "working hypothesis" which is tested on the initial trials of a problem (the working-H rule). Information obtained from testing the working H is used to evaluate it as well as other subset Hs.

Levine describes another response rule, the "majority rule." With this rule, S's overt responses are determined by a combination of subset Hs; S makes "that response which, he believes, has the best chance of success [1970, p. 403]."

Examined here are the effects of problem difficulty on Ss' use of working-H and majority rules. Difficulty is manipulated through orthogonal combination of (a) exposure or nonexposure of problem cards and feedback information during problem solving, and (b) two or three values per stimulus dimension. Memory aids facilitate solutions in hypothesis-testing problems (e.g., Eimas, 1970); problem difficulty in certain concept-identification paradigms is inversely related to intradimensional variability (e.g., Haygood, Harbert, & Omlor, 1970; Schultz & Dodd, 1972).

METHOD

Procedure

The procedure replicates Levine's (1966) standard paradigm except that three additional problem conditions are studied. Following detailed instructions, Ss receive four practice problems for which correct solutions represent each stimulus dimension. Eight 16-card test problems, selected at random from a pool of 16, are presented to each S. For each problem, Cards 1, 6, and 11 are feedback or outcome trials (OTs) in which E informs S about the correctness of his response. Cards 2-5, 7-10, and 12-15 are blank trials (BTs) during which E provides no information about responses.

Feedback following S's responses on Cards 1 (OT_1), 6 (OT_2), and 11 (OT_3) is presented according to the eight different combinations of "right" (+) and "wrong" (−). The order of such sequences for the eight test problems is selected randomly for each S. Feedback on Card 16 (OT_4) is "right" on half of the trials; no feedback is presented on remaining trials. Ss in two-value conditions receive one problem involving each feedback sequence. For three-value conditions, sequences (+++) and (++−) are omitted because of informational inconsistencies that result from reduction of nine three-value BT cards to just four, as explained below. Therefore, two feedback sequences, selected randomly from those beginning with "wrong" on OT_1 are repeated for each S in three-value conditions.

In exposed conditions, each problem card remains face upward in front of S as it is presented. Feedbacks on OTs are indicated by a plus sign for "right" and a minus sign for "wrong," each placed next to the appropriate exposed OT card.

Ss' responses to each problem card are recorded. Patterns of responses during the three BT probes per problem are determined to be consistent or not consistent with an unidimensional H such as "large size." Inconsistent response patterns during a BT probe are further inspected for instances of majority rules. Majority-rule patterns reflect the intersection of at least three unidimensional values across four BT responses (Levine, 1970, p. 403).

Stimuli

Problem cards were 15.2 cm x 10.2 cm white cards on which two or three letters were pasted for two- and three-value problems. Each letter on a particular problem card represented one value on dimensions of color, letter, position, and size. Arrangement of stimulus values on each problem card was dictated by Levine's rules for "internally orthogonal" problem-card construction (1966, p. 333).

Applied to three-value problems, such rules produce nine cards instead of four in two-value problems. Since the concern is hypothesis testing under various conditions of problem-card construction, but with comparable formal structure across problems, four three-value cards were chosen for BT probes so they duplicate as far as possible the relationships between and within dimensions for two-value BT-probe problem cards. For example, between any two two-value cards, values on two dimensions change, two do not (a 2:2 shift). In three-value problems, all BT cards represent 2:2 shifts except one pair that

*This paper is based on a master's thesis submitted to the Graduate College of the University of Vermont. An earlier version was presented at the annual convention of the Eastern Psychological Association, Boston, April 1972. Thanks are due A. E. Goss, who sponsors this paper and takes full editorial responsibility, D. C. Howell, and E. D. Neimark for their comments and assistance.

†Current address: Department of Psychology, Douglass College, Rutgers University, New Brunswick, New Jersey 08903.

involves a 3:1 shift. Between each OT problem card and each BT problem card, there were 2:2 shifts for two-value problems and 1:3 shifts for three-value problems.

Subjects

Ss were 48 undergraduate volunteers who were tested in individual S-paced sessions that lasted approximately 1 h. Ss were paid $2 at the completion of a session. Ss were assigned randomly in equal numbers to each of the four conditions with sexes evenly distributed throughout.

RESULTS AND DISCUSSION

Formal structural uniformity of all problems across conditions produces several three-value problems that end with more than one correct H after three OTs for "ideal" hypothesis testers. Therefore, a response during OT_4 in three-value problems is considered a correct solution if it (a) is predicted by H_3 and (b) has not been infirmed logically by any previous information.

There were 1152 BT-probe response patterns, of which 82.7% are consistent with unidimensional Hs. Of an equal number of OT responses, 85.7% are predictable from the immediately preceding H. Analyses of variance indicate that neither consistent Hs nor predictable OT responses is influenced by problem difficulty.

Effects of Problem Conditions

Correct solutions in exposed conditions exceed those in nonexposed conditions, $F(1,40) = 14.42$, $p < .01$ (Table 1). Likewise, illogical Hs (i.e., Hs previously infirmed by preceding problem information) occur more often in nonexposed than in exposed conditions, $F(1,40) = 16.52$, $p < .01$ (Table 1). Thus, problems are easier to solve with memory aids; intradimensional variability has no effect on problem difficulty.

The facilitative effect of memory aids on performance in exposed conditions is supported also by an interaction of Exposure by BT Probe for illogical Hs, $F(2,80) = 8.10$, $p < .01$. Illogical Hs increase during successive BT probes within problems of nonexposed conditions ($M_{BT1} = 0.96$, $M_{BT2} = 1.46$, $M_{BT3} = 2.50$) but decrease slightly within problems of exposed conditions ($M_{BT1} = 0.87$, $M_{BT2} = 0.71$, $M_{BT3} = 0.67$). Apparently, Ss are less likely to select illogical Hs when reference to previously presented problem cards is possible.

Interactions of Values per Dimension by Exposure with Sex occur for correct solutions, $F(1,40) = 10.51$, $p < .01$, and for illogical Hs, $F(1,40) = 7.11$, $p < .05$. Table 2 shows that females perform better than males in two-value nonexposed and in three-value exposed conditions, whereas males are superior to females in two-value exposed and in three-value nonexposed problems. There is no apparent explanation for these interactions.

Table 1
Means for Measures of H Testing in Problem Conditions

	Problem Condition			
	Two-Value		Three-Value	
Measure	Exposed	Not Exposed	Exposed	Not Exposed
Correct Solutions	6.83	4.83	5.83	5.42
Illogical Hs	0.75	1.94	0.75	1.33
Majority Rules	1.33	2.00	2.75	2.42
Working-H Rules	18.33	14.50	17.58	14.33

Response Rules

Levine's definition of working H is not completely explicit with respect to its effect on subsequent OT responses. However, he does suggest that such Hs are retained or rejected according to informative feedback. Thus, a working H is defined here as a logically consistent H, inferred from S's response pattern during a BT probe, which exactly predicts S's response on the following OT. (Responses on OTs following majority-rule response patterns are similarly predictable.)

Of the 1152 BT-probe response patterns, 67.4% are working Hs, 8.9% are majority rules. Majority rules are distributed equally among conditions (Table 1). Working Hs occur more frequently in exposed than in nonexposed conditions, $F(1,40) = 8.61$, $p < .01$ (Table 1).

There were 108 inconsistent response patterns following OT_1, 48.1% of which are majority rules. In the two-value nonexposed condition which corresponds to Levine's (1966) paradigm, 55.5% of the inconsistent response patterns are majority rules, compared to 38.8% in Levine (1970). Majority rules in all conditions occur more frequently than expected by chance.

In summary, 91.6% of all BT probes represent either consistent Hs or majority rule responses. This percentage corresponds favorably with 92.4% in Levine's (1966) results, which were based on more than twice as many BT probes. Furthermore, Levine's conception of two kinds of response rules is supported, with working-H rules occurring more frequently than majority rules. Finally, support for the subset sampling assumption is presented in Table 3, which shows the H-pool size within each problem condition.

Effects of Informative Feedback on Hypothesis Changes

Levine (1966) is explicit about predicting effects of informative feedback on S's hypothesis-testing behavior.

Table 2
Means of Correct Solutions and Illogical Hs for Females (F) and Males (M) in Problem Conditions

		Problem Condition			
		Two-Value		Three-Value	
Measure		Exposed	Not Exposed	Exposed	Not Exposed
Correct Solutions	F	6.17	5.83	6.17	3.50
	M	7.50	3.83	5.50	5.33
Illogical Hs	F	1.00	1.78	0.50	1.83
	M	0.50	2.11	1.00	0.83

Table 3
Size of H Pool in Problem Conditions Following "Wrong" Feedback According to OT_i

Problem Condition	OT_1	OT_2	OT_3
Two-Value, Exposed	5.05	2.23	1.17
Two-Value, Not Exposed	4.37	2.80	1.88
Perfect Focusing	4.00	2.00	1.00
Three-Value, Exposed*	7.48	4.60	2.48
Three-Value, Not Exposed*	8.47	8.10	3.67
Perfect Focusing*	8.00	4.00	2.00

**Values computed only for those feedback sequences beginning with (−).*

If H_i is confirmed by "right" on OT_{i+1}, then H_{i+1} will be the same as H_i; if H_i is infirmed by "wrong" on OT_{i+1}, then H_{i+1} will not be the same as H_i. In other words, Ss should change Hs following "wrong" and retain Hs after "right." The following conditional probabilities from Levine (1966) strongly support these assumptions: $P(H_i = H_{i+1} \mid -) = 0.02$ and $P(H_i = H_{i+1} \mid +) = 0.95$. Levine's criteria for computing these probabilities are: (a) "right" or "wrong" occurs on OT_{i+1}, which intervenes between H_i and H_{i+1}; (b) H_i and H_{i+1} are both consistent Hs; and (c) the response on OT_{i+1} is predictable from H_i (1966, p. 334).

These criteria applied to the present results yield $P(H_i = H_{i+1} \mid -) = 0.024$ and $P(H_i = H_{i+1} \mid +) = 0.893$ across all problem conditions. These probabilities agree with Levine's. However, violation of Levine's criterion that OT_{i+1} responses must be predicted by the immediately preceding H_i (the predictability criterion) leads to an extension of Levine's model.

Violation of the predictability criterion yields $P(H_i = H_{i+1} \mid -) = 0.060$ and $P(H_i = H_{i+1} \mid +) = 0.810$, the latter especially discrepant from Levine's values. Violation of this criterion is not important for Levine's data since 97.5% of OT_{i+1} responses were predicted by preceding H_is. In the present results, however, OT_{i+1} predictability is at 85.7%, representing 987 cases. Evidently, some Ss who demonstrate consistent H_is on the ith BT probe test some alternative H on the subsequent OT_{i+1}.

Of 250 cases in which $H_i \neq H_{i+1}$ with "wrong" intervening, 2.8% are excluded when the predictability criterion is applied. Of 51 cases in which $H_i \neq H_{i+1}$ with "right" intervening, 49% of such H changes are excluded when the predictability criterion is applied.[1] In other words, for half of the cases where $H_i \neq H_{i+1}$ with "right" intervening, Ss' OT_{i+1} responses are not predictable from H_i. Thus, some Ss change Hs before OT responses; others change Hs after OT responses.

In general, changes of Hs are not random. Of 301 such switches, 78.7% are shifts from logical Hs (i.e., Hs confirmed by previous problem information) or illogical Hs on the ith BT probe to logical Hs on the ith + 1 probe. Nor are H changes unevenly distributed among either Ss or problem conditions. With regard to H changes with intervening "right" on OT_{i+1}, 82% are shifts from logical or illogical Hs to logical Hs.

It may be concluded that Ss do not always perform according to the assumptions of Levine's model. There is support for Suppes and Schlag-Rey's (1965) contention that Levine's "no-change-if-no-error" assumption is not completely correct.

Some of the discrepancy between these and Levine's results is resolved through an extension of the model. Suppose that, following H_i, S decides to test not-H_i (i.e., the complement of H_i—an intradimensional shift) on the subsequent OT_{i+1}. For example, if H_i is "large," S might test "small" instead on OT_{i+1}. Such responses would be reflected in S's protocol as a violation of the predictability criterion. If testing not-H_i results in "right" feedback, then S will discard H_i. If not-H_i is "wrong," then S will retain H_i. Thus, Levine's model can be expanded to include the decision matrix indicated below for S's choice of H_{i+1} given the possible outcomes of his hypothesis-testing behavior on OT_{i+1}. Shown also are conditional probabilities and, in parentheses, the number of cases on which they are based.

OT_{i+1} Feedback	H Tested on OT_{i+1}: H_i	Not-H_i
"Right"	Keep H_i .893 (243)	Reject H_i 1.000 (25)
"Wrong"	Reject H_i .976 (249)	Keep H_i .588 (17)

Of the 25 responses on OT_{i+1} that are not predicted from the preceding H_i when intervening feedback is "right," all represent cases in which not-H_i is tested. The proposed extension of the model completely erases the discrepancy between the two computations of $P(H_i = H_{i+1} \mid +)$ for the present results described earlier.

REFERENCES

Eimas, P. D. Effects of memory aids on hypothesis behavior and focusing in young children and adults. Journal of Experimental Child Psychology, 1970, 10, 319-336.

Haygood, R. C., Harbert, T. L., & Omlor, J. A. Intradimensional variability and concept identification. Journal of Experimental Psychology, 1970, 83, 216-219.

Levine, M. Hypothesis behavior by humans during discrimination learning. Journal of Experimental Psychology, 1966, 71, 331-338.

Levine, M. Human discrimination learning: The subset sampling assumption. Psychological Bulletin, 1970, 74, 397-404.

Schultz, R. F., & Dodd, D. H. Intradimensional variability in concept identification: A replication, extension, and partial clarification of the Haygood, Harbert, and Omlor findings. Journal of Experimental Psychology, 1972, 94, 321-325.

Suppes, P., & Schlag-Rey, M. Observable changes of hypotheses under positive reinforcement. Science, 1965, 148, 661-662.

NOTE

1. Computation of H changes for both "right" and "wrong" ignores majority rules. Inclusion of majority rules would decrease the proportion of H changes following "wrong."

(Received for publication June 10, 1974.)