Memory & Cognition
1981, Vol. 9 (3), 277-282

Age of acquisition, imagery, recall, and the limitations of multiple-regression analysis

PETER E. MORRIS
University of Lancaster, Bailrigg, Lancaster LA1 4YF, England

The first part of the paper describes the limitations of the stepwise multiple-regression technique used by Gilhooly and Gilhooly (1979) to assess the roles of age of acquisition and other variables in memory. The high intercorrelations between the variables make the technique inappropriate for the identification of those variables that influence recall and recognition. An experiment in which age of acquisition and imagery were manipulated is then reported. Better free recall occurred for late-acquired words, and it is suggested that this resulted from their greater distinctiveness and saliency.

Gilhooly and Gilhooly (1979) report four experiments in which multiple-regression analysis was used to assess the effects of word age of acquisition and other potentially relevant variables upon a range of verbal tasks. The present paper is in two parts. The first discusses the limitations of multiple-regression analysis in the study of those properties of words that determine their memorability. The second describes an experiment in which the influence of age of acquisition and imageability upon free recall were examined.

THE LIMITATIONS OF MULTIPLE-REGRESSION ANALYSIS

Gilhooly and Gilhooly (1979) were faced by the perennial problem of anyone interested in the relationship between the properties of words and their recall, namely, that a large number of variables have been found to be related to recall and these variables are themselves intercorrelated, often to a high degree. Meaningfulness (Noble, 1952), frequency (Gregg, 1976), familiarity (Underwood & Schulz, 1960), imagery (Paivio, 1971), and concreteness (Richardson, 1975) are all known to covary with recall and/or recognition. The intercorrelations of many of these variables are high; for example, imagery correlates .81 with concreteness and .52 with meaningfulness and familiarity correlates .58 with frequency and .45 with meaningfulness (Gilhooly & Hay, 1977). If age of acquisition is introduced as a new potential variable, many more intercorrelations occur, since age of acquisition correlates −.66 with familiarity, −.60 with imagery, −.58 with frequency, −.45 with meaningfulness, and −.43 with concreteness (Gilhooly & Hay, 1977). It is not surprising that, given this extent of intercorrelation, it is very difficult to factorially manipulate any particular variable, especially age of acquisition, and hold constant the remainder. Gilhooly and his associates are themselves undertaking the arduous task of collecting data on a very large pool of words, which may eventually make possible the orthogonal manipulation of the variables, but in Gilhooly and Gilhooly (1979), they adopted a stepwise multiple-regression analysis to investigate the effects of the variables mentioned above, and some others, upon memory.

The stepwise multiple-regression technique involves first obtaining a correlation matrix for the criterion variable (e.g., recall) and the predictors (e.g., frequency, imagery) and then carrying out successive multiple-regression analyses by beginning with the predictor that correlates most highly with the criterion and adding in the other predictors, one by one, in the order of their respective correlations with the criterion. The influence of each new variable is assessed by the improvement in the multiple-regression coefficient R.

Multiple regression is a very powerful and flexible statistical tool. Since analyses of variance and covariance can be considered as merely special cases of multiple regression (Cohen & Cohen, 1975), it is perhaps surprising that multiple regression has not been more widely used by psychologists. However, all statistical techniques have limitations in the information with which they can provide the investigator, and unfortunately, the stepwise multiple-regression technique is no exception. If the object is to identify the appropriate set of variables that needs to be considered to obtain a satisfactory prediction of the criterion performance, then, given the qualifications relevant to all correlational techniques (e.g., a linear relationship between predictor and criterion), this form of analysis is very useful. We can conclude, for example, from Gilhooly and Gilhooly's (1979) Experiment 4 that the prediction of recognition memory will not improve by taking other variables into account if frequency, age of acquisition, and concreteness have been controlled. The problem is that most research in the area is carried out to identify those variables that influence memory, not to predict the recognition or recall of particular word lists. It is rare for pragmatic prediction of the probability of recall of lists of random nouns to be required. What is needed

0090-502X/81/030277-06$00.85/0

is an understanding of the dimensions upon which encoding takes place, the circumstances that determine the relative importance of these dimensions, and their interaction with subject strategies. As a technique for identifying which variables influence memory, stepwise multiple-regression analysis has considerable limitations and dangers that are not mentioned by Gilhooly and Gilhooly (1979) and which may mislead some of their readers.

The crucial problem is that when two predictors are correlated both with each other and with the criterion, the shared variance will be included in the multiple-regression analysis when the first of the predictors is introduced. The order of the introduction of the predictors into the stepwise analysis is therefore vital insofar as it determines the apparent influence of a given variable. To take a specific example from Gilhooly and Gilhooly's (1979) Experiment 4, frequency correlated with recognition memory −.45 and with age of acquisition .41, whereas frequency and age intercorrelate by −.68. By beginning the stepwise multiple regression with frequency and following with age, Gilhooly and Gilhooly concluded that frequency accounts for 21% of the criterion variance and age only 2%. If they had begun with age of acquisition, age would have apparently accounted for 17% of the variance and frequency only 5%. Clearly, very different interpretations might be drawn from these two analyses. Gilhooly and Gilhooly (1979, p. 222) concluded that no effect was found for age of acquisition in Experiment 4. This conclusion is valid only if there is good reason for considering frequency to be the primary variable and any influence of age of acquisition to be the result of the latter's covariation. If there were good theoretical or empirical reasons for the order of inclusion of predictors into a stepwise multiple-regression analysis, then there would be less risk of misinterpretation. However, this was not so for Gilhooly and Gilhooly's (1979) analyses, nor is it likely to be until better theories about the nature of the encoding processes are available. Given the lack of theory to guide the order of inclusion of predictors, Gilhooly and Gilhooly followed the logical procedure of including the predictors in the order of their correlation with the criterion. Unfortunately, there are several reasons why this may lead to false conclusions.

The first problem is the reliability of the correlations. For their enormous free recall experiment using very large numbers of subjects and words, Christian, Bickley, Tarka, and Clayton (1978) obtained a reliability of their recall measure of only .57 using the Spearman-Brown formula. Recall, and probably recognition also, are not highly reliable measures. When correlations between predictors and recall or recognition are found to be similar, it is well within the range of experimental error for the rank of these correlations to be changed in a subsequent experiment. Given the unreliability of the criterion measure, it would not be surprising if a replication of Gilhooly and Gilhooly's (1979) Experiment 4 found a slightly higher correlation between age and recognition, shifting from .41 to, say, .44, whereas the frequency correlation might change marginally from −.45 to, say, −.43. Then, the conclusion from a stepwise multiple regression would be that age of acquisition was the fundamental variable and that frequency made little or no contribution.

The second problem is that the correlations between predictors and the criterion are influenced by the intercorrelations with the other predictors and by those predictors with the criterion. Suppose, for example, that the analysis includes a variable that is perfectly correlated with the criterion, when other variables are controlled. This variable, which is perhaps one of the causes of the criterion performance, may have a very low correlation with the criterion in the initial correlation matrix if it happens to covary negatively with another variable or variables that are also positively related to the criterion, or if it correlates with a variable or variables that happen to be negatively related to the criterion. To illustrate this, if the variable perfectly correlated with the criterion "c" is represented by "a" and the covarying variable by "b," then the partial correlation formula is

$$r_{ac.b} = \frac{r_{ac} - (r_{ab}.r_{cb})}{[(1 - r_{ab}^2)(1 - r_{cb}^2)]^{1/2}}. \quad (1)$$

We have defined $r_{ac.b} = 1$. To simplify the illustration, consider the particular case in which $r_{ab} = -r_{cb.} = x$. Formula 1 then becomes

$$1 = \frac{r_{ac} - [x(-x)]}{1 - x^2}, \quad (2)$$

which becomes

$$r_{ac} = 1 - 2x^2. \quad (3)$$

Clearly, the observed correlation between Variable a and the criterion will decrease as x increases. It would be zero when x = .707. Thus, the observed correlation may give a false impression of a smaller relationship between the criterion and the predictor if a mixture of positive and negative intercorrelations occurs. This is the rule rather than the exception in the study of word properties and memory, and the danger is obvious. A lowering of the correlation of a variable and the criterion in the initial correlation matrix lessens the variable's likelihood of being selected early in the stepwise multiple-regression analysis and decreases the probability of importance being ascribed to it, given the absorption of variance by other covarying predictors included. Even

if selected early, the variable will be ascribed less than its true influence because the apparent relationship with the criterion will be suppressed by the counter-effect of the negatively influencing covariable, and the variance accounted for, as indicated by the observed correlation with the criterion, will underestimate, perhaps considerably, the true relationship of the variable and the criterion.

The third problem is the converse of the preceding one. The observed correlation between a predictor and the criterion will be influenced if the predictor correlates positively with another predictor or predictors that are positively related to the criterion. A negative correlation is, of course, similarly inflated by other negative correlations. The more variables with which the predictor correlates and which are themselves correlated with the criterion, the higher will be the observed correlation, even though the true relationship may be small or zero. Even if Variable X has a zero correlation with the criterion when other variables are controlled, if the variable correlates by, say, .67 with another variable that also correlates .67 with the criterion, then the observed correlation for Variable X and the criterion will be .44. If Variable X correlates similarly with another variable that correlates .67 with the criterion, the observed correlation for Variable X will rise to .53, and so on. With some underlying relationship between Variable X and the criterion, the observed correlations are, of course, higher still. Here the danger is that, because of its intercorrelations with other variables, a predictor may be falsely promoted in the rank order of the variables, with the subsequent unwarranted attribution of influence on the criterion. Not only will the variable be promoted above others in the order of introduction into the stepwise analysis, but the amount of the variance in the criterion for which it accounts will also be overestimated. Given the mass of high intercorrelations in any study such as that of Gilhooly and Gilhooly (1979), it is virtually impossible to determine how the observed correlations are being inflated or deflated.

The main danger of the highly intercorrelated nature of the variables that may influence recall and recognition is that the wrong variables may be identified as those likely to be causally related to performance and the true determinants may be made to appear irrelevant. There may also be problems in interpreting the direction of any relationship apparently identified. The following particular example was chosen because it is relevant to the experiment described below.

In Experiment 3, Gilhooly and Gilhooly (1979) report a negative relationship between age of acquisition and recall (−26) in the initial correlation matrix. Those word variables that do appear to be positively related to recall are imagery (.42), frequency (.33), and concreteness (.31). All three of these variables are negatively correlated with age of acquisition (−.59, −.48, and −.45, respectively). The partial correlation of age of acquisition and recall with imagery and frequency controlled ceases to be negative and becomes a small positive correlation of .10, which might increase slightly if concreteness was also controlled. What is important is that there is nothing in the stepwise multiple-regression analysis that would identify such a change in the direction of the correlation.

Finally, although not a criticism of the multiple-regression technique, it is worth pointing out that it will be most sensitive if the predictors have a linear relationship with recall. However, it is quite possible that some of the variables that will influence word recall will be important only when they are at their strongest. So, for example, emotionality may be a largely irrelevant variable except for words of highly emotional connotations, or imagery may be important only for highly imageable items when either spontaneous images may occur or the nature of the material suggests an imagery strategy to the subject. In such cases, the random sampling undertaken for the multiple-regression analysis will include few items from extremes on the variables, and the variables may be dismissed. Of course, factorial manipulation of high and low sets from the variables runs the danger of overemphasizing the general importance of such variables.

To summarize, the high intercorrelations between variables make initially attractive any technique that does not require experimental control of the variables, but it is the high intercorrelation between the variables that makes it virtually impossible to identify which are really contributing to memory performance.

It is important to emphasize that the criticisms above are directed to a particular application of multiple-regression analysis. Multiple regression will often allow a more powerful analysis of data; for example, it is preferable to analysis of variance when the latter blocks items together, losing the within-block variance, which is retained in multiple regression. The wide range of applications of multiple-regression analyses are well illustrated by Cohen and Cohen (1975). However, even multiple regression does not provide a satisfactory method for unpacking the relative contributions of intercorrelated variables to performance.

AGE OF ACQUISITION, IMAGERY, AND RECALL

The previous section has emphasized the problem of experimentally controlling word properties in order to examine the relationship between age of acquisition and recall. However, using the nouns made available by Gilhooly and Hay (1977) for 205 five-letter words, it is possible to control the more potent known covariables and still manipulate age of acquisition. In the present study, two lists were prepared. The first consisted of items either early or late in age of acquisition, and imagery, frequency, and meaningfulness were controlled. In the second, imagery was manipulated and frequency and age of acquisition were controlled. Meaningfulness could not be controlled in the latter list, but Christian

et al. (1978) found no relationship between meaningfulness and free recall when other variables were controlled. Gilhooly and Gilhooly (1980) have demonstrated the validity of the Gilhooly and Hay (1977) ratings of age of acquisition.

At the time that the research was carried out, prior to the publication of Gilhooly and Gilhooly's (1979) study, it was expected that if age of acquisition was related to free recall, words with an earlier age of acquisition would be better recalled. There was little theoretical justification for this prediction, but when the latency of picture naming was measured (Carroll & White, 1973), earlier acquired words were apparently retrieved more quickly from semantic memory. Also, Paivio (1976) reported that age of acquisition correlated negatively with recall. However, Paivio's study involved no attempt to control for the influence of confounding variables such as frequency and imagery. It is not difficult to suggest reasons why later acquired words may be easier to learn. They may, for example, be associated with more specific situations, and they may be associated with more salient factors in the life of the undergraduate subject. The experiment was therefore exploratory.

Because age of acquisition correlates highly with imageability, it was possible that age of acquisition could account for the known relationship between imageability and recall. Hence the second condition was included, in which imagery was manipulated while age of acquisition was controlled.

Method

Subjects. Thirty 1st-year undergraduate students at Lancaster University were tested in two groups prior to a practical class.

Materials. Two lists were prepared. In the first (the A list), 12 pairs of words were chosen from the Gilhooly and Hay (1977) norms for 205 five-letter nouns. The pairs varied in their age of acquisition, but they were matched for frequency (Thorndike & Lorge, 1944) and imagery. Meaningfulness was also controlled. For the 12 early age-of-acquisition words, the means were, for rated age of acquisition, 2.55, for frequency, 39, for imagery, 4.95, and for meaningfulness, 5.08. For the 12 late-acquired words, the means were age of acquisition, 5.12, frequency, 39, imagery, 4.93, and meaningfulness, 4.72.

The second list (the I list) was selected from the remaining pool of words in the Gilhooly and Hay (1977) nouns. Twelve pairs were selected, matching for age of acquisition and frequency and varying imagery. For the 12 high-imagery words, the means were imagery, 5.56, age of acquisition, 3.75, and frequency, 32. For the 12 low-imagery words, the means were imagery, 2.80, age of acquisition, 3.84, and frequency, 32.

Two sets of booklets, one containing the A list and the other containing the I list, were prepared. One word was typed on each page, and the page order was randomized for each copy of the booklets.

Procedure. All of the subjects were tested on both lists, half being tested first with the A list and half with the I list. A tape-recorded bleep sounded every 5 sec. The subjects were instructed to turn over a page at a time, when the bleep sounded. They were to try to remember the words, which they could recall in any order they wished.

To eliminate short-term memory effects, at the end of the subjects' study of each list, two eight-digit numbers were read out, and the subjects recalled each number before they attempted to write the words from the list. After the subjects had recalled as many words as they thought they could, their recall sheets were removed and the second list was tested.

Results

For List A, the numbers of early and of late age-of-acquisition words recalled by each subject were calculated, as well as the number of subjects recalling each word. There was better recall of words acquired late than for words acquired early (means: late, 7.73; early, 6.27). This difference was significant when tested by an analysis of variance of the subjects' recall [$F(1,29) = 12.866$, $p < .01$, $MSe = 2.568$] and when analyzed for words recalled [$F(1,22) = 6.177$, $p < .025$, $MSe = 13.06$] and a min F' analysis (Clark, 1973) was significant [min $F'(1,41) = 4.173$, $p < .05$].

For the I list, a similar analysis produced a higher mean recall for high-imagery than for low-imagery words (high imagery, 6.87; low imagery, 5.30). However, while this difference was significant when analyzed by subjects [$F(1,29) = 12.159$, $p < .01$, $MSe = 3.02$], it was nonsignificant when analyzed by words [$F(1,22) = 3.821$, $p > .05$, $MSe = 24.09$] and a min F' analysis was nonsignificant [min $F'(1,37) = 3.016$, $p > .05$].

Discussion

Words acquired late were better recalled than were those acquired early in life. This result was also found in a preliminary study with the same set of items by Whittle (Note 1). In the earlier part of this paper, it was pointed out that when imagery and frequency were partialled out from the correlation between age of acquisition and recall in Gilhooly and Gilhooly's (1979) Experiment 3, the result was a small positive correlation that would indicate that later acquired items were better recalled. Finally, it should be noted that the relatively high correlation between age of acquisition and recognition found by Gilhooly and Gilhooly (1979) was also in the direction of better performance for later acquired words. This accumulation of evidence suggests that words acquired later in life are better remembered than are those acquired early.

Assuming the conclusion that later acquired words are more easily remembered is valid, why should this be so? There is sufficient evidence (e.g., Carroll & White, 1973; Gilhooly & Gilhooly, 1979) to suggest that in tasks that basically involve the use of semantic memory, words that are acquired early in life are more easily accessed. The cause of the reverse finding for episodic memory tasks is therefore not likely to be some product of the way late-acquired items are stored in semantic memory. For example, while any model that assumed that the more recently an item was entered into semantic memory, the earlier it would be found in a memory search would be compatible with the episodic memory data, it would contradict the semantic memory data.

It is probable that the basis of the better recall of late-acquired words results from the later acquired

words' having properties that facilitate distinctive encoding. By definition, we are considering words that were learned relatively late in life but yet are as frequently encountered in adult literature as are the control words. At the least, such words will have been encountered in fewer situations, being known for fewer years, and more specific past experiences are likely to be aroused and more distinctive new episodic traces formed in the memory experiment. The late-acquired words will have had less opportunity to be associated with varying encoding situations and different meanings attached to the words, so that they will be less susceptible to encoding variability between learning and testing (e.g., Morris, 1978). Also, there will be reasons why the words are acquired later. They will deal with adult life and may, therefore, be more salient for the young adult subject, arousing, again, more specific memories and producing more distinctive episodic encoding. The late-acquired words will probably not be more emotionally arousing than earlier acquired words, but they will be more relevant to the current activities of the subjects.

In the present study, frequency was controlled, but normally high-frequency words that tend to be acquired earlier than low-frequency words are better free recalled. This suggests that frequency is a more powerful variable than age of acquisition and that it normally swamps the influence of the latter. Age of acquisition, if uncontrolled, will tend to reduce the apparent influence of frequency on free recall.

The results for the manipulation of imagery in the present experiment are not sufficiently reliable to confirm that imageability affects recall independently of age of acquisition. Such difference as does occur is, however, in the expected direction, and it is likely that if the imagery variable could have been more efficiently manipulated by selection from a larger pool with a wider range of imageability and more homogeneity on other variables, a clear difference for imageability would have emerged. Indeed, the negative correlation between age of acquisition and imagery means that if age is not controlled, it will tend to suppress the observed relationship between imageability and recall, so that the covariation of age of acquisition and imagery cannot explain past observations of a relationship between imageability and recall.

At this point, it is worth returning to Gilhooly and Gilhooly's (1979) Experiment 4, since the negative correlation between age of acquisition and imagery is probably the cause of one of their anomalous results. In the initial correlation matrix, a significant negative correlation between imagery and recognition memory occurs; that is, words that are more easily imaged are apparently more difficult to recognize. This conflicts with past research (e.g., Jones & Winograd, 1975; Morris & Reid, 1974), in which imageability was found to have little or no effect upon the correct recognition of old items, but in which far more false positives occurred to new low-imagery words than to high-imagery words: a result attributed by Morris and Reid (1974) to the greater semantic similarity they found between low-imagery words. Given Gilhooly and Gilhooly's (1979) method of testing, a forced choice between an old and new item, little effect of imagery was to be expected. When the covarying influences of frequency and age of acquisition are partialled out, the observed correlation between imagery and recognition is .02; that is, no relationship remains between imagery and recognition.

To return to the theme of the earlier part of this paper, the stepwise multiple-regression analysis may hint, as in the example just given, at relationships that do not exist. Conversely, the analysis may lead to the overlooking of relationships that do exist, such as that between later age of acquisition and higher recall and recognition. It is often a useful technique, but it cannot be used for unravelling the causes of differences in memory from irrelevant covariables. The experimental control and manipulation of the variables is a better method, but that requires enormous effort in collecting data on the various variables. If and when the considerable efforts of K. J. Gilhooly and associates make such data available, a far better understanding of the relationship between the properties of words and their memorability should result.

REFERENCE NOTE

1. Whittle, A. *The effect of the age of acquisition variable on free recall.* Unpublished study, Lancaster University, 1978.

REFERENCES

Carroll, J. B., & White, M. N. Word frequency and age of acquisition as determiners of picture-naming latency. *Quarterly Journal of Experimental Psychology,* 1973, **25**, 85-95.

Christian, J., Bickley, W., Tarka, M., & Clayton, K. Measures of free recall of 900 English nouns: Correlations with imagery, concreteness, meaningfulness, and frequency. *Memory & Cognition,* 1978, **6**, 379-390.

Clark, H. H. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior,* 1973, **12**, 335-359.

Cohen, J., & Cohen, P. *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, N.J: Erlbaum, 1975.

Gilhooly, K. J., & Gilhooly, M. L. Age-of-acquisition effects in lexical and episodic memory tasks. *Memory & Cognition,* 1979, **7**, 214-233.

Gilhooly, K. J., & Gilhooly, M. L. The validity of age of acquisition ratings. *British Journal of Psychology,* 1980, **71**, 105-110.

Gilhooly, K. J., & Hay, D. Imagery, concreteness, age-of-acquisition, familiarity, and meaningfulness values for 205 five-letter words having single-solution anagrams. *Behavior Research Methods & Instrumentation,* 1977, **9**, 12-17.

Gregg, V. Word frequency, recognition and recall. In J. Brown (Ed.), *Recall and recognition.* London: Wiley, 1976.

Jones, S., & Winograd, E. Word imagery in recognition memory. *Bulletin of the Psychonomic Society,* 1975, **6**, 632-634.

MORRIS, P. E. Frequency and imagery in word recognition: Further evidence for an attribute model. *British Journal of Psychology,* 1978, **69**, 69-75.

MORRIS, P. E., & REID, R. J. Imagery and recognition. *British Journal of Psychology,* 1974, **65**, 7-12.

NOBLE, C. E. An analysis of meaning. *Psychological Review,* 1952, **59**, 421-430.

PAIVIO, A. *Imagery and verbal processes.* New York: Holt, Rinehart, & Winston, 1971.

PAIVIO, A. Imagery in recall and recognition. In J. Brown (Ed.), *Recall and recognition.* London: Wiley, 1976.

RICHARDSON, J. T. E. Imagery, concreteness and lexical complexity. *Quarterly Journal of Experimental Psychology,* 1975, **27**, 170-182.

THORNDIKE, E. L., & LORGE, I. *The teacher's word book of 30,000 words.* New York: Columbia University Press, 1944.

UNDERWOOD, B. J., & SCHULZ, R. W. *Meaningfulness and verbal learning.* Philadelphia: Lippincott, 1960.

(Received for publication October 23, 1979;
revision accepted September 25, 1980.)