MONTANA STATE UNIVERSITY

The effects of subject's reading achievement level in a recognition memory task
by Anne Thompson-Gibbs

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Psychology
Montana State University
© Copyright by Anne Thompson-Gibbs (1982)

Abstract:
Studies of recognition memory usually examine the effects of variables related to the stimulus materials, one of the most common being the meaningful ness of brief letter strings. The subject in a recognition study may use meaningfulness as a cue in the decision process. Conceivably the subject's reading achievement level may directly affect letter string recognition due to differences in accessing letter string codes during the recognition process. A signal detection paradigm was used to obtain measures of performance for the two reading achievement groups in a recognition task varying the meaningful ness of CVC trigrams. Non-parametric measures were employed to evaluate differences in sensitivity and response bias. While there were no differences in the levels of sensitivity as a function of the subject variable, there were significant differences in response bias, suggesting that the less capable reader does not process letter strings in the same manner as the skilled reader.

THE EFFECTS OF SUBJECT'S READING ACHIEVEMENT LEVEL IN A RECOGNITION MEMORY TASK

by

Anne Thompson-Gibbs

A thesis submitted in partial fulfillment
of the requirements for the degree

of

Master of Science

in

Psychology

MONTANA STATE UNIVERSITY
Bozeman, Montana

December 1982

MAIN LIB.
N378
T3795
cop.2

APPROVAL

of a thesis submitted by

Anne Thompson-Gibbs

This thesis has been read by each member of the thesis committee and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the College of Graduate Studies.

12/27/82 — Date

W D Short — Chairperson, Graduate Committee

Approved for the Major Department

12/27/82 — Date

W D Short — Head, Major Department

Approved for the College of Graduate Studies

12-28-82 — Date

Michael P Malone — Graduate Dean

STATEMENT OF PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a master's degree at Montana State University, I agree that the Library shall make it available to borrowers under rules of the Library. Brief quotations from this thesis are allowable without special permission, provided that accurate acknowledgment of source is made.

Permission for extensive quotation from or reproduction of this thesis may be granted by my major professor, or in his absence, by the Director of Libraries when, in the opinion of either, the proposed use of the material is for scholarly purposes. Any copying or use of the material in this thesis for financial gain shall not be allowed without my written permission.

Signature Anne Thompson Gibb

Date Dec. 16 1982

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Studies of recognition memory usually examine the effects of variables related to the stimulus materials, one of the most common being the meaningfulness of brief letter strings. The subject in a recognition study may use meaningfulness as a cue in the decision process. Conceivably the subject's reading achievement level may directly affect letter string recognition due to differences in accessing letter string codes during the recognition process. A signal detection paradigm was used to obtain measures of performance for the two reading achievement groups in a recognition task varying the meaningfulness of CVC trigrams. Non-parametric measures were employed to evaluate differences in sensitivity and response bias. While there were no differences in the levels of sensitivity as a function of the subject variable, there were significant differences in response bias, suggesting that the less capable reader does not process letter strings in the same manner as the skilled reader.

INTRODUCTION

Studies of recognition memory usually take into consideration the effects of variables related only to the stimulus materials. When the stimulus materials consist of letter strings or words, these variables often include approximation to the English language, meaningfulness, pronounceability, and familiarity. Underwood and Schulz (1960) suggested that there are a number of relationships between meaningfulness, pronounceability, and familiarity, and that one or all of these may play a part in the recognition of letter strings and words. The subject in a recognition study would use these factors as cues aiding the decision process.

Peterson, Peterson, and Miller (1961) found positive relationships between meaningfulness and short term memory with the single presentation of trigrams and three letter words. Keppel and Underwood (1967) varied response meaningfulness in a paired associate paradigm and found a positive relationship between meaningfulness and short term memory in a recognition memory task.

Raser (1970) studied the effects of meaningfulness utilizing a paired associate paradigm allowing for separation of both the stimulus and response meaningfulness. Using the four possible combinations of high and low meaningfulness stimuli, as defined by Noble (1961), he found that meaningfulness was an effective variable in recognition performance. Raser indicated that recognition was more likely when both members of a given pair were high in meaningfulness.

In an unpublished study, Thompson-Gibbs (1972) utilized a signal detection paradigm, presenting old and new CVC trigrams serially and sequentially, varying meaningfulness across four conditions. Trials where both old and new CVCs were high in meaningfulness yielded the best recognition performance.

While it is well established that meaningfulness provides effective cues to the subject in a recognition study, subjects' utilization of the cues may be directly affected by a subject variable not generally considered. Wallach (1963) approached the problem of subject variables in his study on the relationship of spelling achievement to the perception of letter strings. He found positive relationships between high levels of spelling achievement and the recognition of letter strings. Spelling and reading achievement are directly related to one another (Jastak, Bijou, and Jastak, 1965), as spelling is more or less the reverse of reading, translating sound into graphic symbols. Subjects' ability to recognize English words out of context, then, may be a potentially confounding subject variable that should be considered in recognition memory studies. This skill, generally defined as skilled recognition reading, or the ability to read individual words out of normal sentence structure, combines several subskills or components that may be directly related to recognition memory performance. These subskills are discussed by Eleanor Gibson in her 1969 analysis of reading skills.

The first of these is the acquisition of language, not only comprehending its meaning, but perceiving its various segments and orders of combination. The second skill is visual discrimination of the various letters, noting the contrasts and similarities in their form. Both these skills are originally developed during the preschool years and reach maturation during the early school years. Neither needs to be considered in a word recognition task unless the subject population is known to have defects in visual perception or is considered learning disabled.

The decoding of written language is the third skill. At this reading level the subject reads out in silent verbalization units what is directed by the written units. This may be an important skill in the recognition of letter strings and words, especially if the verbalization unit is not a word, but rather, a string of letters. If the letter string can be coded as meaningful, i.e., one that can be associated with known words or word-like patterns, the subject is more likely to recognize that sequence when it is presented to him as a recognition stimulus.

Jackson (1980) suggests that the better reader is more efficient in the processes of accessing letter codes, and that the unfamiliar or new item will be processed more quickly by the better reader. A unique item, i.e., one that is an unfamiliar letter pattern, may be processed serially by the less skilled readers due to their decreased automatization of recognition skills. This is supported by Healy's (1981) discussion of the rules readers use to process reading units. She suggests that a letter string is processed as a unit when it is normally contained in a frequent word, but that when the letter string is unfamiliar, it is processed in single letter units.

Gibson's fourth subskill related to the recognition of new words is the ability to read in higher order units . . . reading in chunks, utilizing all of the structural regularities found in the written language. It is this skill that facilitates the learning of new words, allowing the reader to attach meaning to the various combinations of written symbols. This skill is continuously developed and improved upon with practice, leading to the differences in performance that are measured by tests of word recognition.

Skilled recognition reading as discussed here is measured by one of the subtests of the Wide Range Achievement Test (WRAT) devised by Jastak, Bijou, and Jastak (1965). During development and standardization of the WRAT, research indicated that the test scores correlate highly with the verbal scores of various other individual intelligence tests, and with other standardized tests measuring academic achievement.

With the tendency for researchers to use readily available subjects and small groups for recognition studies, perhaps the ability to recognize words out of context should be considered a subject variable. Subjects are often a select group, i.e., college students who tend to be at the upper end of the reading achievement scales for the general population. College students do, however, exhibit a wide range of intellectual and academic skills and cannot be considered to be a matched subject group for studies that rely heavily on one specific skill. The use of a word recognition test such as the WRAT would allow the experimenter to screen subjects subsequently making comparisons between subject populations identified by differences in reading achievement levels and, by doing so, avoid what may prove to be a confounding variable in recognition studies that use letter strings as stimulus materials.

Since it is possible that more skilled readers process letter strings in units as suggested by Jackson (1980), Healy (1981), and Gibson (1969), while the less skilled reader processes letter strings sequentially, there may very well be differences in performance between the two groups. Most recognition studies allow only a very brief time for both learning trials and recognition trials. The reader who cannot process a particular item as a unit simply does not have time to enter that item into the memory code as a meaningful bit of information. In the recognition task, therefore, there is little information present in the memory code that is of value to the subject in making the Old or New decision. This would increase both the subject's tendency to make errors and his tendency to guess.

This study, therefore, concentrates on the effects of a possible subject variable, i.e., the ability of the subject to recognize words out of the normal context as measured by the reading subtest of the WRAT. Using a signal detection paradigm (Green and Swets, 1966; Shontz, 1967), data trials were conducted utilizing CVC trigrams of different meaningfulness values in the four possible combinations of high and low meaningfulness.

Previous studies of recognition performance which involved varying the meaningfulness of CVC trigrams as the independent variable

(Raser, 1970, and Thompson-Gibbs, 1972) have yielded consistent patterns of recognition performance. It is expected that the following patterns would again occur.

1. Recognition performance in trials where both the Old and New CVCs were high in meaningfulness would reflect the highest degree of sensitivity to the various stimulus features.
2. Recognition performance for those trials where the Old CVCs were high in meaningfulness and the New CVCs low in meaningfulness would reflect sensitivity of a somewhat lower level.
3. Recognition performance for trials where the Old CVCs were low in meaningfulness and the New CVCs high in meaningfulness would reflect an even lower level of sensitivity to the features of the stimulus materials.
4. In the trials where both the Old and New stimulus materials were low in meaningfulness, the recognition performance would reflect a very low level of sensitivity.

Differences in recognition performance occur not only in response to the meaningfulness levels of the stimulus materials, but also appear to be related to the contrast effects evident in the two conditions where Old and New stimuli are of different values.

The major hypotheses of this study are: 1) Previous findings will continue to hold true regarding recognition performance when meaningfulness is varied across conditions; and 2) subjects in the highest reading achievement group will display consistently better recognition performance under all four stimulus conditions. The Low reading achievement group will display significantly lower measures of performance.

METHOD

Subjects

Eighty subjects were selected from an original pool of 120 students enrolled in psychology courses at Montana State University. Subjects were selected on the basis of their score on Jastak, Bijou, and Jastak's (1965) Wide Range Achievement Test. Administration of the WRAT was the first task presented in the experimental session. Subjects were assigned to two achievement groups designated as High WRAT and Low WRAT. The forty subjects with the highest WRAT scores were designated the High group, while the forty subjects with the lowest scores were designated as the Low group. The only experimental requirement was that subjects had normal visual acuity with or without corrective lenses.

Each of the original 120 subjects was scheduled for a thirty minute time block, and participated in the data collection trials whether or not he/she was subsequently assigned to one of the two experimental groups.

Stimulus Materials

CVC trigrams used as stimulus materials were identified by selecting 200 High and 200 Low meaningfulness CVC trigrams from Noble's (1961) scaled lists. The CVCs were chosen to exclude English words and known abbreviations.

The CVCs were prepared for presentation by typing the individual trigrams on acetate film, then mounting the film in standard 2X2" slide mounts.

Eight trays of slides were prepared for the experimental task using the four possible pairings of meaningfulness. The four conditions were designated as High-High, High-Low, Low-High, and Low-Low. The first value, in each case, refers to the meaningfulness level of the Old stimulus, and the second to the New stimulus.

Two Carousel trays were assembled for each condition; each contained 25 Old CVCs in random order, then 25 New CVCs interspersed randomly among repeats of the 25 Old CVCs for a total of 50 stimulus slides for the recognition task. Each subject was given two series of recognition trials, both with the same meaningfulness condition for a total of 100 trials per subject.

Old CVCs were defined as those CVCs presented in learning trials; the New CVCs were those added to the list of Old CVCs and presented during recognition trials. No trigram was used more than one time in any condition. As shown in Table 1, the meaningfulness values assigned by Noble (1961) for each CVC were summed for Old and New CVCs in each tray and mean meaningfulness values computed for each condition.

Table 1

Mean Meaningfulness Value for Each CVC Condition by Tray

Tray	Condition	Mean Old	Mean New
1	High-High	3.41	3.32
2	High-High	3.53	3.49
1	High-Low	3.36	1.27
2	High-Low	3.29	1.26
1	Low-High	1.27	3.39
2	Low-High	1.26	3.42
1	Low-Low	1.25	1.27
2	Low-Low	1.27	1.26

Total possible range: 1.06 to 4.78.

One tray was used in the practice trials. This tray was assembled using 50 CVCs from the middle range of Noble's scaled lists. No CVC from the practice set was used in any of the eight sets of experimental materials.

Apparatus

Stimulus materials were presented on a rear projection screen (12x15 inches) via dual Carousel projectors equipped with automatic shutters and a timing device known as the Lafayette T-2K constant illumination tachistoscope. The first project projected the CVC on the screen while the second provided comparable interval illumination to the screen to minimize afterimage effects. The screen was placed at eye level with the subject seated approximately 36 inches from the screen.

Illumination of the experimental room during learning and recognition trials consisted of low level lighting sufficient only to allow the experimenter to record the data.

Procedure

The word list from the Reading portion of the Wide Range Achievement Test, Level II was administered individually to each of the 120 volunteer subjects. The 40 highest scoring subjects were assigned to the High WRAT group; the 40 lowest to the Low WRAT group. The scores and data trials from the 40 subjects in the middle range were discarded and not included in any experimental analysis.

The mean score on the Wide Range Achievement Test for the High WRAT group was 15.6, with a range of 13.2 to 17.1, whereas the mean for the Low WRAT group was 11.2 with a range of 7.7 to 12.2. The scores expressed above represent the academic grade level of subjects' recognition reading skills.

Each of the subjects also participated in the recognition portion of the study. To facilitate counterbalancing the order of presentation of stimulus materials, a running mean was kept for the WRAT scores after the first 20 subjects had completed data trials. Subjects that followed were arbitrarily assigned to High and Low WRAT groups depending on the score being above or below the then current mean WRAT score.

A signal detection paradigm was chosen as the most effective way of obtaining measures that were sensitive to differences in performance between the two reading achievement groups. This paradigm, proposed by Green and Swets (1966), is formulated on a theory that assumes psychophysical judgments are based on underlying neural activity which can vary in magnitude. This activity can be produced by an external stimulus (the signal) but can also occur when no signal (or a false signal) is present. Gleitman (1981) provides a most simple explanation of the subject's task in the detection experiment:

> ". . . the subject's task in a detection experiment . . . . He must decide whether a given sensory process should be attributed to the signal superimposed on background noise or to the background noise along . . . . On the average, the signal (plus background noise) produced a process of greater magnitude than does the noise alone, an average value which rises with increasing signal strength. The subject's decision problem arises because the magnitude of the two sets of sensory processes fluctuates. . . . the more intense the stimulus, the less likely it is that such confusion will occur.
>
> . . . . The best they can do is decide that they have a sensation--that their internal sensory experience is produced by a signal rather than by noise alone. And in this decision they can be wrong."

Following a technique discussed by McNicol (1972), a rating scale response was utilized rather than the standard Yes or No answer. The use of such a scale provides the most efficient means of obtaining information regarding the sensitivity of the subject to differences in the stimulus materials.

Prior to the presentation of learning and recognition trials, each subject was handed a copy of the instructions. The experimenter went over these instructions verbally while the subject read his/her copy of the instructions. After the subject had an opportunity to become familiar and comfortable with the rating scale, the experimenter presented the two blocks of practice trials using the tray of medium value CVC trigrams. The first of these trials consisted of ten CVCs in the learning trial and twenty in the recognition trial. The second practice trials consisted of fifteen and thirty CVCs, respectively. The subject was encouraged to ask questions regarding the use of the rating scale and verbally reinforced for using the scale efficiently.

The instructions for the experimental task (Appendix A) were modified from a previous study (Thompson-Gibbs, 1972).

The stimulus conditions for the data trials were then presented via the Lafayette projection system. The 25 CVCs in the learning trials (Old CVCs) were presented sequentially for 1/125th second at one second intervals. After a fifteen second interval between learning and recognition tasks, the fifty slides for the recognition task were presented for 1/125th second at four second intervals. During the fifteen second interval between learning and recognition trials the subject was instructed to count backwards from 100 by 3s out loud to minimize the rehearsal of the Old CVCs.

The four stimulus conditions were counterbalanced for order of presentation within the stimulus conditions as well as for order of stimulus conditions. This counterbalancing occurred for both reading groups.

During presentation of the recognition stimuli, the subject's task was to respond to each CVC shown with a confidence rating of one through six, with responses of one, two, and three indicating an Old CVC (one that had been seen in the learning trials) with confidence levels of Very Sure, Fairly Sure, and Guess (unsure), respectively. Responses of four, five, and six indicated that the CVC was New (not seen in the learning trials) and with respective confidence levels, of Guess, Fairly Sure, and Very Sure. It should be noted that a response of three or four is different only in terms of the Old or New portion of the response.

At the end of each experimental presentation, each subject was asked to complete a post-experimental questionnaire (Appendix B).

RESULTS AND DISCUSSION

Measures of Sensitivity and Response Bias

In the process of tallying responses for each of the two stimulus conditions (Old or New CVC) the raw data from each subject was organized in a two-by-six matrix. Data analysis followed procedures designed to illustrate any differences in sensitivity and response bias across the four meaningfulness conditions and between the two reading achievement groups. Such an analysis yields an in-depth measure of subject performance that is not possible when merely comparing correct response rate.

Sensitivity in the present study is defined in terms of recognition performance. In the language of the signal detection paradigm, presenting an Old CVC, i.e., one that occurred during learning trials, constitutes the "Signal" trial. A New CVC (one that has not been presented before), constitutes a "Noise" trial. The more Old CVCs that are recognized, the more sensitive the observer is to qualities unique to the stimulus materials. A primary factor affecting the sensitivity of the subject to CVCs varying in meaningfulness is presumed to be the number of associations that can be made to the CVC (Noble, 1961). Also related to sensitivity is the ability of the subject to process all of the components of the CVC in such a manner that the meaningfulness factor can operate.

An additional factor was suggested by subjects who participated in Thompson-Gibbs 1972 study. This is a "contract effect" that occurs when Old and New CVCs differ in meaningfulness values. Old CVCs that are high in meaningfulness will appear different to the subject when they occur along the New CVCs that are low in meaningfulness and vice versa. This difference is in the content of the letter string, since CVCs high in meaningfulness often follow patterns found in common words, while low meaningfulness CVCs often appear to have no relationship to the letter strings found in words.

Response bias, on the other hand, is the manner in which the subject uses the rating scale. Subjects in the recognition study responded using a six-point rating scale. Differences in the meaningfulness combinations of the four stimulus conditions may result in subjects using the scales differently. Under one experimental condition subjects may spread out their responses to Signal trials more than to Noise trials or vice versa. The mean category response may shift more for Signal trials than for Noise trials under certain experimental conditions. In this study, this is what is meant by response bias. Subjects may use the scale differently under the four conditions in a systematic way. Measurement of such systematic shifts may provide additional information on how meaningfulness or reading ability influences subject performance in a Short-Term Memory task such as that used in the present study.

It must be pointed out that this definition of response bias does not refer to the type of response set individuals display when responding on a rating scale (Guilford, 1954). In response sets, a tendency to use extremes of the scale, the center of the scale, or a normal distribution of responses is determined by the subjects personality or influenced by the instructions. Sensitivity and response bias measures in this study are not sensitive to this type of set.

Three measures of sensitivity and response bias were used in analyzing the performance of the two reading achievement groups. The measure of sensitivity is the area under the Receiver Operating Characteristic (ROC) curve (P_A), as defined by Green and Swets (1966) and McNicol (1972). This generally accepted measure provides an analysis sensitive to the effects of meaningfulness, association values related to meaningfulness, and other distinctive features of the CVC trigrams used as stimulus materials. This measure should also be sensitive to any processing differences related to the subject variable.

The measure of response bias used was a modification of the Chi-square Goodness of Fit Test. This test measures response bias as defined earlier. Shontz (1967) discusses this measure in his study of factors affecting the processing of sequentially presented form parts. He found this measure to be sensitive to systematic changes in the use of the rating scale as a function of the experimental variables.

A third measure, an analysis of the Mean Rating Category Values, should provide information regarding both sensitivity and response bias. Shontz (1967) has established that this measure is useful in the interpretation of information regarding response bias data.

Area Under the ROC Curve - $P_{(A)}$

Plotting the probabilities of correct recognitions (Hit Rates) versus "false" recognitions of New stimulus materials (False Alarms) from the rating scale data on the unit square yields the Receiver Operating Chracteristic Curve discussed by Green and Swets (1966) and McNicol (1972). As discussed earlier, the area under the ROC curve has become a generally accepted measure of sensitivity as it is independent of changes in response bias. This non-parametric measure is useful when assumptions regarding the parametric measures of sensitivity (d') and response bias (β) cannot be met. Plots of the probability data from this experiment on double probability paper indicated that the assumptions for parametric measure had not been met. The area measure ($P_{(A)}$) has been shown to be comparable to d' (McNicol, 1972) and was chosen as the best measure of sensitivity.

Area values were computed for each subject. These values were used as data points for a two-way ANOVA for the area under the ROC curve as a measure of sensitivity. The two-way ANOVA Table is illustrated in Table 2. There were significant differences, $F(1, 72) = 24.20$, $p < .001$, in sensitivity as a function of meaningfulness conditions indicating that all subject groups were in some way cognizant of, and using, the features of the stimuli in reaching decisions. The mean area values for the High WRAT group were: High-High, .796; High-Low, .895; Low-High, .773; and Low-Low, .612. Mean area values for the Low WRAT group were .766; .344; .737; and

.737, respectively. There were no significant differences in the sensitivity of the two reading achievement groups, $F\ (1, 72) = 1.23$, $p > .05$.

Table 2

Analysis of Variance for Area Under the ROC Curve As a Measure of Sensitivity

Source	SS	DF	MS	F
Meaningfulness (R)	0.622	3	0.207	24.200***
Reading Level (B)	0.011	1	0.011	1.232
R B Interaction	0.010	3	0.003	0.396
Individuals	0.617	72	0.009	---
Total	1.260	79	---	---

***Significant beyond the .001 level.

The relative differences in sensitivity between stimulus conditions and the two reading achievement groups are illustrated in Figure 1. It should be noted that it is generally not appropriate to combine raw data from individuals as illustrated in Figure 1 as large differences in the slopes of the individual curves exist (McNicol, 1972). It is included here for illustration purposes only.

In previous studies (Raser, 1970; and Thompson-Gibbs, 1972), distinct differences in the pattern of recognition performance were found when meaningfulness was varied as a stimulus variable. Superior recognition performance occurred when both Old and New CVCs were high in meaningfulness. In this study, subjects appear to have utilized

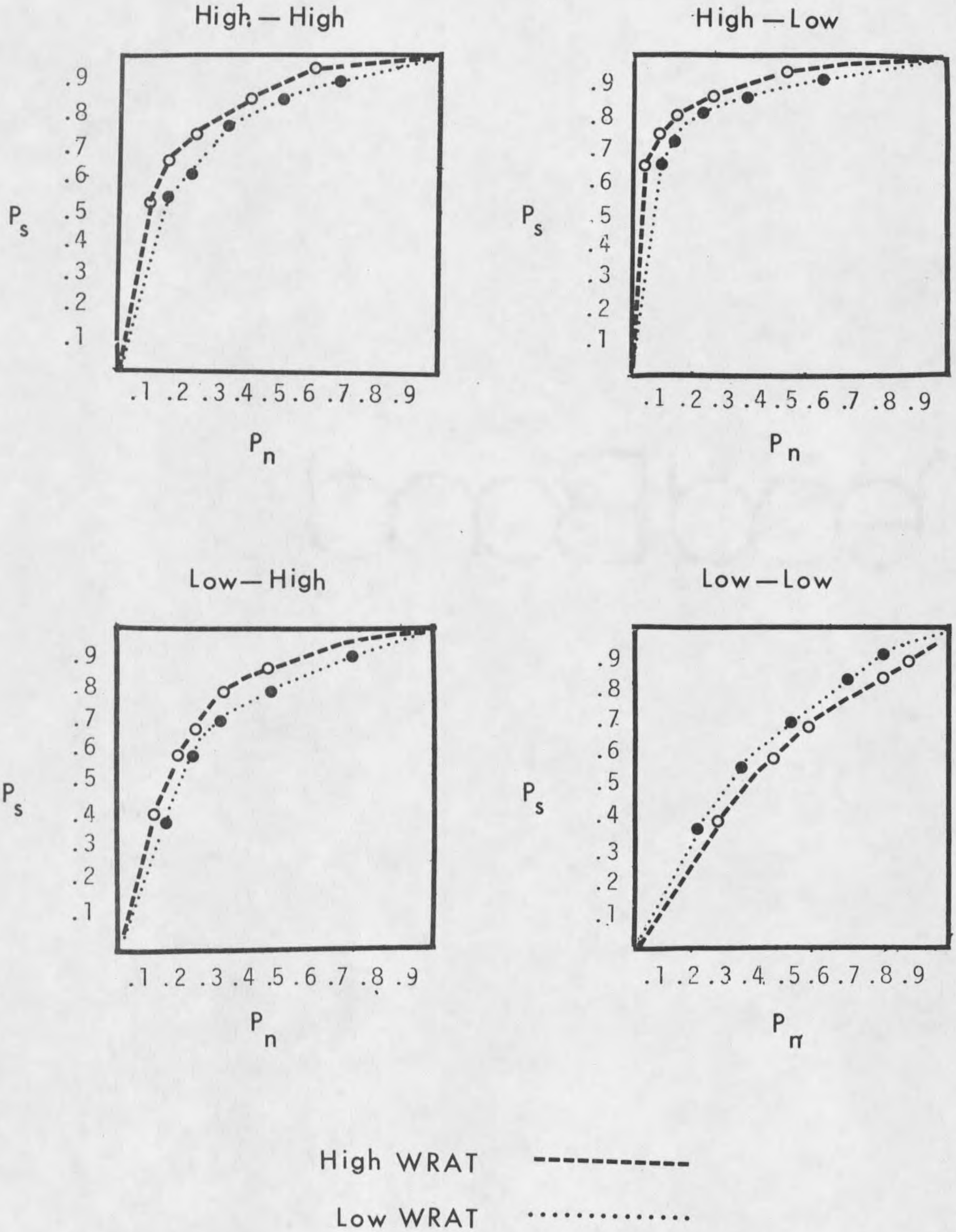

Figure 1

ROC Curves for Meaningfulness Conditions contrast effect in addition to the stimulus properties related to meaningfulness. Recognition performance was superior for both groups when the Old CVCs were high in meaningfulness and New CVCs were low. Inspection of the representations of the ROC curves for each meaningfulness condition clearly illustrates the decline of sensitivity in the following manner:

1. Subjects were most sensitive to the High-Low condition. Not only were subjects sensitive to the properties of the stimulus CVCs that provided cues for later identification, but they seemed to be able to pick up the contrast in the two levels of meaningfulness.
2. The second level of sensitivity was to the High-High condition. In this case, subjects were sensitive to the similar properties of both Old and New CVCs, apparently making comparisons with accessible representations in the memory codes.
3. The third level of sensitivity was to the Low-High condition. Under this condition, the Old CVCs allow for few associations that might aid the observer in his decision process. The fact that the New CVCs are high in meaningfulness and are therefore more wordlike in nature, probably provides a contrast effect that offers the major cue for the subject's decision process.
4. As expected, the lowest level of sensitivity was present in the Low-Low condition. There is little information contained in the stimulus that allow for the formation of

associations with other information nor is there any contrast information available.

An inspection of Figure 1 suggests that there were slight differences in recognition performance between the two reading achievement groups across all four meaningfulness conditions. While not statistically significant, $F\ (1,\ 72) = 1.232$, $p > .05$, these differences in sensitivity may be related to the manner in which subjects process the stimulus materials. Jackson (1980) suggests the less skilled reader is more likely to process the individual components of the stimulus serially rather than as a unit. What is not clear is at what level the degree of reading achievement (related to chronological age, academic level, and reading ability), affects the sensitivity of the subject to the properties of the stimulus.

The possibility should be considered that only those subjects who were reading at the lower end of the range were processing the components of the stimulus differently from other subjects. This would create the differences in sensitivity that are evident in the pictorial representation of the ROC curves. If only those subjects at the lower end of the range of WRAT scores are performing differentially, however, there would not be sufficient differences between the groups to be statistically significant. Additional studies separating subjects into groups differing in degree of reading ability as outlined above would be necessary to determine where differences in sensitivity, if any, would exist.

Chi Square Values as a Measure of Response Bias

Chi-square values can be used in testing the Goodness-of-Fit of a theoretical curve to an observed frequency distribution (McNemar, 1969). The Chi-square Goodness-of-Fit technique enables us to test whether the frequencies of an experimental distribution depart from those of a theoretical curve sufficiently to be regarded as chance departures.

A modification of the Chi-square Goodness-of-Fit test as used by Shontz (1967) was employed to assess the systematic variations in response bias as a function of meaningfulness and reading achievement level. Since the assumptions for the parametric measure (β) could not be met, this non-parametric measure was chosen as the most suitable means of evaluating response bias because it can provide data that reflect differences in the use of rating scales across conditions.

The frequency data were organized as shown in Figure 2. In this case, the ordering of confidence levels is reversed for the New CVC presentations. This new matrix pairs those cells that are equivalent in confidence and accuracy. As a result, as sensitivity changes, the two frequency distributions can be expected to change concurrently. The shape of the two distributions, when no response bias is present, will be similar, and the Chi-square value will be small. If, however, there are systematic shifts in response bias across conditions, the Chi-square values will become larger as the two frequency distributions differ systematically. As this measure is insensitive to both sensitivity and variation bias, it serves as an independent measure of response bias.

STIMULUS

RESPONSE CATEGORIES	OLD	NEW	
1	27	22	6
2	10	17	5
3	0	3	4
4	1	3	3
5	8	3	2
6	4	2	1

Figure 2

Sample Frequency Data as Organized for Computation of Chi-Square Measure of Response Bias

Response bias, as discussed here, is not the type of response set characterized by a subject's consistent idiosyncratic use of the rating scale. This measure is sensitive, however, to changes in the use of the scale due to experimental conditions. As an example, there may be changes in the distribution of responses on Signal trials (Old CVCs) without concomitant changes in Noise trial distributions. If sensitivity changes across conditions but the use of the rating scale remains constant for Signal and Noise trials, the responses will be distributed in similar patterns, and the resultant Chi-square Goodness-of-Fit test value will be small. If response distributions are different for Signal and Noise conditions the resultant Chi-square value will be large. Differences in Chi-square values across the different meaningfulness and subject conditions provide the measure of response bias.

Chi-square values were calculated for each subject. These values were then transformed using $\sqrt{X} + \sqrt{X+1}$ to stabilize the variance of distribution of the Chi-square values (Snednecor and Cochran, 1980). The transformed values were analyzed using the same ANOVA model as was used on the analysis of the area under the ROC curve. It must be pointed out that these values are used only as data points in the ANOVA and have no other significance in the overall analysis of results.

The results of the analysis are shown in Table 3. The nonsignificant $F\ (3, 72) = 2.059$, $p > .05$, indicates there were no significant changes in the use of the rating scale as a function of meaningfulness. There were, however, obvious systematic differences in the use of the rating scale as a function of the subject variable. The significant difference $F\ (1, 72) = 4.495$, $P < .05$ found in reading achievement level indicates that there were systematic differences in the way the two reading achievement groups used the rating scale. Means for the High WRAT group were: High-High, 3.59; High-Low, 3.36; Low-High, 3.4; and Low-Low, 4.39. Means for the Low WRAT group were 4.34, 4.49, 3.64, and 4.62, respectively. This measure, however, while providing evidence that there were differences in response bias as a function of reading achievement level, does not provide clues to the manner in which they differ. These differences are more clearly illustrated in the discussion of the Mean Rating Category Values that follows.

Table 3

Analysis of Variance for Chi Square (2x6) as a Measure of Response Bias

Source	SS	DF	MS	F
Meaningfulness (R)	9.783	3	3.261	2.059
Reading Level (B)	7.102	1	7.120	4.495*
R/B Interaction	3.006	3	1.002	0.633
Individuals	114.032	72	1.584	--
Total	113.942	79	--	--

*$p < .05$

Mean Rating Category Values

The analysis of the differences in mean confidence levels for each condition provides an information measure for both response bias and sensitivity. By adding the response condition (Old or New CVC) as additional variables in the ANOVA, it is possible to obtain an analysis sensitive to evidence that mean confidence levels are affected by stimulus conditions.

As sensitivity changes, the mean value of the rating category should decrease for the Old response category, and conversely, increase for the New response category. To provide a measure where the two mean values are positively related to one another, and to avoid negative values in the ANOVA, the value of the New mean rating is subtracted from seven (n rating categories plus one). As senstivity to stimulus conditions changes, therefore, the two means would now tend to shift upward or downward simultaneously. The difference between mean rating category values for Old and New stimuli is the index of sensitivity. Differing distributions of response within each of these means across stimulus conditions indicate changes in response bias. Figure 3 illustrates these shifts in value and provides some evidence of the manner in which subject groups differ.

To more clearly illustrate the differences between the subject groups, the transformations of the Old response category data were not used when plotting the data for Figure 3. The data points reflect the original mean category rating values for Old CVCs.

A final type analysis of variance was conducted to provide information sensitive to interactions between reading achievement groups, stimulus conditions (meaningfulness) and mean confidence ratings for each response condition (Old or New). The four-way ANOVA (r(BC)D) provides a measure of the differences in sensitivity and response bias that may exist between the two reading achievement groups, and any interaction that may be present between the mean confidence ratings, stimulus conditions, and reading achievement groups.

This measure is redundant with the two measures previously reported. As indicated in Table 4, this ANOVA does provide additional information regarding the differences in performance between the two reading achievement groups. Of particular interest is the significant interaction between meaningfulness conditions and response category ratings, $F\ (3, 72) = 3.350$, $\underline{p}<.02$, indicating that there were

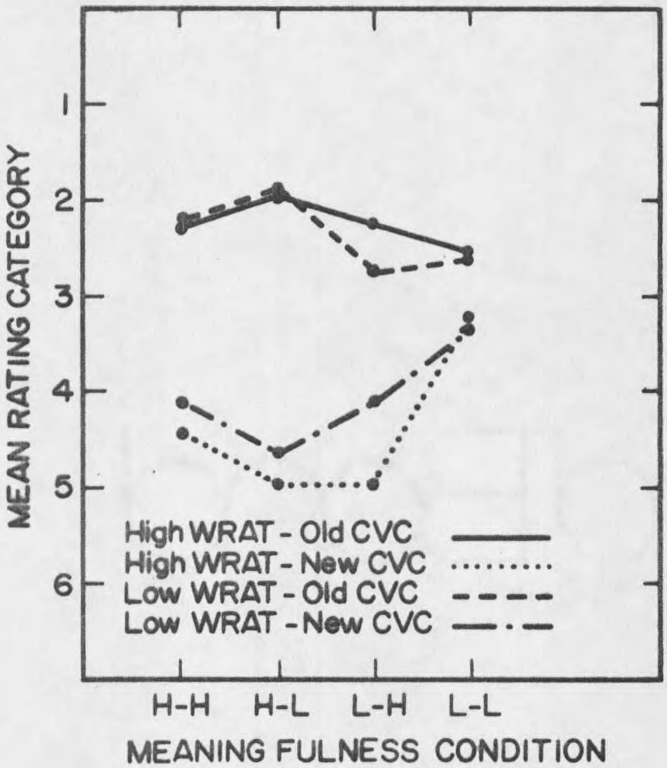

Figure 3

Mean Response Categories for Old and New CVC Conditions

When the response was 1, 2, or 3, the subject believed the CVC was Old; when responding 4, 5, and 6, the subject believed the CVC was New.