STIMULUS GENERALIZATION AS A FUNCTION OF TRAINING IN VERBAL LEARNING, AS MEASURED BY A RECOGNITION TEST

A THESIS

Presented to

the Faculty of Graduate Studies and Research

University of Manitoba

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts



---

by

William Russell Muir

August 1963

ABSTRACT OF THESIS

The evidence for the hypothesis that stimulus generalization rises and then falls in the course of verbal learning was reviewed. As it was concluded that very little support could be mustered for this postulate that was free of major criticism, a "recognition" test which avoided this criticism was devised to test the hypothesis.

This was done by having each of 10 groups of 14 subjects practice by a standard anticipation recall method with a list of eight paired associates for either 1, 2, 3, 4, 5, 6, 8, 10, 12 or 14 trials, respectively. Immediately after each subject finished his pre-assigned amount of practice he was shown, one at a time, the eight correct pairs randomly interspersed with eight mispairings of the correct stimuli and responses. The subject was instructed to respond "Same" if he recognized the pair as correct, "Different" if he recognized it as incorrect, and "Don't Know" if he was uncertain. In each group of 14 subjects, all of the 56 possible mispairings of stimuli and responses were each presented once. A response of "Same" to a mismatched pair was counted as a generalized response.

The results of this experiment indicate that generalization decreases monotonically as a function of practice, contrary to the hypothesis considered earlier. It was also found that the recognition test yielded higher learning scores than did the recall test in the early stages of learning, although this difference disappeared as training continued.

It was concluded that the recognition test used in this study measured the purely associative phase of learning, avoiding the necessity

of prior response acquisition before the subject could respond to the stimulus. It was also tentatively concluded that stimulus generalization rises and falls in the usual verbal learning experiment because a sizeable amount of response acquisition must take place before the subject can respond. Further research was called for to confirm this conclusion by studying the effects of other factors which were uncontrolled in this study.

TABLE OF CONTENTS

CHAPTER		PAGE
I.	INTRODUCTION AND HISTORICAL BACKGROUND	1
	Gibson's Verbal Learning Theory	1
	Review of the Theory	1
	Support for Gibson's Theory	3
	Conditioning experiments	4
	Verbal learning experiments	5
	Criticism of Gibson's Theory	5
	The Recognition Test	15
	Statement of the Problem	18
II.	EXPERIMENTAL METHOD	20
	Material	20
	Design	22
	Subjects	22
	Procedure	23
III.	RESULTS	26
	Comparison of First and Second Recognition Test-trials	26
	Generalized Responses as a Function of Training	36
	Anticipation Recall	42
	Comparison of the Recognition and Anticipation Recall Measures	42
	Within-subjects comparison	44
	Between-subjects comparison	47

CHAPTER PAGE

IV. DISCUSSION 51

V. SUMMARY AND CONCLUSIONS 58

BIBLIOGRAPHY 60

LIST OF TABLES

TABLE		PAGE
I.	Mean Number of "Same", "Different", and "Don't Know" Responses	27
II.	Changes in Mean Difference Scores between Tests 1 & 2	34
III.	Proportions of "Different" and "Same" Responses on Trial 1 to Correct Pairs Containing Stimuli Which Elicited "Same" Responses to Incorrect Pairs	40
IV.	Relation between "Same" and "Don't Know" Responses to Mismatched and Correctly Matched Pairs Containing the Same Stimulus Items	41
V.	Correlation between Number of Responses Correct or Incorrect on Recall Test and Number of Responses Correct or Incorrect on Recognition Test	45
VI.	Correlation between Number of Responses Correct or Incorrect on Recall Test and Number of Responses Correct or Incorrect on Recognition Test, Adjusted for Extra Learning and for Guessing	48

LIST OF FIGURES

FIGURE		PAGE
1.	Mean "Same" Responses to Correct Pairs on Tests 1 & 2	28
2.	Mean "Same" Responses to Incorrect Pairs on Tests 1 & 2	29
3.	Mean "Different" Responses to Correct Pairs on Tests 1 & 2	30
4.	Mean "Different" Responses to Incorrect Pairs on Tests 1 & 2	31
5.	Mean "Don't Know" Responses to Correct Pairs on Tests 1 & 2	32
6.	Mean "Don't Know" Responses to Incorrect Pairs on Tests 1 & 2	33
7.	Pooled Mean "Same" Responses to Incorrect Pairs	38
8.	Mean Number of Overt Errors on Anticipation Recall Test	43
9.	Comparison of Mean Recall, Mean Recognition and Corrected Mean Recognition Scores	50

CHAPTER I

INTRODUCTION AND HISTORICAL BACKGROUND

When an organism has learned to produce a distinctive response to a particular stimulus, similar but different stimuli will also elicit this response, although the organism has had no training with these other stimuli. This phenomenon, known as stimulus generalization, has been shown to occur in a wide variety of learning situations involving both human and animal subjects (Kimble, 1961).

I. GIBSON'S VERBAL LEARNING THEORY

Review of the Theory

Gibson (1940) has proposed a theory which predicts and explains a number of experimental findings in the field of verbal learning by the application of principles derived from the classical conditioning situation. Laboratory verbal learning typically involves the association by a human subject of a specified verbal response to a given verbal stimulus. The items used as stimulus and response are either meaningful words or groups of letters arranged to make meaningless "words" (nonsense syllables). These words may be arranged consecutively, so that each word serves as a response for the word before and a stimulus for the one following (serial learning); or with the words in pairs, with one designated as a stimulus and the other as a response (paired-associate learning). This thesis is concerned with the occurrence of stimulus generalization in paired-associate verbal learning.

Starting with ten basic pastulates which were either "empirically demonstrated in the conditioned response situation" or "incapable of

empirical demonstration, and [which] must rest on confirmation of the relationships which they predict in common" (1940, pp. 205-206), Gibson derived 17 propositions which predicted a number of phenomena in verbal learning. This study is concerned with Gibson's Postulate 4, which states that "Generalization will increase to a maximum or peak during the early stages of practice with a list, after which it will decrease as practice is continued" (1940, p. 206). Generalization is defined as "the tendency for a response Ra learned to [stimulus] Sa to occur when [stimulus] Sb (with which it has not been previously associated) is presented" (1940, p. 204). Gibson does not specify the relationship between stimuli which will result in generalization, but simply assumes that it will be caused by any continuum along which stimuli can be said to be "similar".

Her explanation of this postulate can be paraphrased as follows: Before a subject's first attempt to learn a paired-associate list, the bonds between stimuli and responses should be negligible if he has not experienced the pairings before. After he has been exposed to the pairs for a few trials, the association between each stimulus and its correct response will increase, resulting in a growth of habit strength of the correct responses. However, if the stimuli are similar to each other the subject will tend to confuse them, causing him to associate wrong (or generalized) responses to a given stimulus as well. In this way, the habit strength of both the correct and the incorrect responses will increase in the initial trials of paired-associate learning, resulting in interference with the correct responses and a rise in overt errors.

But as the frequency of the correct responses increases, and with it the frequency of the generalized ones, differentiation between the stimuli will also increase. This is because the correct responses are reinforced when they are produced, increasing their habit strength; while the incorrect ones are not, resulting in their eventual extinction. The net effect, then, will be that generalization will increase in the early stages of paired-associate learning with the growing habit strength of the correct responses. It will reach a peak and then subsequently decrease as differentiation increases.

Regarding the measurement of generalization in paired-associate learning, in her 1940 article Gibson says only that generalization represents the tendency for responses other than the correct one to occur to a given stimulus. However, in a subsequent study (Gibson, 1942), she states that "Although there undoubtedly were frequent tendencies to generalize which did not materialize in overt errors, the number of cases of overt generalization may be taken as a good indicator of the generalizing tendencies" (p. 194), and "By a generalization error is meant a wrong response which is actually the appropriate response to another stimulus item in the list" (p. 191).

Support for Gibson's Theory

As was stated earlier, Gibson supports her postulates with two types of evidence: (1) Principles derived from classical conditioning experiments and (2) experiments using verbal material to test the predictions of the postulates. In the case of Postulate 4 which is being considered here, Gibson relies on both categories.

Conditioning experiments. Pavlov (1927) reported that upon presenting his dogs with a musical tone similar to one that they had been trained to salivate to, a weak response would occur. With subsequent repetitions of the generalized stimulus and further training with the conditioned stimulus, the generalized response would increase until it equalled the conditioned response in magnitude, and then decrease to zero. Razran (1949) examined a number of Pavlovian experiments, equated for procedure, and analyzed the pooled results statistically (which apparently Pavlov had never done). He found that the magnitude of the generalized response, expressed as a percentage of the conditioned response strength, reliably increased and then decreased.

Beritoff (1924) reports that while conditioning dogs to give a motor response to a given tone, he found that they would initially respond to increasingly dissimilar tones, but eventually this generalization would stop spreading or decrease. However, his description presents no evidence of controlled experimental design, statistical analysis or even consistent effects.

Hovland (1937), using human subjects, conditioned a GSR to a tone of a given intensity, with four groups receiving 8, 16, 24 and 48 reinforcements respectively. When a tone of different intensity was presented to each group it was found that relative generalization increased to a peak at 16 reinforcements, and then dropped and levelled off.

Verbal learning experiments. Gibson (1942) designed an experiment to test her postulate by having subjects learn nonsense syllable responses paired with "nonsense forms" (geometric forms having no conventional names) as stimuli. When overt errors from three lists were pooled and Vincent curves drawn, it was found that generalization increased sharply for the first three tenths of the learning curve, remained constant for another tenth, and then gradually decreased. The changes, although regular, were quite small, with an increase of approximately 1.1 errors and a decrease of approximately 2.0 errors being reported out of a possible 36 responses. No statistical tests were reported.

Gagné (1950), in a similar study not mentioned by Gibson, varied the order of the pairs among several groups of subjects in an attempt to minimize and maximize generalization. All groups showed an increase in overt errors followed by a decrease. No test of the reliability of the increase was reported, but from the published summary of the data it would appear to be significant.

Criticism of Gibson's Theory

Although an initial rise and subsequent fall in the frequency of overt errors in the learning of a paired-associate list has been generally accepted as a widely occurring empirical phenomenon (Murdock, 1958), Gibson's explanation and the experimental method that she used to test this explanation have been criticized on a number of points:

(1) Underwood (1961) has questioned the validity of the support that Gibson's 1942 experiment (and, indirectly, Gagné's 1950 study) gives to the theory, because they used forms, not words,

as stimuli. "The theory derives from principles of conditioning, the experiments used forms as stimuli, and yet the results of her experiments are said to be directly relevant to a theory of verbal learning. This series of extrapolations has always been a puzzle to me, although, of course, there is no puzzle if the theory actually 'works' for verbal material" (p. 200). However, there seems to be no serious doubt that the frequency of overt errors usually rises and then falls in the course of learning a paired-associate list by standard anticipation recall methods (Murdock, 1958). It would appear, then, that in this respect the theory "works" for verbal stimuli as well as for forms, resolving Underwood's puzzle.

(2) Murdock (1958) points out that Gibson (1940) does not distinguish between absolute and relative generalization in her theory. Absolute generalization refers to the intensity or frequency of the generalized response, while relative generalization is the ratio of generalized to conditioned or "correct" responses. This distinction is important, as the two measures seldom co-vary. According to Kimble (1961), "Functions of this form [generalized responses increasing to a peak and then decreasing in conditioning studies], when they have been reported, have always been for relative generalization, never for absolute generalization." (p. 335) Hovland's (1937) experiment is a case of this. The absolute generalization curve continues to increase slightly between 16 and 24 trials, and between 24 and 48 trials, while relative generalization drops sharply after 16 trials and continues to decrease slightly after 24 trials. This simply represents the fact that, in the initial part of conditioning, the

generalized response increases more rapidly than does the conditioned one, while the reverse occurs in the latter stages.

However, in Gibson's (1942) experiment, although absolute generalization (as overt errors) increases at first and then drops, relative generalization continually decreases in a negatively accelerated curve, because the number of correct responses always increases more rapidly than the number of incorrect ones. In Gagné's (1950) study relative generalization shows an initial negligible increase (from a ratio of 3.11 to 3.27 over two trials, i.e., approximately 5%) and then a sharp decrease, again negatively accelerated. So it can be seen that because of the confusion between absolute and relative generalization, Gibson's postulate is only vaguely analogous to the conditioning situation exemplified by Hovland (1937).

(3) Hovland's (1937) experiment differs from the usual verbal learning situation on still another point, making comparison even more difficult. This concerns the amount of training with the generalized stimulus that the subject receives prior to testing.

In Hovland's (1937) procedure the subject was tested with the generalized stimulus after a pre-arranged number of conditioning trials with the conditioned stimulus alone. So his generalization curve represents the subjects' initial reaction to a generalized stimulus which has not been presented previously, after completion of a specified amount of training with the conditioned stimulus alone. Each experimental group, then, had the same amount of experience with the generalized stimulus (i.e., on the generalization test only) regardless of the varying amounts of reinforcement that they received. (This is also true of the Pavlovian experiments summarized by Razran [1949]).

However, the subject in the typical verbal learning experiment is, in effect, exposed to both the conditioned and generalized stimuli all through training. According to Gibson (1940), if two stimuli from a paired-associate list elicit the same response, one correctly and the other incorrectly, then they are defined as the "right" (or conditioned) stimulus and the generalized stimulus, respectively. But the subject receives the same amount of practice with each one, for both stimuli are presented once on each trial from the start of training. The only difference in the treatment that they are given is that the conditioned response is reinforced when it occurs, while the generalized one is not. Because the generalized stimulus is presented on each trial, it follows that the various experimental groups will differ in the amount of experience that they have with it, because they are tested at different stages of learning.

It is obvious, then, that the two learning paradigms are hardly comparable. In the conditioning procedure used by Hovland (1937) the subjects had no experience with the generalized stimulus before conditioned response strength had been brought to an appreciable level. In the typical verbal learning experiment, however, the subjects are trained with the generalized and conditioned stimuli concurrently from the beginning, which almost certainly would result in an interaction in the development of conditioned and generalized response strengths.

This is an important point, for the growth of discrimination through the elicitation and subsequent nonreinforcement of generalized

responses during learning is a crucial process in Gibson's (1940) theory. For a valid comparison of the paired-associate situation with conditioning procedure it would be necessary to find an experiment where subjects were presented with both the conditioned and generalized stimuli from the outset of training, and where results are reported throughout in terms of absolute generalization. Unfortunately, few studies could be found that met these conditions, and the results reported offer only indifferent support to Gibson's hypothesis.

An instrumental conditioning study supporting Gibson's hypothesis was done by Raben (1949), who trained rats to run along runways of different brightnesses for a food reward. All animals were given eight trials of training with the brightest runway, and then assigned to one of five groups. Subsequent to this, each animal ran unrewarded for three trials out of every four on a runway that was darker (generalized stimulus) than the one that it was trained on (conditioned stimulus), while continuing to run every fourth trial on the original runway for reward. The ratio of the brightnesses of the positive and negative runways varied among the groups, for a control group there being no difference.

Although the response strength (measured by running speed) to the positive stimulus was at a fairly substantial level when discrimination training began, it continued to increase for all groups. The group with the darkest runway differentiated quickly, and generalized responses, which had started at the same level as the conditioned responses, dropped almost immediately. As the brightness became more similar the response strength to the generalized stimulus increased more before dropping, with no differentiation at all being shown to the identical stimuli.

Hilgard, Campbell & Sears (1937) used an eye-blink conditioning procedure to study the development of discrimination between two stimuli, small windows in which lights appeared. Among 14 subjects tested over three days, they found that three groups became roughly differentiated: (i) Immediate discriminators, with whom differential responding was established almost from the beginning of training (N = 5); (ii) Gradual discriminators, who did not discriminate until the third day, when responses to the negative stimulus decreased after an initial increase (N = 4); (iii) Nondiscriminators, for whom the level of response was the same to both stimuli for all three days (N = 5). When the results reported for these three groups are pooled and a composite curve is drawn, the percent frequency of response to the negative stimulus increases on the second day and drops slightly on the third, as the Gibson theory (1940) predicts.

However, the interstimulus interval (between the positive conditioned and unconditioned stimuli), although held constant over the three days for any given subject, was not the same for all subjects, and it was found that a longer interval favoured the rapid development of discrimination. This would seem to indicate that the rise and subsequent fall in generalization was caused largely by the subjects in the intermediate interval range, while those that received longer intervals discriminated quickly and those with shorter intervals failed to discriminate at all.

Two recent studies (Hartman & Grant, 1962; Kimmel & Pennypacker, 1963) have demonstrated that variation of the CS-UCS interval does have significant effects on the course of differential conditioning. Even more important it was found that controlling the CS-UCS interval eliminated almost all of the increase and subsequent decrease in generalized response.

Hartman & Grant (1962), using an eyelid conditioning procedure and CS-UCS intervals of 400, 600, 800 and 1000 msec. found no instance in which generalization increased and then decreased. However, at the end of the 44 trials of training reported for the negative stimulus, generalization in the 400 and 600 msec. groups appeared to be still increasing, so it is possible that a decrease in generalization might have occurred had training been continued. Kimmel & Pennypacker (1963), in a similar study using a GSR conditioning procedure and CS-UCS intervals of 0.25, 0.50, 1.00 and 2.00 sec. found a clear-cut rise and fall in generalization only for the shortest interval. Again, however, training with the negative stimulus continued only for a short time (18 trials), so that it is possible that generalized response strength might have eventually dropped if training had been continued.

In conclusion, a review of the few differential conditioning studies where training with positive and negative stimuli commenced simultaneously and absolute generalized response strength is reported throughout training revealed very little support for Gibson's (1940) hypothesis. Raben (1949) showed that generalized responses by rats to a brightness dimension will increase and then decrease with practice in an instrumental conditioning situation; however, her subjects started discrimination training with some prior learning with the positive stimulus alone. Hilgard, Campbell & Sears (1937) found that generalized eye-blink responses first increase and then decrease with training, but variations in CS-UCS interval between subjects made it impossible to decide whether the changes in generalization were due to practice

alone. Hartman & Grant (1962) and Kimmel & Pennypacker (1963) showed that, with the CS-UCS interval controlled over a range of 0.25 to 2.00 sec., generalization increased and decreased only in one instance, when the interval was 0.25 sec. However, in both these studies only a limited amount of training was given the subjects, and it is possible that Gibson's postulate might have been confirmed had conditioning been continued.

(4) Gibson (1942) has been criticized for using frequency of overt errors as a measure of generalization. Murdock (1958) points out that recent studies of the "response-acquisition" phase of learning (Hovland & Kurtz, 1952; Underwood & Schulz, 1960) have cast doubt on this matter. These studies suggest that paired-associate learning can theoretically be divided into two stages: response-acquisition and associative learning. Before a subject can learn to produce a number of specific responses to appropriate stimuli, it is obvious that these responses must be in his repetoire and that they must be differentiated from other responses not associated with this particular set of stimuli. Logically, then, in a paired-associate task a subject must acquire (when the responses are unfamiliar) or isolate (when familiar responses are used) the proper set of responses before he can learn to associate them with their stimuli. In practice the two stages can and probably do overlap, but the evidence (Underwood & Schulz, 1960) indicates that prior familiarization training with the responses to be learned in a paired-associate task facilitates the learning of the pairs. Thus it seems reasonable to conclude that in the initial stages of paired-associate learning the subject is occupied more with the acquisition of the responses than with associating them with their respective stimuli.

As Murdock (1958) points out, this initial response-acquisition affects Gibson's (1942) frequency of overt errors measure of generalization. At the start of the typical paired-associate learning experiment, because it is likely that the subject does not know the specific set of responses that he must give, he produces a large proportion of irrelevant responses from previous learning (intrusions) or fails to respond. Murdock (1958) considers that the true explanation of why errors rise and then fall in verbal learning is that these intrusions and failures to respond interfere with generalization in the early stages of list acquisition, and then gradually decrease as the proper responses are learned. If this interference were controlled, he feels, it could be shown that generalization is maximized at the start of learning, not in the middle. To test this argument, he has proposed to have subjects learn a paired-associate list by the usual anticipation recall method, except that they would be given a list of the responses and forced to respond to all stimuli, even if they had to guess. In this way the confounding effects of omissions and intrusions early in learning could be eliminated.

Gibson (1959) points out that Murdock (1958), in avoiding the problem of response learning by his method of testing generalization, has created another problem. By allowing only two classifications of response -- correct and incorrect (or generalized), he is in effect saying that a response that has not been learned correctly must have been learned incorrectly, i.e., that the inability to produce a correct response is due to generalization alone. Thus "generalization" is equated with "absence of learning", which is not at all how Gibson (1940) defined it.

Runquist (1959) has also criticized the use of frequency of overt errors as a generalization measure, both in the Gibson (1942) and the Murdock (1958) methods of measurement. He contends that generalization is an interference phenomenon, and as such is not measured accurately by incorrect responses, which can be considered as intrusions of interfering learning. Citing evidence from retroactive inhibition and associative interference studies (Underwood, 1945; Spiker & Holton, 1958) which show that interfering tendencies are usually indicated by omissions, not intrusions, he concludes that neither the Gibson nor the Murdock methods can adequately differentiate between an omission caused by absence of learning and one caused by generalization.

Summarizing the general conclusions that we can draw from these criticisms of Gibson's theory, we find that:

(1) It is a well-established empirical fact that, in the learning of a list of paired-associates, the frequency of overt errors rises to a maximum after the initial stages of practice and then decreases. This is consistent with Gibson's predictions.

(2) In comparing verbal learning experiments with conditioning studies, Gibson fails to distinguish between <u>absolute</u> stimulus generalization, which is measured by the magnitude of the response to the generalized stimulus, and <u>relative</u> stimulus generalization, which is the ratio of the strength of the generalized responses to the strength of the conditioned responses. When it is realized that these two measures do not necessarily co-vary, it is found that her postulate is not supported by the conditioning studies from which she derived it.

(3) Moreover, it was found that the experimental paradigm of these conditioning studies differed from Gibson's in that the subject had no experience with the generalized stimulus during conditioning. The only

relevant experiments which could be found that satisfied this condition supported her hypothesis only in very limited circumstances.
(4) It was concluded that frequency of overt errors is not a satisfactory measure of generalization, either by the method proposed by Gibson (1942) or by Murdock (1958).

In brief, it would seem that there is almost no relevant evidence to confirm Gibson's generalization hypothesis that is free from major criticism, either from the conditioning analogy or from empirical tests with verbal learning. This thesis attempts to test Gibson's postulate in the verbal learning situation, using a method to measure generalization that will try to avoid the criticisms mentioned in the previous discussion.

The Recognition Test

In a recent re-examination of her theory, Gibson (1959, p. 340) suggests that "an identification experiment, with only the responses 'same' or 'different', seems ideal" for studying generalization, although she doubts that further study of generalization in paired-associate learning would be fruitful. However, it seems possible to combine this suggested method of Gibson's with the paired-associate situation in a modification of a test that is currently being used in verbal learning studies at the University of Manitoba as a measure of degree of learning.

This test consists of presenting the subject with a number of stimulus-response pairs, one at a time, the pairs being either the same as ones from a list that the subject is attempting to learn, or mismatches of the stimuli and responses from that list. The subject is told that

some of the pairs will be the same as the ones that he has been trying to learn, while others, although containing the same words, have been mispaired. He is instructed to respond "Same" if he recognizes the pair, "Different" if he thinks that it is a mismatch, and "Don't Know" if he is uncertain.

This measure is referred to as a recognition test throughout this thesis, as it is felt that this is the most descriptive title that could be applied. However, this should not be confused with a wide variety of other types of tests that have been given the same name (e.g., Luh, 1922; Postman, Jenkins & Postman, 1948; Postman & Rau, 1957; Murdock, 1963), but which could be better described as "selection" or "identification" tests. No published reports mentioning the recognition test used in this study as a testing method in paired-associate learning have been found, but the following summary shows why it is felt that this test would be suitable for testing Gibson's postulate.

(1) On a priori grounds it would appear that this recognition test should be free of the effects of the response-acquisition phase of learning. The subject need not have the response available or even have to select the response that he gives. All he must do is say whether he considers that an associative connection is correct or not.

The fact that the proposed test measures associative learning only, and not response learning is probably the most important reason for using this recognition test as a measure of generalization. It should be pointed out that this feature distinguishes the recognition test used here from the identification type of recognition test mentioned earlier, in that these identification tests measure response learning only, and not the associative strength between the members of the pair.

The recall test measures response learning and associative strength. It can be seen, then, that this property of the proposed recognition test gives it an advantage over other methods of measuring generalization.

(2) Battig (1959) has pointed out that a method which provides independent measures of generalized and correct response strength, as is done in the conditioning paradigm (where correct response strength is measured by the frequency of response to the conditioned stimulus, while generalization is estimated from frequency of response to the generalized stimulus), is needed to settle the question of Gibson's hypothesis. Gibson's (1942) and Murdock's (1958) procedures did not do this, as both correct and generalized response strengths were estimated from the same set of stimuli. The proposed test satisfies Battig's condition, as generalized response strength is measured by the number of "Same" responses to one set of stimuli (the mismatched pairs), while correct response strength is measured by the number of "Same" responses which are elicited by the separate set of correctly matched pairs.

(3) From unpublished studies carried on at this laboratory there is reason to believe that the recognition test measures the same learning process as does recall. There is some indication, however, that recognition can detect weaker associative bonds than recall.

(4) It will be recalled that both Gibson's (1942) and Murdock's (1958) methods of measuring generalization were criticized because neither could adequately distinguish between absence of learning and generalization. This was because both of these factors would tend to produce either omissions in Gibson's method or overt errors in Murdock's. The proposed recognition test, however, should be able

to distinguish between generalization and absence of learning through the provision of the "Don't Know" response category. This third response category would provide a measure of absence of learning, so that this would not be confounded with generalization.

(5) Because the subject can be tested with both correctly and incorrectly matched pairs containing the same responses or stimuli, the consistency of his responding can be measured to obtain a more complete view of the response tendencies operating.

To conclude, it can be seen that the recognition test proposed in this thesis would seem to avoid many of the criticisms that have been leveled at other methods of measuring generalization. The most important improvement is the elimination of the necessity of response learning. As Murdock (1958) has pointed out, response learning could very well account for Gibson's (1942) finding that overt errors rise before they decrease in a recall test of verbal learning. The present test also avoids a disadvantage of Murdock's (1958) method of controlling response acquisition. Because Murdock allows only two categories of response (correct and incorrect), responses that have not yet been learned are indistinguishable from generalized responses. The proposed recognition test, however, can detect unlearned responses. In addition, this recognition test would seem to be able to measure generalization more sensitively than other methods, to be able to provide a measure of generalization independent of the measure of correct response strength, and to provide additional information about the generalized responses.

II. STATEMENT OF THE PROBLEM

It has been shown that Gibson's hypothesis that generalized response tendencies in paired-associate learning increase and then decrease as a function of practice has not been adequately tested empirically, because no satisfactory measure of generalization has been available. This study will attempt to determine the course of generalized errors as training with a paired-associate list progresses, by means of a recognition test. Subjects, after being given a predetermined number of learning trials on a list designed to produce a large amount of stimulus generalization will be presented with a number of stimulus-response pairs, made up of correctly or incorrectly matched items from the list that they are attempting to learn. They will be instructed to respond "Same" if they recognize the pair, "Different" if they feel that the pair is mismatched, and "Don't Know" if they are uncertain. By comparing the relative frequencies of misrecognitions of the incorrectly matched pairs for subjects who had different numbers of learning trials previous to testing, changes in generalization can be measured.

CHAPTER II

EXPERIMENTAL METHOD

I. MATERIAL

A list of eight paired-associates was constructed, consisting of consonant-vowel-consonant nonsense syllables as stimuli and frequently-used two-syllable adjectives as responses. The adjectives were selected by several judges so as to be as dissimilar to each other as possible, both semantically and structurally. The nonsense syllables were taken from Appendix A of Underwood & Schulz (1960), and were selected so as to have relatively low meaningfulness values (between 33-40% on the Glaze scale and between 60-72% on the Krueger scale). All of the nonsense syllables contained the vowel "O", but no consonant was used more than once in the eight syllables. It was assumed that the low meaningfulness and common vowel would make the initial discriminability of the stimuli low, resulting in conditions conducive to high stimulus generalization. Conversely, the high discreteness and meaningfulness of the responses was designed to minimize response generalization and the necessity of extensive response learning.

Four different orders of the eight pairs were arranged for anticipation learning. Except that the last two pairs in an order never occurred among the first two pairs in the next order, the sequence of the pairs was randomly varied to prevent serial learning.

To construct the recognition tests, an eight by eight matrix of all the possible combinations of the stimuli and responses was drawn up, comprising the eight correct pairings and 56 incorrect pairings. From this matrix 14 recognition tests were constructed. A Latin square (Fisher

& Yates, 1953) was used to ensure that the following conditions applied to these tests:

(1) Each test contained four correctly paired and four mismatched stimuli and responses.

(2) Each stimulus and each response was used once and only once in each test.

(3) Each of the possible 56 mismatchings of stimuli and responses occurred exactly once among the 14 tests.

(4) Over the 14 tests, each stimulus and each response appeared exactly seven times as an item in a mismatch; and each appeared exactly seven times in a correct pairing (i.e., each correct pair was repeated seven times).

(5) The 14 tests could be arranged into seven complementary pairs of tests, in that each pair of tests contained all eight correct pairs without repetition, and also all eight stimuli and eight responses, without repetition, arranged in eight different mismatched pairs.

The four orders of the anticipation learning lists and the 14 recognition tests were arranged on a Lafayette multilist variable speed memory drum so that the exposed list could be rapidly changed by opening and closing appropriate shutters.

A practice list of four one-digit-number stimuli paired with four two-digit-number responses was made up, with two different orders of an anticipation learning list, and a recognition test composed of two correct pairs and two mismatchings. These were used as examples to illustrate the instructions given to the subjects for the two types of test.

II. DESIGN

Ten groups of 14 subjects each attempted to learn the eight word-pairs by the anticipation-recall method for either 1, 2, 3, 4, 5, 6, 8, 10, 12 or 14 trials. After completing the number of learning trials assigned to his group each subject was given one of the 14 recognition tests, followed by a second recognition test using the test complementary to the first one. All 14 tests were used once as a first test and once as a second test for each group. Also, all subjects were tested on all eight correct pairings of the items (four in the first test, four in the second), and on eight of the 56 possible mismatches.

The experimental conditions for each subject (number of learning trials, initially-presented order of learning list and recognition tests used) were determined in advance and recorded on a data sheet. The 140 sheets were then shuffled and the subjects were assigned to the different conditions according to the sheet that was on top of the pile when they reported for testing.

III. SUBJECTS

The subjects were 140 University of Manitoba students (57 males and 83 females) enrolled in the introductory psychology course, who took part in the experiment to fulfill a laboratory requirement. All subjects had previously served in another verbal learning experiment where they had learned and recalled from two to four lists of paired adjectives. Experienced subjects were used to reduce the variability of performance due to "learning to learn" behaviour.

IV. PROCEDURE

The subject was seated in front of the apparatus, and the following instructions were read to him:

"This experiment is similar to one that you did earlier, in that you are going to learn pairs of words. However, as there are some important differences between this experiment and the earlier one, and as you may have forgotten some of the procedure, I'll give you complete instructions, as though you've never done this before.

"I want you to learn a list of eight word-pairs that I'm going to show you. Each pair consists of a short nonsense word and a meaningful word. The nonsense words are simply three letters combined to make words that mean nothing, and they serve as cues for learning the meaningful words. In one of these slots I'm going to show you each nonsense cue-word for two seconds, followed by the pair of words for two seconds. I want you to learn which meaningful word goes with which cue-word. After I've shown you all eight pairs once, when you see a cue-word alone in the slot, try to call out the meaningful word that goes with it, before the pair appears. You should try to learn as many pairs on each trial as you can, with the objective of learning them all. After you've reached a perfect trial, I may have you continue for a short time, so don't stop until I say so.

"At some point during this experiment, I'm going to have you look at another part of this slot where I'll show you another list of eight word pairs. All of the <u>words</u> in this second list will be the same as in the list you're trying to learn, but some of them may be <u>paired</u> differently; that is some new pairs may have been made with the same

words. I'll show you each one of these pairs for four seconds, and I want you to call out 'same' if you recognize it as a pair from the previous list and 'different' if you think that it has been changed. If you don't know whether it is the same or different, say 'don't know'. Don't make wild guesses. You don't have to be absolutely sure of recognizing a pair as the same or different to answer that way. But you should be reasonably sure. If you have no idea whatsoever if the words have been switched or not, say 'don't know'. And remember -- any number of pairs may have been switched, ranging from all of them to none, so judge each pair separately.

"Do you have any questions?"

(After any questions had been answered).

"Fine, I'll give you some practice now with pairs of numbers, instead of words."

(Here the subject was shown the first order of the practice anticipation list).

"Now you should try to anticipate the second numbers before they appear."

(Here the second order of the practice anticipation list was shown).

"Now I'll show you a list with some of the pairs switched. Say 'same' if you recognize the pair as one of the ones that was on the first list; 'different' if you think that the numbers have been switched; and 'don't know' if you can't tell whether it is the same or different."

(Here the subject was given the practice recognition test. If he did not respond to the first pair, he was reminded. At the finish he was asked):

"Fine. Do you have any questions before you start learning the word-pairs?"

(After questions, if any, had been answered).

"O.K. Now we'll start with the experiment."

The subject was then given the number of anticipation learning trials that he had been assigned on his data sheet, starting with the order specified. Each stimulus word was shown for two seconds, followed by the pair of words for two seconds. All correct responses, mismatched responses, intrusions and omissions were recorded. A six-second interval was interposed between trials.

After the subject had completed the specified number of learning trials he was told,

"Now I'm going to give you the recognition test. Call out 'same', 'different', or 'don't know' to each one of the pairs that you see."

The subject was then shown the first recognition test assigned to him on the data sheet by closing the anticipation learning list shutter and opening one of the recognition test shutters. Each pair was shown for four seconds, and all responses recorded.

Following this the subject was told, "Fine. Now I'll give you another recognition test. Again I want you to call out 'same', 'different', or 'don't know', just like you did last time."

The subject was then shown the second recognition test, the same procedure being followed here as for the first recognition test.

CHAPTER III

RESULTS

I. COMPARISON OF FIRST AND SECOND RECOGNITION TEST-TRIALS

Table I shows a summary of the mean number of responses to the first and second recognition tests in each of the three response categories, "Same", "Different" and "Don't Know". The responses to the correctly matched and incorrectly matched pairs are shown separately for each of the 10 learning trials on which recognition tests were given.

The first step taken in the analysis of the data was to determine whether the results were different for the two recognition tests. If the results on the two tests were found to be the same they could be pooled to form more stable measures. Comparisons of the results of Tests 1 and 2 for each of the six combinations of response category and type of pair are shown graphically in Figs. 1-6.

A preliminary examination of the data showed that there were marked differences in variability of response between the different experimental groups. This was largely because the majority of subjects tested on the later trials were at or near the minimum or maximum limits of the different response combinations. For this reason an analysis of variance across trial groups could not be used. Instead, the differences between the scores on the two tests over all 10 test-trials were computed for each subject in each of the six combinations of response category and type of pair. Then t tests for correlated samples were used to

TABLE I

MEAN NUMBER OF "SAME", "DIFFERENT", AND "DON'T KNOW" RESPONSES

		Trial	1	2	3	4	5	6	8	10	12	14
Response	Type of Pair	Test No.										
"Same"	Correct	1	2.57	2.21	3.07	3.07	3.07	3.57	3.29	3.57	3.64	3.86
		2	2.21	2.21	2.43	3.07	3.00	2.71	3.21	3.29	3.50	3.64
	Incorrect	1	0.64	0.64	0.50	0.14	0.43	0.29	0.43	0.36	0.00	0.00
		2	0.93	0.43	0.86	0.21	0.36	0.00	0.43	0.21	0.14	0.00
"Different"	Correct	1	0.43	0.86	0.43	0.43	0.14	0.29	0.29	0.36	0.21	0.07
		2	0.86	0.50	0.79	0.36	0.50	0.64	0.50	0.43	0.14	0.36
	Incorrect	1	2.64	2.29	2.86	3.57	3.21	3.36	3.21	3.07	3.71	3.93
		2	2.21	2.79	2.43	3.50	3.07	3.86	3.29	3.50	3.71	4.00
"Don't Know"	Correct	1	1.00	0.93	0.50	0.50	0.79	0.14	0.43	0.07	0.14	0.07
		2	0.93	1.29	0.79	0.57	0.50	0.64	0.29	0.29	0.36	0.00
	Incorrect	1	0.71	1.07	0.64	0.29	0.36	0.36	0.36	0.57	0.29	0.07
		2	0.86	0.79	0.71	0.29	0.57	0.14	0.29	0.29	0.14	0.00

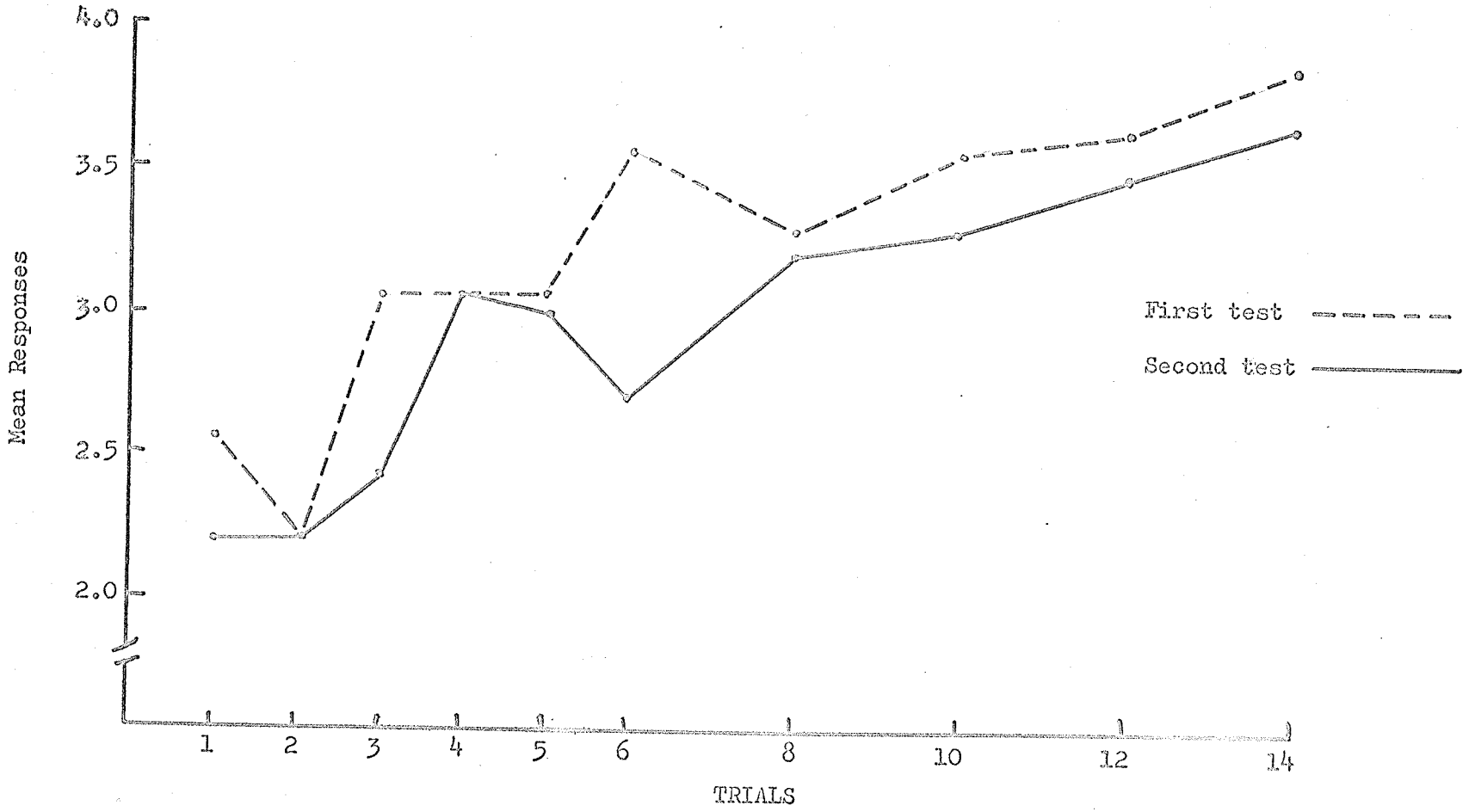


Figure 1. Mean "Same" responses to correct pairs on Tests 1 & 2.

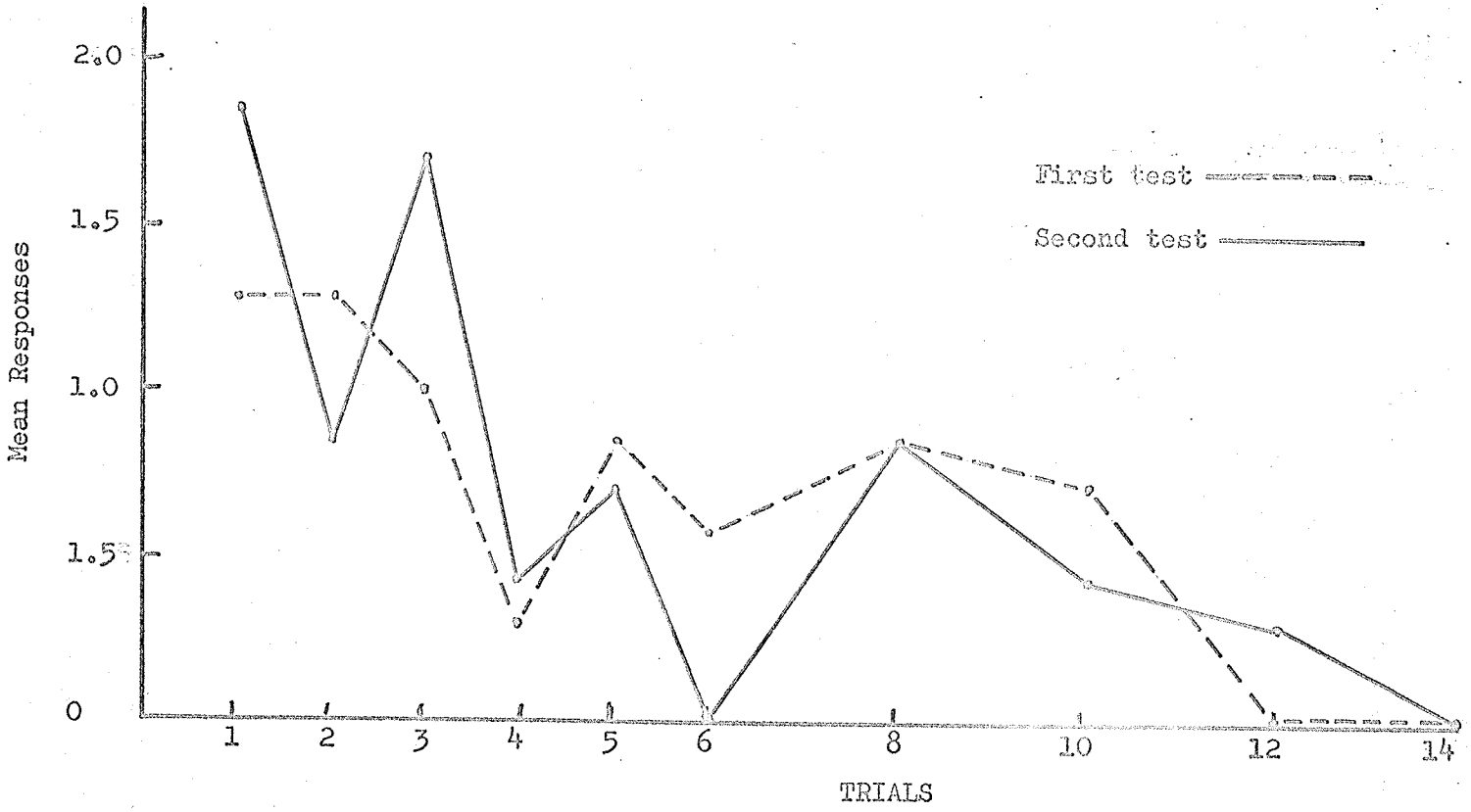


Figure 2. Mean "Same" responses to incorrect pairs on Tests 1 & 2.

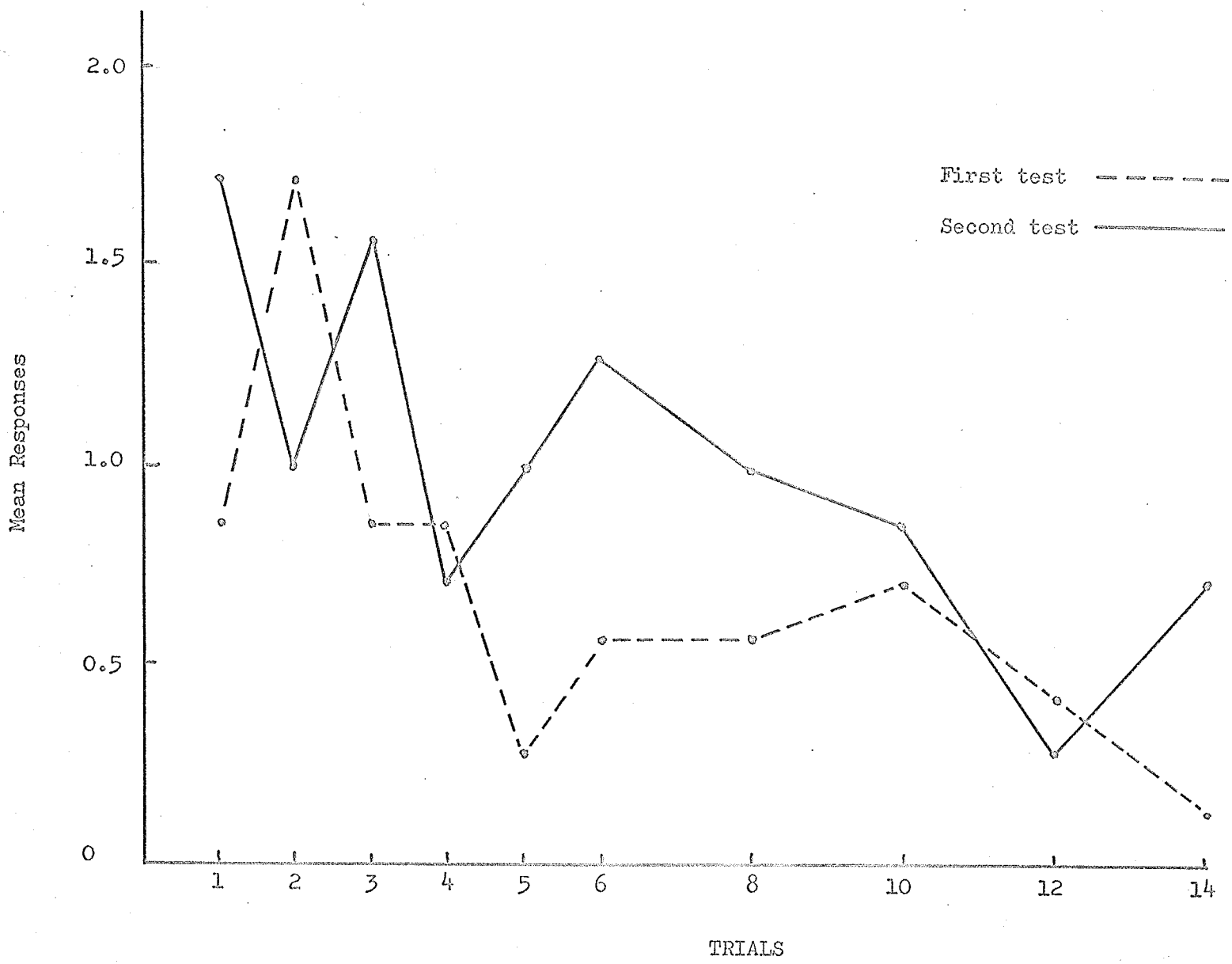


Figure 3. Mean "Different" responses to correct pairs on Tests 1 & 2.

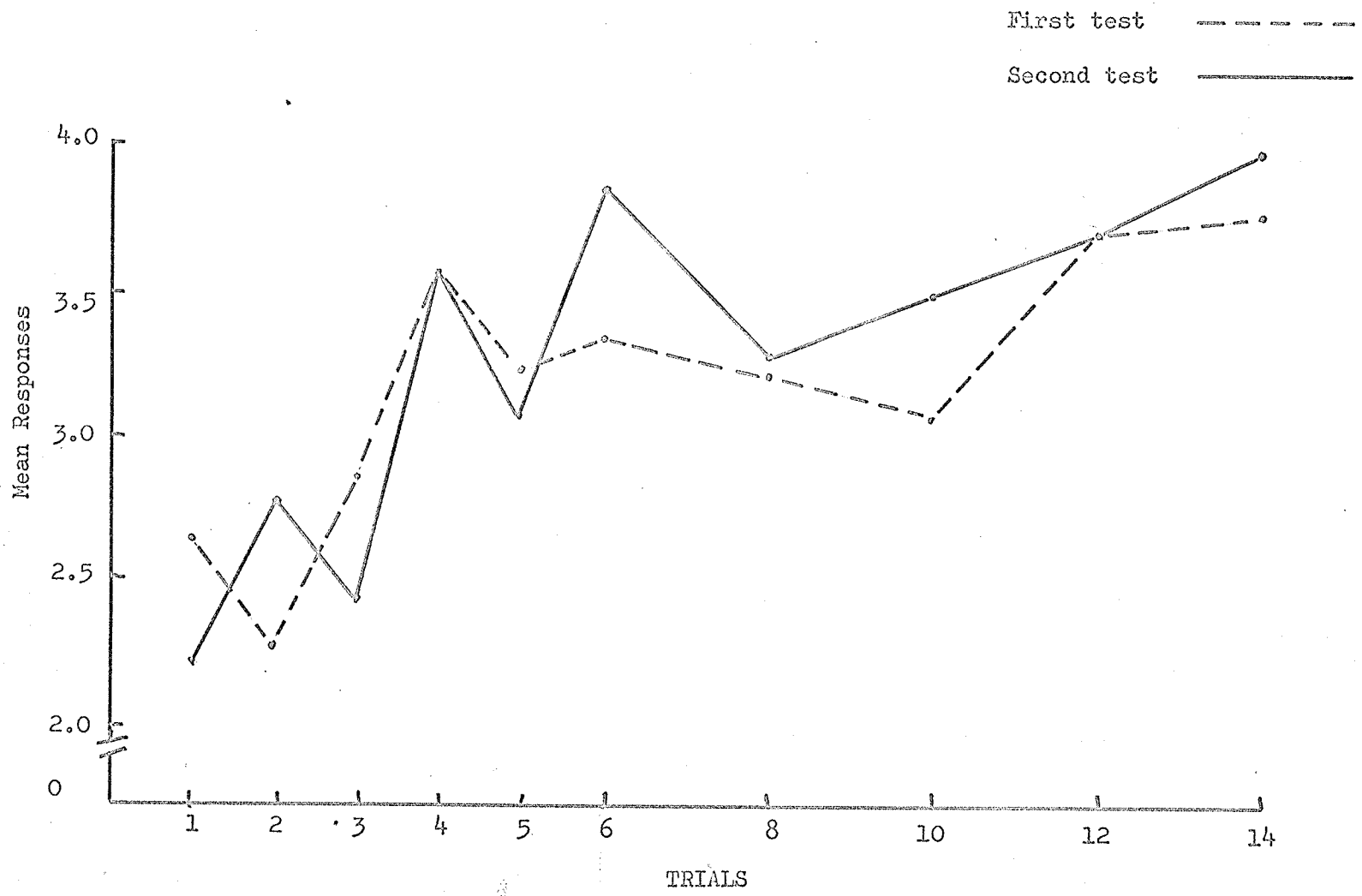


Figure 4. Mean "Different" responses to incorrect pairs on Tests 1 & 2.

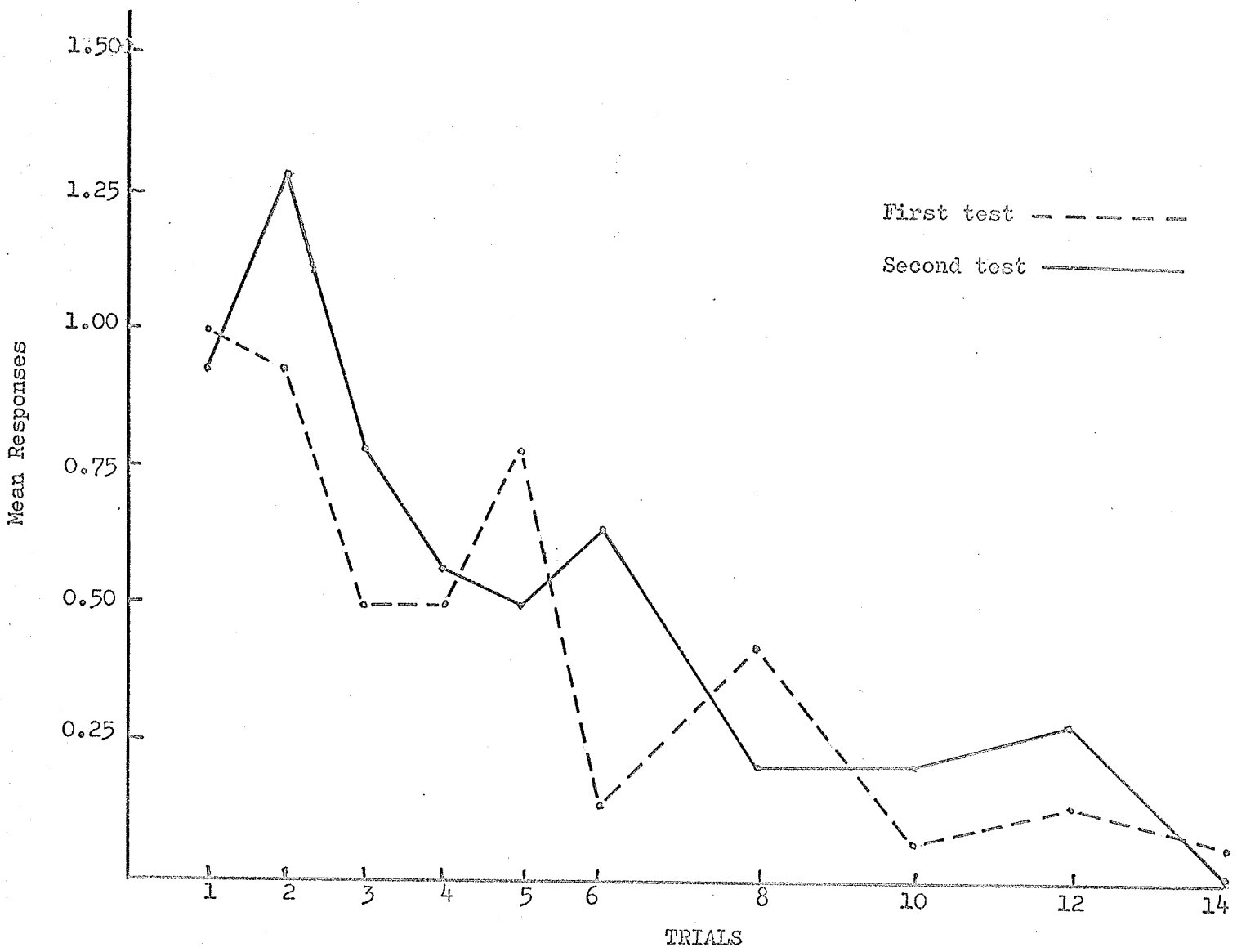


Figure 5. Mean "Don't Know" responses to correct pairs on Tests 1 & 2.

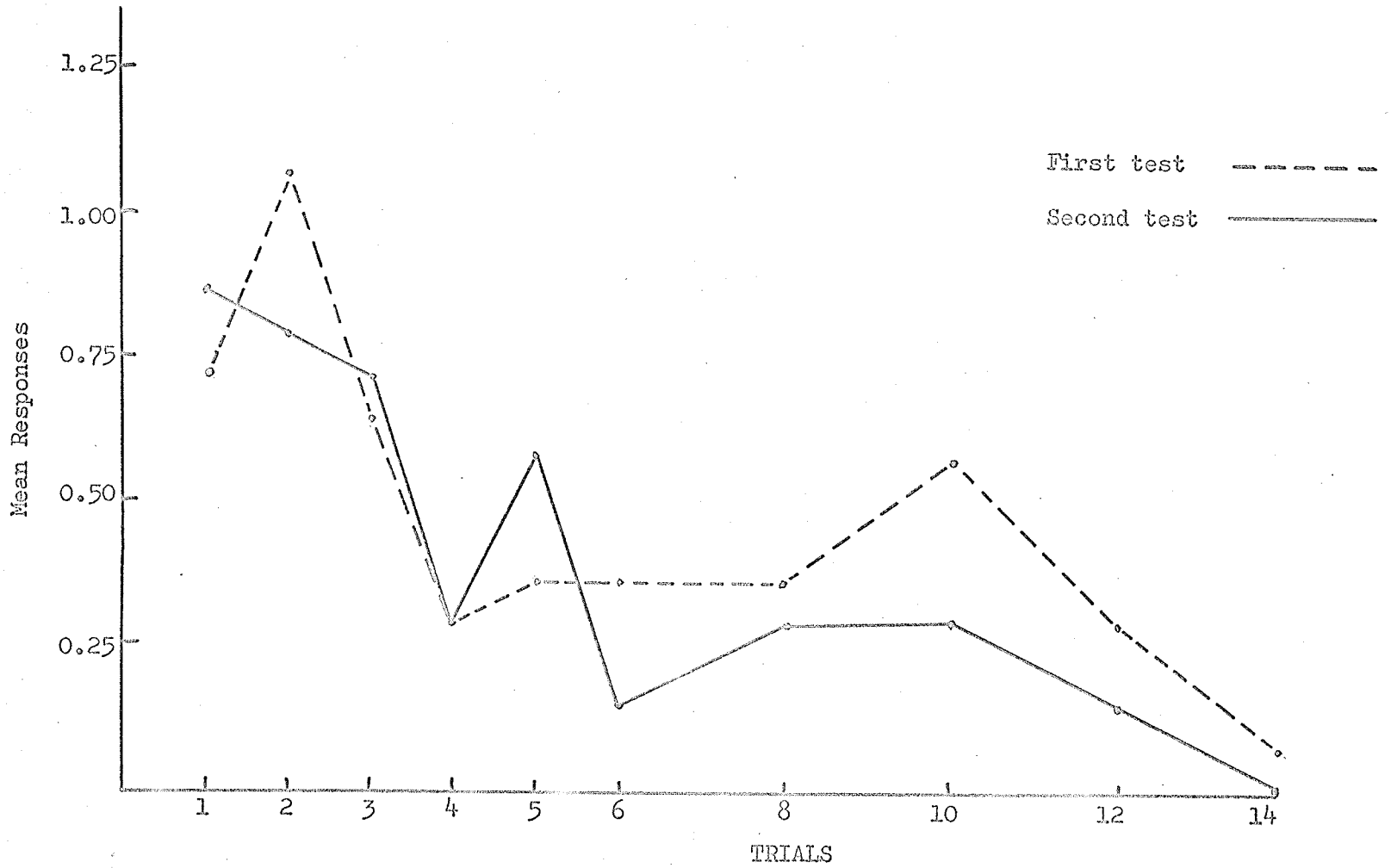


Figure 6. Mean "Don't Know" responses to incorrect pairs on Tests 1 & 2.

determine if the mean difference were significant (Table II, Column A).[1] As an examination of Figs. 1-6 suggested that in some cases there might be an interaction between tests and trials, the data were also tested to determine whether any differences occurred early or late in learning that had not been detected by the test over all 10 trials. To do this the mean difference scores over the first five and the last five test-trials were tested separately for significance with t tests for correlated samples (Table II, Columns B and C).

TABLE II

CHANGES IN MEAN DIFFERENCE SCORES BETWEEN TESTS 1 & 2

		A	B	C	D	E
Response	Type of Pair	Overall Change	Change on Trials 1-5	Change on Trials 6-14	Difference between Trials 1-5 and 6-14	F Ratio of Variances of Trials 1-5 and 6-14
"Same"	Correct	-0.264**	-0.214	-0.314**	-0.100	1.68*
	Incorrect	+0.014	+0.086	-0.058	-0.134	3.57**
"Different"	Correct	+0.157*	+0.143	+0.171*	+0.027	1.33
	Incorrect	+0.050	-0.114	+0.214*	+0.328	3.16**
"Don't Know"	Correct	+0.107	+0.071	+0.143	+0.071	1.79**
	Incorrect	-0.064	+0.029	-0.157**	-0.186	3.61**

*$p < .05$ **$p < .01$

These differences and their associated probability values are presented in Table II. A positive score indicates an increase in response from Test 1 to Test 2, while a negative one shows that a decrease occurred.

[1] Because this comparison was between the paired scores of each subject, rather than between the performances of the different groups, the objection to parametric tests because of heterogeneity of variance between groups given earlier does not apply here.

It can be seen that, on the last five test-trials, the frequency of "Different" responses to correct pairs increased significantly from the first to the second test while "Same" responses to correct pairs decreased in the same period. For the incorrect pairs, "Different" responses increased on the second test during the last five test-trials, while "Don't Know" responses dropped. None of the difference scores for the initial five trials proved significant, although the overall difference increased significantly for the "Different" responses to the correct pairs, while the "Same" responses to the correct pairs decreased.

The fact that no significant changes occurred between Test 1 and Test 2 during Trials 1-5, while four out of six measures showed significant increases or decreases on Trials 6-14 suggests that performance on the two tests was more similar during the early stages of learning than during the later ones. Testing this hypothesis would necessitate a between-groups comparison. As it will be recalled, the variances of the different groups were found to be heterogeneous, which violates one of the assumptions of parametric tests between these groups. However, t tests were tentatively used to test the hypothesis that the mean difference scores for the first five test-trials were different from those of the second five test-trials. Although the heterogeneity of the variances of the samples would tend to inflate the value of t in these tests, no significant differences were found (Table II, Column D).

The occurrence of significant changes from Test 1 to Test 2 in the last five test-trials and not in the first five can be explained by the fact that variability of response was greater during the initial

trials than during the later ones. This greater variability for the first five test-trials would tend to obscure mean changes between the two tests. In support of this interpretation, F ratios for each of the six response-pair categories (where F = variance on Trials 1-5 divided by variance on Trials 6-14) are presented in Column E of Table II. In all six cases it can be seen that the variance on Trials 1-5 was greater than on Trials 6-14, although for "Different" responses to correct pairs the difference was not significant. It would seem, then, that the differences between the first and second recognition test scores became more stable during the last five test-trials. However, it is difficult to conclude whether the non-significance of the difference scores on the first five trials is due to this greater variability, or to a genuine lack of mean difference.

Regardless of this difficulty, however, it can be concluded from these results that interposition of a recognition test-trial between learning and a second recognition test-trial at least at times has significant effects on performance on the second test. It is obvious that caution must be exercised in pooling the results of Test 1 and Test 2 for a given pair-category combination.

II. GENERALIZED RESPONSES AS A FUNCTION OF TRAINING

One of the two pair-category combinations in which no significant changes in performance from the first to the second test were found was "Same" responses to incorrect pairs, and this combination also had the smallest mean difference of the six studied. This is the particular

pair-category combination which was intended to furnish the measure of stimulus generalization to be used in this thesis. As there was no difference in performance between Tests 1 and 2 for this measure, the results of the two tests were pooled to yield a more stable score for each subject. These pooled results are shown graphically in Fig. 7.

An analysis of variance trend test could not be used with these data because of the marked heterogeneity of variance between the groups, so a non-parametric test was necessary. A Kruskal-Wallis one-way analysis of variance by ranks indicated a highly significant difference between the scores for the different trial groups ($H = 35.45$ after correction for tied ranks; $p < .001$). Mann-Whitney U tests showed that the decrease between Trials 1 and 14 (Fig. 7) was significant ($p < .004$). None of the Mann-Whitney Us for the observed increases (i.e., from Trials 2 to 3, 4 to 5, and 6 to 8) in Fig. 7 were significant. Thus it would appear that the number of "Same" responses to the incorrect pairs decreased monotonically as a function of practice.

There are a number of ways in which this finding can be interpreted. (1) It is possible that these incorrect recognitions are a function of stimulus generalization, which is at a maximum early in learning and subsequently subsides. (2) Although an attempt was made to eliminate response generalization by using distinctive response items, the possibility still exists that the incorrect recognitions are due to confusion of responses, which decreases with practice. (3) The incorrect recognitions can be considered as mislearned associations, due perhaps to prior learning, which were strongest in the early stages of learning, and which were subsequently extinguished through

nonreinforcement. (4) It is possible that the instructions were not successful in reducing guessing behaviour to a negligible level. If this were the case, guessing would be expected to occur most frequently in the early stages of practice, and would decline steadily as more correct responses were learned.

Although these four hypotheses appear equally likely from the evidence presented in Fig. 7, the plausibility of each can be checked by testing certain deductions that can be inferred from each hypothesis. These deductions concern the manner in which the subjects respond to the correct pairs containing the stimulus and response items of the mispairings that they incorrectly recognized.

It will be recalled that the subjects were shown each stimulus and each response item twice on the recognition test: once when the correct stimulus-response pairing was presented, and once when each item was mispaired with another item. For every incorrect pairing there are two correct pairs, one containing the stimulus item and the other the response item (e.g., for the incorrect pairing Sa-Rb, we have the two correct pairs Sa-Ra and Sb-Rb). By examining the subject's recognition responses to these correlated correct pairs we can test the above hypotheses.

All of the tests reported below were done by separating the eight correct pairs in the recognition tests into two groups - the experimental group, comprised of those correct pairs containing an item (stimulus or response item, as the case might be) that had occurred in a mispairing that elicited a misrecognition; and, as a control group, those correct pairs which contained items occurring in mispairings that had not caused

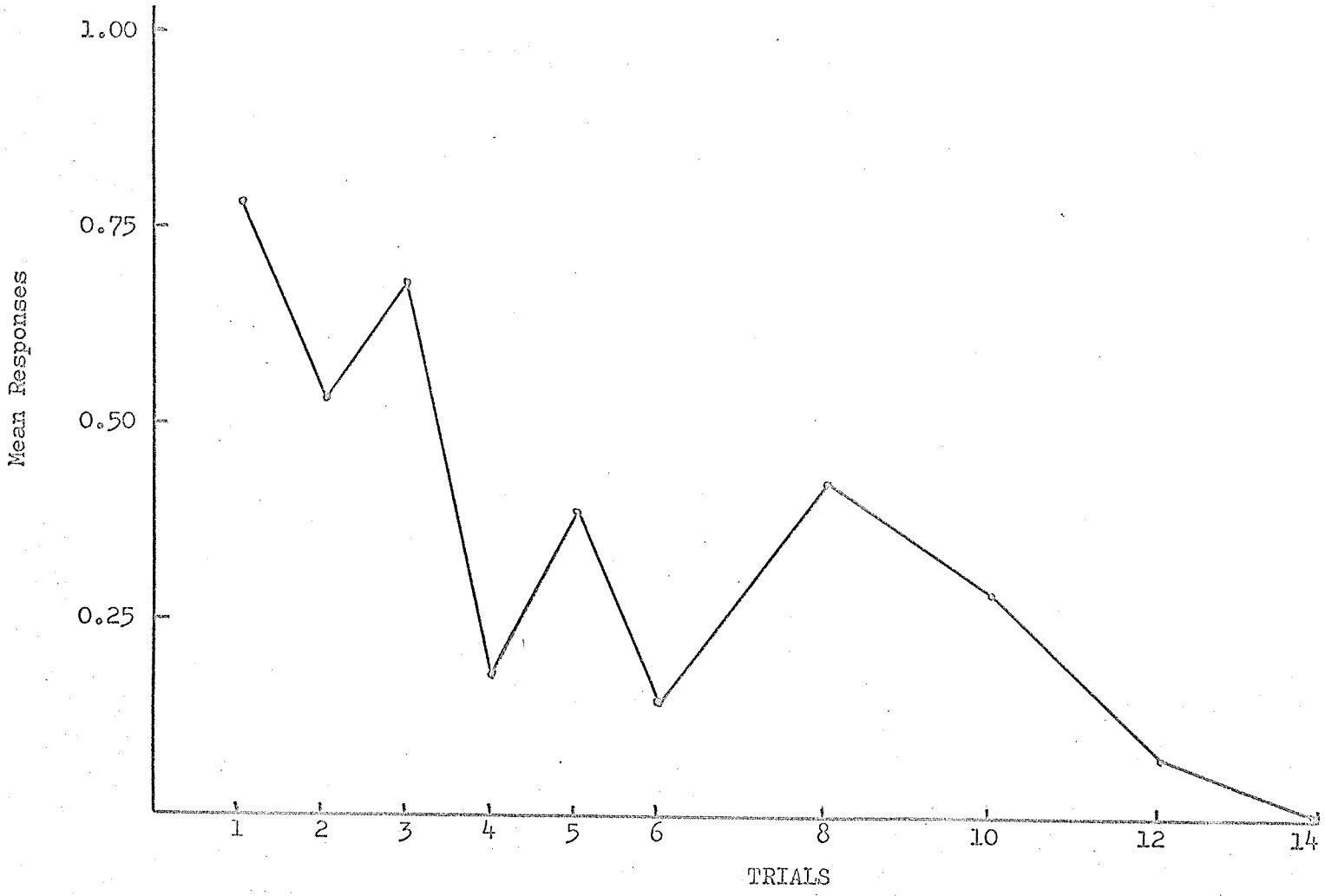


Figure 7. Pooled mean "Same" responses to incorrect pairs.

a misrecognition. The proportion of "Same", "Different" or "Don't Know" responses to each of these groups was compared. A subject was assigned to a "More" category if he responded more (with "Same" for example) to the group correlated with the misrecognitions than he did to the control group, and to a "Less" caregory if the converse occurred. If the two proportions were equal the results were not used in the test. The probability of the occurrence of a given response distribution was determined by comparison with the binomial distribution.[2]

The above four hypotheses can be stated in terms of this analytical scheme as follows:

(1) If the incorrect responding to the mismatched pairs is due to stimulus generalization (i.e., to the subject's confusing the stimulus item in the mismatched pair with one in a pair which he has previously learned correctly), then it should be found that the subject can recognize the pair in which the response term is correctly coupled (e.g., if stimulus generalization from Sa to Sb has caused misrecognition of the pair Sa-Rb, then the pair Sb-Rb should have been previously learned, and should have been correctly recognized). This deduction was tested by comparing the proportion of "Same" responses to the experimental group of correct pairs (i.e., those containing response items which had elicited generalized responses) with the proportion of "Same" responses

---

[2]In the statistical tests which follow in this section, all samples of $N \leq 20$ were compared with an exact binomial distribution. For samples with $N > 20$, a normal approximation to the binomial distribution was used.

given to the control group (i.e., those not containing response items which had elicited generalized responses). If the hypothesis were true then it should be found that the former proportion is significantly greater than the latter.

TABLE III

NUMBER OF SUBJECTS RESPONDING "SAME", "DIFFERENT", OR "DON'T KNOW" MORE OR LESS TO CORRECT PAIRS CORRELATED WITH RESPONSE ITEMS CAUSING GENERALIZED RECOGNITIONS (EXPERIMENTAL PAIRS) THAN TO CORRECT PAIRS CORRELATED WITH RESPONSE ITEMS NOT CAUSING GENERALIZED RECOGNITIONS (CONTROL PAIRS).

	"Same"			"Different"			"Don't Know"		
Trials	More	Less	Ratio	More	Less	Ratio	More	Less	Ratio
1-2	10	8	1.25	7	8	0.88	3	13	0.23*
3-5	10	5	2.00	6	6	1.00	5	8	0.63
6-14	8	5	1.60	2	7	0.29	3	7	0.43
Total	28	18	1.56	15	21	0.71	11	28	0.39**

* $p < .01$ **$p < .005$

Table III shows the results of a median test of this deduction. The data from different groups have been pooled to give reasonably large Ns. Also shown are similar tests for "Different" and "Don't Know" responses. It can be seen that although there was no significant difference indicated between the overall frequencies of "Same" responses made to the experimental pairs and to the control pairs, all three comparisons show a trend in the predicted direction.

(2) If it were true that the misrecognitions were caused by response generalization, a similar argument could be extended to the proportion of "Same" responses made to the experimental pairs containing the stimulus items (e.g., if pair Sa-Rb were misrecognized due to confusion of Rb for Ra, then pair Sa-Ra should have been correctly recognized). Table IV

shows the results of a median test of this deduction. It can be seen that there is no significant difference between the two frequencies.

TABLE IV

NUMBER OF SUBJECTS RESPONDING "SAME", "DIFFERENT" OR "DON'T KNOW" MORE OR LESS TO CORRECT PAIRS CORRELATED WITH STIMULUS ITEMS CAUSING GENERALIZED RECOGNITIONS (EXPERIMENTAL PAIRS) THAN TO CORRECT PAIRS CORRELATED WITH STIMULUS ITEMS NOT CAUSING GENERALIZED RECOGNITIONS (CONTROL PAIRS).

	"Same"			"Different"			"Don't Know"		
Trials	More	Less	Ratio	More	Less	Ratio	More	Less	Ratio
1-2	10	7	1.43	6	10	0.60	4	12	0.25*
3-5	10	6	1.67	1	11	0.09**	6	8	0.75
6-14	5	8	0.63	2	6	0.33	6	4	1.50
Total	25	21	1.19	9	27	0.33***	16	24	0.67

*$p < .05$ **$p < .005$ ***$p < .001$

(3) If the incorrect recognitions can be accounted for by mislearning (e.g., because the subject, perhaps due to a previous association, considered pair Sa-Rb to be correct, and pairs Sa-Ra and Sb-Rb to be incorrect), it would be expected that the subjects would respond "Different" to the correctly matched stimulus and response items which had been involved in misrecognitions. In Table IV it can be seen that there is a significant tendency over all trials for the subjects to respond in the opposite direction to this deduction when the correctly paired response terms are studied. From Table III it can be seen that there are fewer "Different" responses to the experimental pairs than to the control pairs, although the difference is not significant. However, the pooled "Different" responses to both the generalization-producing stimulus and response items are significantly less frequent ($p < .002$) than are given to the control pairs. It would seem that a mislearning hypothesis cannot account for the misrecognitions.

UNIVERSITY OF MANITOBA LIBRARY

(4) If the misrecognitions are due to guessing randomly at some of the incorrect pairs because the subjects did not know the correct pairings then either or both of two situations should be found.

a) The subjects should have responded "Don't Know" more to the correctly paired items (stimulus and response) that were misrecognized than to the controls.

b) There should be no difference in the rate of responding "Same" and "Different" between the experimental pairs (stimulus and response) and the controls.

It can be seen from Table III that over all trials the subjects responded "Don't Know" significantly less to the correctly paired response items which had caused misrecognitions than to the control pairs. The same significant trend can be seen over Trials 1-2 in Table IV for the stimulus items. In addition, the pooled "Don't Know" responses of Tables III and IV are significantly lower ($p < .002$) for the generalization-producing items than for the controls. Thus deduction (4a) from the guessing hypothesis, that there is a higher level of "Don't Know" response to the experimental pairs than to the controls, is contradicted by the data.

Concerning deduction (4b) of the guessing hypothesis, it has been shown earlier that the overall level of "Different" responses to the correctly paired generalization-producing items (stimulus and response) is significantly lower than for the controls. The most reasonable conclusion seems to be that deduction (4b), that both "Same" and "Different" responses should not differ between experimental and control pairs, is not supported by the data.

III. ANTICIPATION RECALL

Fig. 8 shows the mean number of overt errors given on the recall test on each trial. Each point is based on the total number of subjects who had

been tested on that given trial. (Because different groups continued anticipation learning for different numbers of trials before their recognition test, the points representing earlier trials are based on larger Ns than later trials, with no results being obtained for Trial 14). It can be seen that the curve rises to a peak around the third trial and then decreases, which is in accord with the results found by Gibson (1942) and Gagné (1950).

IV. COMPARISON OF THE RECOGNITION AND ANTICIPATION RECALL MEASURES

In measuring the communality of performance on the recall and recognition tests a problem was encountered, necessitating indirect methods of measurement. It will be remembered that the subjects were given the recognition test following their last scheduled anticipation learning trial. A recall score was thus available for each subject from the results of the anticipation learning trial before the recognition test. However, it must be realized that each anticipation trial is two things -- first, a test of what had been learned on the previous trials, and, second, a study or learning trial. If a subject had been assigned to complete n trials of learning, then his recognition score would measure his performance after n trials of learning, but his last recall score would be measuring performance after n-1 trials. It can be seen that a given subject's recall score is always one trial behind his recognition score.

To obtain a complete view of the communality of the two tests, two methods were used. The first compared the correlated recall and recognition scores for each subject (within-subjects comparison) to determine if the particular subject responded consistently to the particular pairs on the different tests. The second compared the mean number of correct responses as measured by the two tests on each trial, by comparing the

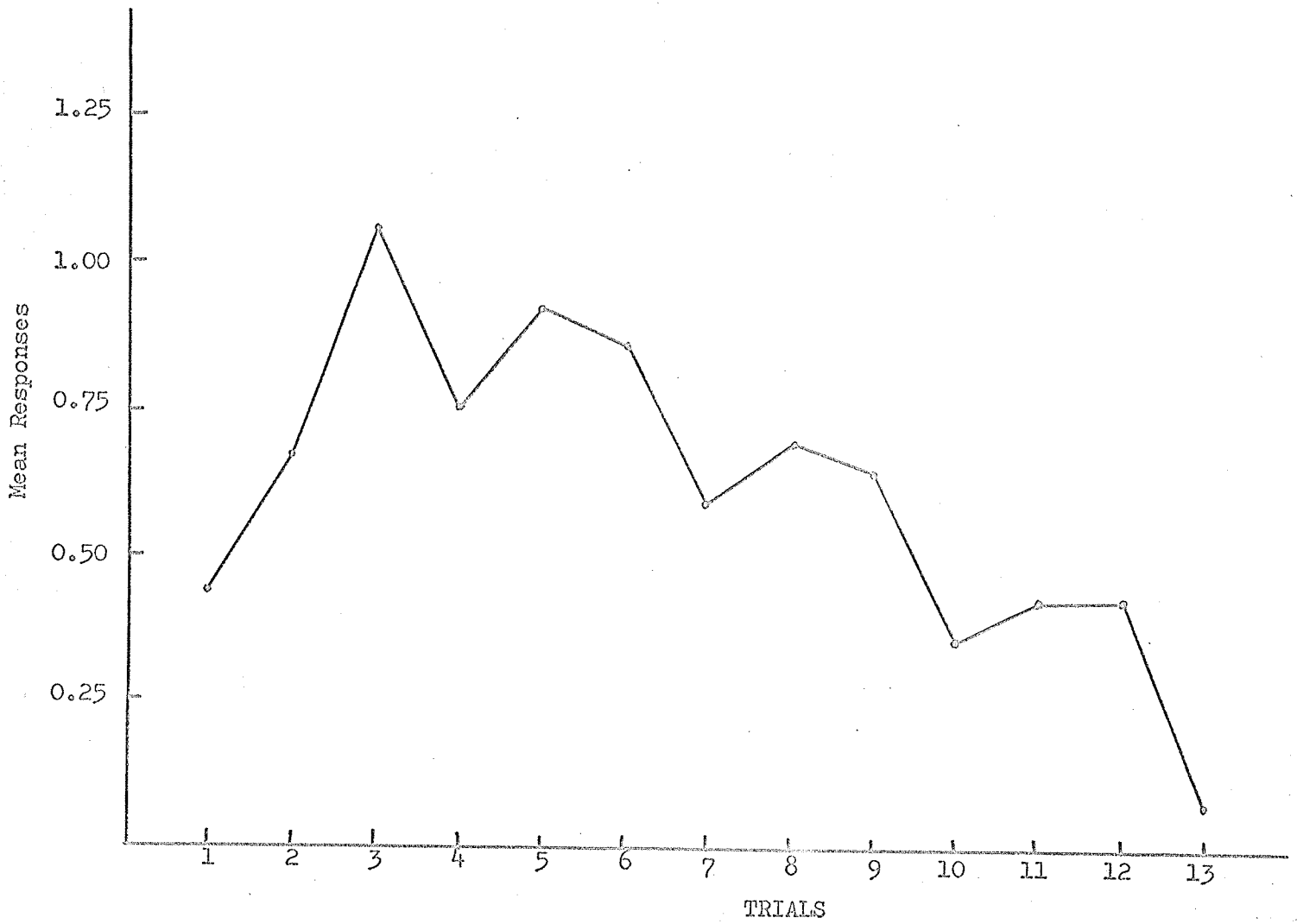


Figure 8. Mean number of overt errors on anticipation recall test.

performance of different subjects (between-subjects comparison) so that the number of pre-test trials could be equalized for recall and recognition.

Within-subjects Comparison

For the within-subjects comparison, paired scores of the responses given on the two tests to each of the eight pairs were obtained for each subject. Only the first recognition test score was used, as it will be recalled that the level of response changed on the second test in some categories. For the purpose of this analysis the recall test responses were scored as "Correct" if they had been correctly anticipated, and as "Incorrect" if they were incorrectly anticipated or if no response was given. The recognition test responses were scored as "Correct" if they were correctly recognized (i.e., the subject responded "Same" to a correct pair or "Different" to an incorrect pair); and as "Incorrect" if incorrectly recognized (i.e., the subject responded "Same" to an incorrect pair or "Different" to a correct pair), or if the subject responded "Don't Know". A fourfold contingency table was made (Table V) comprising all of the paired responses for all subjects except those tested immediately after the first trial. These were not included because no recall test had been given to them. The phi coefficient for this contingency table was +.327, which would indicate only a low correlation between the two measures.

TABLE V

CORRELATION BETWEEN NUMBER OF RESPONSES CORRECT OR INCORRECT ON RECALL TEST AND NUMBER OF RESPONSES CORRECT OR INCORRECT ON RECOGNITION TEST

		Recognition Test		
		Incorrect	Correct	Total
Recall Test	Correct	43	520	563
	Incorrect	145	300	445
	Total	188	820	1008

Φ = +.327

An examination of Table V will show that almost all of the items responded to correctly on the recall test were also correct on the recognition test. However, the large majority of the items not recalled were correctly recognized. Thus, although a correct response on the recall test indicates almost certainly that the item would be correctly recognized, the converse does not hold. In fact, an incorrectly recalled response was more likely than not to be correctly recognized.

There are two reasons which might account for part of this disparity. The first, which has been previously mentioned, is that, in effect, the recall and recognition tests are measuring degree of learning on different trials. An estimate was made of the number of responses acquired on this "extra" learning trial, in order to determine approximately what the level of response would have been without it. This was done by calculating the difference in the level of "Correct" responses (as defined above in connection with Table V) between the recognition tests on Trial 1 and on Trial 14. This difference, it is felt, is a reasonable estimate of the total number of responses acquired on the extra learning trial by the nine groups of subjects concerned. It was found to be 36 responses. This represents an estimate of the total number of correct recognitions by all subjects that resulted from the exposure to the pairs between the time that each subject was given the recall test and when he was given the recognition test.

A second factor which would tend to distort the apparent relation between performance on the recall and recognition tests is the effect of guessing on the recognition test. Although it is suspected that very little guessing occurred, any guessing at all would inflate the proportion of correct responses on the recognition test, as compared to the recall

test. Therefore, a correction was made. This was done by counting the number of pairs that were incorrectly recognized on Trials 2 to 14 and assuming that an equal number had been guessed correctly.[3] This yielded an estimate of 121 correct guesses.

It is possible that these two factors did inflate the number of "Correct" responses shown in Table V for the recognition test, but which category of response they should be assumed to have affected in the recall test is difficult to determine. The items learned between the two tests (i.e., on the last anticipation trial) would presumably have not been correctly recalled on that test. But it is impossible to estimate, in the case of the correct guesses, what proportions should be paired with correct and incorrect recalls. That is, which <u>particular</u> correct recognitions resulted from guessing?

However, if it is assumed that only the items that had been incorrectly recalled were guessed, then we can calculate the maximum possible correlation between the two measures under these conditions. The corrected table is shown in Table VI, from which it can be seen that the phi coefficient in these circumstances is +.607. It will be recalled that a number of assumptions have been made in calculating this figure, which most probably has resulted in an underestimation of the true recognition score. In spite of

---

[3]It can be seen that this operation apparently equates the measure of stimulus generalization used in this thesis with guessing. This correction factor was used, however, because it provided an estimate of the minimum possible recognition score. It is realized that this estimate most probably errs on the conservative side.

this conservative estimate, however, it can be seen that approximately one third of the pairs that were incorrectly recalled were recognized correctly.

TABLE VI

CORRELATION BETWEEN NUMBER OF RESPONSES CORRECT OR INCORRECT ON RECALL TEST AND NUMBER OF RESPONSES CORRECT OR INCORRECT ON RECOGNITION TEST, ADJUSTED FOR EXTRA LEARNING AND FOR GUESSING.

		Recognition Test		
		Incorrect	Correct	Total
Recall Test	Correct	43	520	563
	Incorrect	302	143	445
	Total	345	663	1008

$\phi = +.607$

Between-subjects Comparison

The second approach to studying the relationship between the recall and recognition tests is to find the mean number of correct responses given to the two types of test after equal amounts of training. By using a between-subjects comparison for the two tests the amount of practice previous to each test was equalized. This was done by comparing the mean recognition score of each group with the mean recall score of all subjects tested on the following trial. Again, to find the most conservative estimate of the recognition score it was assumed that just as many pairs had been correctly guessed as had been misrecognized. The recognition test scores were corrected for this guessing by the subtraction of a number of responses equal to the total number of "Same" to incorrect and "Different" to correct pairs on the same trial.

Fig. 9 shows a comparison of the learning curves as measured by the recall, recognition and corrected recognition tests. The points on the recall test curve were based on different Ns, as was explained earlier. Because the recall test given on a particular trial measured learning up to and including the previous trial only, no results are reported for Trial 14. Each point on the recognition test curve represents a different group of subjects, the one that had been tested after that particular anticipation trial.

An examination of Fig. 9 shows that the general level of response is appreciably higher on the recognition test, even when stringently corrected, for the first six trials. On the subsequent trials the uncorrected recognition score tends to be only slightly greater than the recall score, while the correction factor brings the recognition score to below or only equal to the recall score. It would seem that recognition scores are greater than recall scores early in learning, but that this superiority disappears in the later stages of practice with the list.

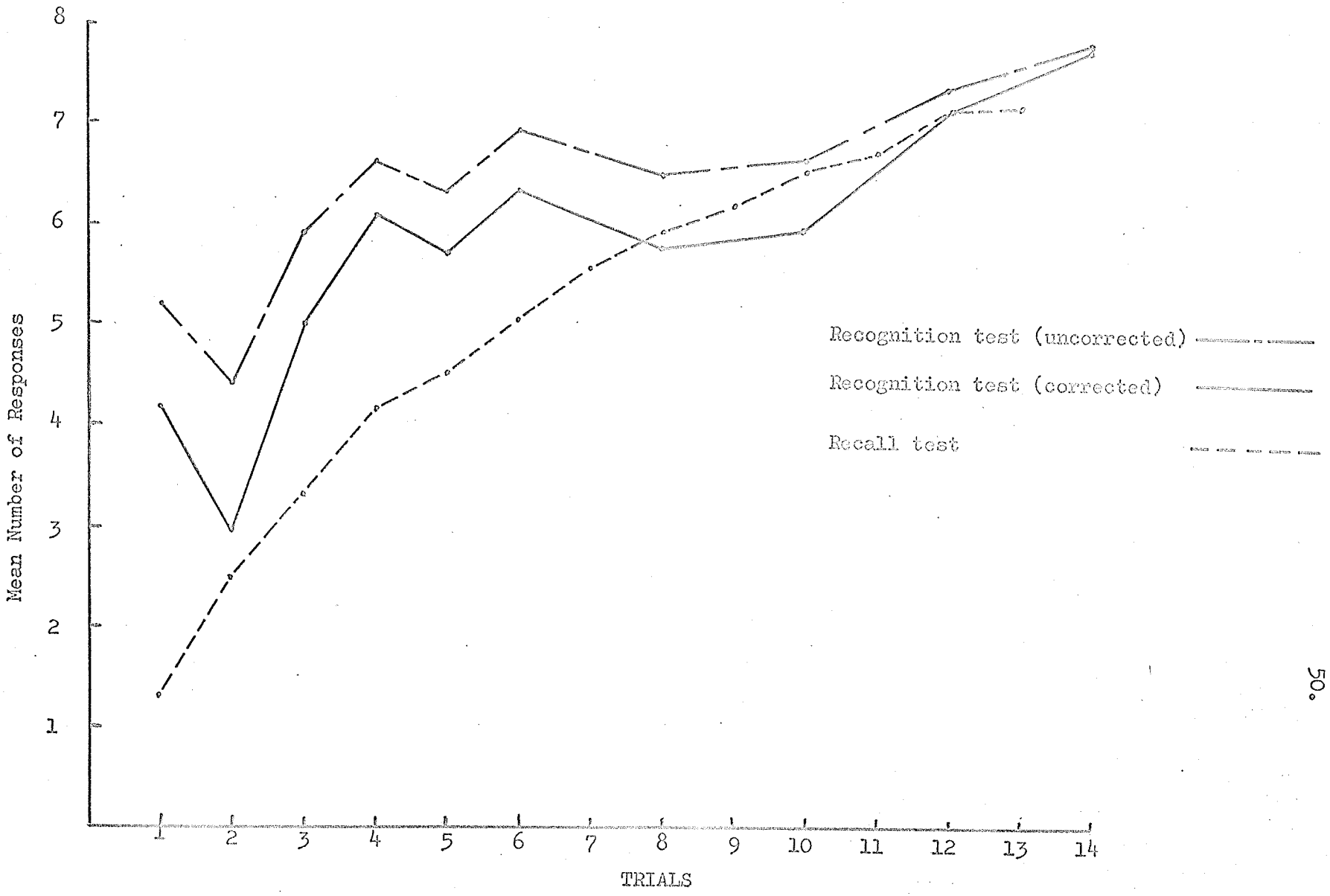


Figure 9. Comparison of mean recall, mean recognition and corrected mean recognition.

CHAPTER IV

DISCUSSION

There are three possible explanations of the marked differences found in this experiment between recall and recognition test performance. The superiority of recognition over the recall measure in the early stages of learning is almost certainly partly due to the elimination of the necessity for response-acquisition prior to associative learning. It will be recalled that one of the major purposes in using the recognition test in this experiment was to control the effects of the subject having to learn the responses before he could begin responding, correctly or incorrectly, to the stimuli. It is to be expected, then, that the recall measure would be under a relative disadvantage, compared to the recognition test, in the early stages of learning.

However, it will also be recalled that presumably there is another difference between the two types of test. As was discussed earlier, there is reason to believe that the recognition test can detect weaker associations than the recall test is capable of doing. It is possible, then, that on the first few learning trials the recognition score would be higher than the recall score because most of the associations at that time would be quite weak. This difference, of course, would gradually disappear as associations became sufficiently robust to rise above the thresholds of both types of test.

It is also possible that the difference in the results of the two tests can be partly attributed to the particular conditions of

this experiment. It will be remembered that subjects were given only 2 sec. to respond to the stimuli on the recall test, while they were allowed 4 sec. to respond on the recognition test. It can be seen that this disparity would give the recognition test an advantage in this experiment that would operate in the same way as the presumably inherent superiority discussed in the previous paragraph.

Although no evidence can be found from this experiment to indicate the effects of these three factors, this could easily be provided by additional experimentation. The relative effects of each could be separated by a response-familiarization procedure similar to that used by Underwood & Schulz (1960), and allowing equal response intervals for both tests. However, previous work (Underwood & Schulz, 1960) indicates that the elimination of the response-acquisition phase of learning by prior response familiarization produces higher recall test performance over the first few trials of practice with a list. It is fairly certain, then, that eliminating the need for response acquisition in this experiment by the use of the recognition test can account for a significant amount of the reported superiority of the recognition test over recall.

Concerning the course of generalization tendencies as a function of practice, no support could be found for Gibson's (1940) hypothesis that generalization increases during the early stages of learning before it decreases. Instead, the measure of stimulus generalization used in this thesis (misrecognitions of incorrect pairs) was found to be at a maximum after the first learning trial, and to decrease monotonically thereafter. This supports the prediction made by Murdock (1958).

The data were examined further to determine if factors other than stimulus generalization might be responsible for the high initial rate of misrecognitions. The hypothesis that subjects had mislearned certain incorrect pairings was rendered unlikely by the finding that fewer "Different" responses were given to the correctly paired generalization-causing stimulus and response items than would be expected from the control items. It was also found that fewer "Don't Know" responses were given to the experimental pairs than to the controls, and that "Same" and "Different" responses to the experimental pairs were not randomly distributed. This would seem to indicate that the subjects were confident that their "Same" responses to the mispaired stimulus and response items were correct, and that they were not guessing randomly.

It must not be construed from this last statement, however, that it has been shown that no guessing whatsoever occurred. Although the evidence indicates that complete guessing to the stimuli in both correct and incorrect pairs does not seem to have occurred, the possibility of the occurrence of what might be called "partial" guessing cannot be disproved. That is, it is quite possible that the subjects, for the pairs which they had learned only slightly, were capable of recognizing the correctly paired items. However, a higher response strength might be necessary for the recognition of mismatched items as incorrect, so that in the early stages of learning these might be guessed at randomly. If this hypothesis were correct the results could be expected to be approximately the same as those reported.

But although the possibility of "partial" guessing cannot be directly discounted by the data, for a number of indirect reasons

this hypothesis would seem unlikely. First, the instructions emphasized to the subjects that they were not to guess. Second, the subjects gave the "Don't Know" response most frequently on the first trial, indicating that this response was used most when it would be expected that the subjects were most uncertain. Third, the occurrence of "partial" guessing (to incorrect pairs only) without the occurrence of completely random guessing (to correct and incorrect pairs) would seem unlikely. However, no evidence that subjects guessed randomly could be found.

Nevertheless, these arguments are only suggestive, and the possibility of "partial" guessing still cannot be abandoned. But if this possibility were to be considered, it might prove awkward to handle theoretically. Guessing behaviour that occurs only to the pairs that are expected to elicit generalized responses would be difficult to separate, conceptually as well as operationally, from generalization.

In spite of the effort made to eliminate response generalization in this study, no evidence could be found to indicate that the effect of stimulus generalization was significantly greater than that of response generalization in producing misrecognitions, although a slight difference was found. In fact, it could not even be shown that the effect of stimulus generalization was significantly greater than chance. It should be pointed out that Gibson (1940, 1942) made no attempt to demonstrate that it was indeed stimulus generalization, and not some other factor, that produced the results that she reported. Underwood (1961) has pointed out that the role that response generalization plays in causing interference in verbal learning was probably underestimated by Gibson. Although an attempt was made in this study to clear up this question,

the results appear to be inconclusive. The most reasonable conclusion seems to be, in view of the elimination of random guessing and mislearning as reasonable explanations for the misrecognitions found in this experiment, that a combination of stimulus and response generalization was the most likely cause of these misrecognitions.

Another objection that might be made to the results of this study is that the experiment could be atypical of the usual paired-associate situation. If this were true (e.g., if the list were unusually easy to learn), it is possible that the increase in generalization that Gibson predicted occurred completely during the first learning trial and hence was not detected. However, this hypothesis does not appear likely when the results of the recall test are considered. Mean correct recalls started at a low level on the first trial and increased smoothly, while overt errors on the recall test showed the usual increase during the early trials and a subsequent decrease. As no deviation from the usual pattern of response is readily apparent on the recall test, it would seem that the failure of the recognition test results to confirm Gibson's prediction is not due to procedural factors specific to this experiment.

This does not rule out the possibility, of course, that an undetected increase in generalization did occur in this experiment. From the results of the recognition test, even when corrected for guessing, it can be seen that over half of the pairs had been learned by the first trial. It is conceivable that if the rate of learning had been retarded on this first trial, an initial rise in generalization might have been found. Although an attempt <u>had</u> been made to slow the rate of learning, the results show

that a considerable amount was learned on the first trial. However, this study has been able to demonstrate that if a rise and fall in generalization does take place early in learning, this peak is not related to the rise and fall of overt errors in recall.

To conclude, the most feasible explanation for the failure of this experiment to confirm Gibson's (1940) hypothesis that generalization rises and then falls would seem to be that this phenomenon occurs only when the subject is preoccupied with response acquisition in the early stages of learning. Although further studies would be necessary to determine the role of other factors such as the time allowed for response on the two types of test, and the relative sensitivities of the two tests, from available evidence it would seem that, in the purely associative phase of learning, generalization decreases monotonically with practice.

From an examination of the various statements of Gibson's theory (1940, 1942), it would seem that she made two implicit assumptions in formulating her postulate about generalization. The first was that all responses are readily available to the subject at all stages of the learning process. Although this was not directly stated, it would seem that it had been assumed, as nowhere does Gibson approach the point that unfamiliar responses, such as the nonsense syllables that she used in her experiment (1942), might be more difficult for a subject to produce than familiar words. This was not an unreasonable assumption at the time, considering that the first suggestion of the response-acquisition phase of learning as distinct from the purely associative phase was not made until 1952, by Hovland & Kurtz. Nevertheless, it is obvious from the previous discussion of this point

that Gibson's theory is weakened by its failure to allow for the effects of response acquisition.

Another assumption that Gibson seems to have made is that the strength of the correct associations is relatively low at the beginning of learning. From this it seemed a logical conclusion that generalized associative strength would necessarily be negligible as well during the early trials. However, the results of this experiment indicate that the low performance on the first trial of the typical verbal learning experiment could be due to the fact that the recall test that is usually employed in these studies taps associative and response learning. This study has shown that if a recognition test, even when corrected for guessing, is used as a measure of learning, over half the associations have been learned after only one trial. The recall score, on the other hand, indicates that only 17% of the associations have been learned. If it could be confirmed that correct associative strength were quite appreciable in the early stages of learning, it would be no surprise if generalized response strength were also high, and that it decreased steadily with further practice. It is clear that this question can only be answered by further study of comparative performance on the recognition and recall tests.

CHAPTER V

SUMMARY AND CONCLUSIONS

Gibson has hypothesized that stimulus generalization in verbal learning shows an initial rise before decreasing. However, as criticisms have been made of the evidence cited as support for this hypothesis, a test was devised which seemed more suited for measuring generalization than those used formerly.

The procedure was as follows: Each of 10 groups of 14 subjects attempted to learn a list of eight paired-associates for either 1, 2, 3, 4, 5, 6, 8, 10, 12 or 14 trials by a standard anticipation recall method. Immediately after completing this practice each subject was shown, one at a time, a random sequence of the eight pairs and eight mispairings of the correct stimuli and responses. The subjects were instructed to respond "Same" if they recognized a pair as correct, "Different" if they recognized it as incorrect, and "Don't Know" if they were uncertain. A generalized response was defined as a "Same" response to an incorrect pair.

Generalization was found to decrease monotonically as a function of practice under these conditions. Several other hypotheses were found that might provide alternate explanations for these results, such as simple mislearning and guessing, but these were rendered unlikely when other aspects of the data were considered.

A comparison of the results of the recall and recognition tests, with correction made for guessing on the latter, showed that recognition learning scores were higher than recall scores in the early part of the learning curve. However, performance on the two tests was almost identical toward the end of practice.

It would seem that in the typical verbal learning experiment, where performance is measured by a recall test the level of response does not indicate the true strength of the associative bonds in the early stages of learning because the subject cannot readily produce any response, correct or incorrect. Because of this it has been generally considered that the associative bonds between stimuli and responses are negligible until a considerable amount of exposure to the pairs has been given the subject. This experiment seems to show that these associative bonds are greater early in learning than the recall test scores would indicate. However, a test like recognition which apparently can tap associative strength directly seems to indicate that comparatively little practice is necessary to produce measurable bonds between stimuli and responses, both correct and generalized.

Further study is required, however, to confirm that the results of this experiment are attributable to the elimination of the need for response learning. Although the evidence from other studies supports this contention, the effects of methodological factors cannot be separated from the effects of response learning in this experiment. Until this study has been repeated with these factors controlled these conclusions can only be held tentatively.

BIBLIOGRAPHY

Battig, W. F. Comment on "Intralist generalization in paired-associate learning," Psychol. Rev., 1959, 66, 338-339.

Beritoff, J. S. On the fundamental nervous processes in the cortex of the cerebral hemispheres. I. The principal stages of the development of the individual reflex: its generalization and differentiation. Brain, 1924, 47, 109-148.

Fisher, R. A., & Yates, F. Statistical tables for biological, agricultural and medical research. (4th ed.), London: Oliver and Boyd, 1953.

Gagné, R. M. The effect of sequence of presentation of similar items on the learning of paired associates. J. exp. Psychol., 1950, 40, 61-73.

Gibson, Eleanor J. A systematic application of the concepts of generalization and differentiation to verbal learning. Psychol. Rev., 1940, 47, 196-229.

Gibson, Eleanor J. Intra-list generalization as a factor in verbal learning. J. exp. Psychol., 1942, 30, 185-200.

Gibson, Eleanor J. A re-examination of generalization. Psychol. Rev., 1959, 66, 340-342.

Hartman, T. F., & Grant, D. A. Differential eyelid conditioning as function of the CS-UCS interval. J. exp. Psychol., 1962, 64, 131-136.

Hilgard, E. R., Campbell, A. A., & Sears, W. N. Conditioned discrimination: The development of discrimination with and without verbal report. Amer. J. Psychol., 1937, 49, 564-580.

Hovland, C. I. The generalization of conditioned responses: II. The sensory generalization of conditioned responses with varying intensities of tone. J. genet. Psychol., 1937, 51, 279-291.

Hovland, C. I., & Kurtz, K. H. Experimental studies in rote-learning theory. X. Pre-learning syllable familiarization and the length-difficulty relationship. J. exp. Psychol., 1952, 44, 31-39.

Kimble, G. A. Hilgard and Marquis' conditioning and learning. New York: Appleton-Century-Crofts, 1961.

Kimmel, H. D., & Pennypacker, H. S. Differential GSR conditioning as a function of the CS-UCS interval. J. exp. Psychol., 1963, 65, 559-563.

Luh, C. W. The conditions of retention. Psychol. Monogr., 1922, 31, No. 142, 1-87.

Murdock, B. M., Jr. Intra-list generalization in paired-associate learning. Psychol. Rev., 1958, 65, 306-314.

Murdock, B. M., Jr. An analysis of the recognition process. In C. N. Cofer, Barbara Musgrave (Eds.), Verbal behavior and learning: Problems and Processes. New York: McGraw-Hill, 1963, Pp. 10-32.

Pavlov, I. P. Conditioned reflexes. (Translated by G. V. Anrep). London: Oxford University Press, 1927.

Postman, L., Jenkins, W.O., & Postman, D. L. An experimental comparison of active recall and recognition. Amer. J. Psychol., 1948, 61, 511-519.

Postman, L., & Rau, L. Retention as a function of the method of measurement. Univer. Calif. Publ. Psychol., 1957, 8, 217-270.

Raben, Margaret W. The white rat's discrimination of differences in intensity of illumination measured by a running response. J. comp. Physiol. Psychol., 1949, 42, 254-272.

Razran, G. Stimulus generalization of conditioned responses. Psychol. Bull., 1949, 46, 337-365.

Runquist, W. N. Remarks on "Intralist generalization in paired-associate learning". Psychol. Rev., 1959, 66, 343-344.

Spiker, C. C., & Holton, Ruth B. Associative transfer in motor paired-associate learning as a function of amount of first-task practice. J. exp. Psychol., 1958, 56, 123-132.

Underwood, B. J., & Schulz, W. S. Meaningfulness and verbal learning. New York: J. B. Lippincott Co., 1960.

Underwood, B. J. The effect of successive interpolations on retroactive and proactive inhibition. Psychol. Monogr., 1945, 59, No. 3.

Underwood, B. J. An evaluation of the Gibson theory of verbal learning. In C. N. Cofer (Ed.), Verbal learning and verbal behavior. New York: McGraw-Hill, 1961, Pp. 197-223.