1979

The Effects of Task and Stimulus Manipulations on Judgement of Knowing Accuracy

Joseph F. King
Loyola University Chicago

THE EFFECTS OF TASK AND STIMULUS MANIPULATIONS ON JUDGMENT OF KNOWING ACCURACY

by

Joseph F. King

A Dissertation Submitted to the Faculty of the Graduate School of Loyola University of Chicago in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

April

1979

ACKNOWLEDGEMENTS

The author wishes to express his sincere gratitude to the director of this dissertation, Dr. Eugene B. Zechmeister, for his help and guidance in the completion of this project. The author is also indebted to the other members of his committee, Dr. Deborah L. Holmes and Dr. Emil J. Posavac. Comments by Dr. John J. Shaughnessy are also acknowledged.

VITA

The author, Joseph F. King, is the son of LaVerne Edward King and Mary (LeSage) King. He was born November 17, 1951, in Kankakee, Illinois.

The author graduated from Bishop McNamara High School in Kankakee, Illinois. In September, 1969, the author entered the University of Illinois at Chicago Circle. One year later, he enrolled at Loyola University of Chicago and in 1973, he graduated Cum Laude with a bachelor of science degree in psychology.

In the fall of 1973, Joseph F. King was granted a research assistantship in experimental psychology at Loyola University of Chicago. In 1976, the author received a masters degree in research psychology from Loyola University of Chicago. The author is a member of the Midwestern Psychological Association.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

LIST OF TABLES

LIST OF FIGURES

CONTENTS OF APPENDICES

INTRODUCTION

Current theoretical descriptions of encoding, or the process by which a memory is established, have emphasized the qualitative characteristics of the learner's cognitive activity. Such concepts as "depth" (Craik & Lockhart, 1972), "meaningfulness" (Jenkins, 1974), and "congruity" (Schulman, 1974) have been used to describe encoding activities which can lead to durable memory traces. According to these viewpoints, remembering is not the result of a stimulus acting on an organism; rather, as Craik and Tulving (1975) have suggested, the mental activity of the learner determines what will be remembered. This new emphasis represents a shift away from concerns with how changes in stimulus characteristics and learning conditions affect learning and retention. Furthermore, this new emphasis is accompanied by several assumptions about the nature of the memory system. Briefly stated, it is assumed that the learner has available a repertoire of learning strategies to be employed in a variety of situations. Also, it is assumed that the learner has the ability to make decisions during learning about how and when these various strategies are best employed. It is from this latter assumption that the issues addressed in this paper arose.

Several sources of information are available to aid the learner in deciding whether or not a learning strategy is appropriate for a given task. First, through a history of processing verbal information,

individuals come to understand their own memory ability. Flavell has coined the term "meta-memory" to refer to knowledge about one's memory (Kreutzer, Leonard & Flavell, 1975). For example, experience with a wide variety of learning tasks will obviously contribute to meta-memory and may allow the learner to direct encoding activities in a manner that is most successful.

Another source of information used to guide encoding activity would be one's judged progress toward a learning goal. If the success or effectiveness of an encoding effort can be assessed during learning, decisions can be made about how subsequent efforts should be allocated. Consider the following situation. Suppose a student is studying for a final examination. It is likely that some of the information is "learned" and some is not. Since the task is to maximize the amount of information that can be retrieved at a later time, it would be to the student's advantage to spend any remaining study time on that information which is not well-learned. The ability to "judge what is known" or to monitor the effectiveness of encoding during learning has been suggested as an important concern for memory researchers in light of the claim that retention is the result of the active processing of information (Tulving & Madigan, 1970). Furthermore, information flow models of the human memory system (e.g., Atkinson & Shiffrin, 1968) have included "control processes" as theoretical constructs that direct the processing of information. It can be suggested that assessing one's progress toward a learning goal constitutes one role of control processes.

It is likely that the study of meta-memory and memory monitoring

has pragmatic as well as theoretical relevance. An efficient learner may have superior ability to judge the success of an encoding effort. If memory monitoring ability could be improved through learning exercises, students who habitually under-study or over-study may become more efficient in allocating study time.

The present research investigates the ability to judge what is known. In the following discussion, evidence will be presented which demonstrates that adult learners can accurately predict what will and will not be recalled. Also, research examining the ability to judge past retrieval success and the ability to judge the ease with which materials can be learned will be reviewed in relation to the predictions of retrieval success. Finally, a framework within which ony may study memory monitoring ability will be outlined.

Judgments of Knowing

To demonstrate the ability to judge what is known, several researchers have asked subjects to make overt predictions of recall or nonrecall during learning. These predictions are referred to as Judgments of Knowing (JKs). In this section, JKs will be formally defined and experiments which have employed the JK task will be reviewed in some detail.

A JK can be defined as the subjectively rated likelihood of the later retention of presently studied information. Accuracy of the JK is determined by comparing the ratings with later retrieval success. Several aspects of this definition deserve special attention. First, the JK is made with the to-be-learned material present. Thus, JKs can be distinguished from the "feeling-of-knowing" judgment which requires

the subject to predict recognition performance for information that cannot be recalled (Hart, 1965). Secondly, the definition is indifferent to the type of retrieval test to be employed. For example, JKs have been requested during paired-associate (PA) learning (Arbuckle & Cuddy, 1969; King, Note 1; Pasko, Note 2), during free recall learning (Lovelace, Note 3; Zechmeister & Shaughnessy, Note 4), and during recognition learning (Groninger, 1976). Finally, the JK is designed to assess the likelihood of retention of specific information. Judgments of the "percentage" of list items to be retrieved are not included in the present definition of a JK (see LaPorte & Nath, 1976, for an example of this task).

Arbuckle and Cuddy (1969) reported a study utilizing the JK task as presently defined. Subjects were shown a long series of short PA lists for study. As each pair was presented, subjects were asked to respond "YES" if they though the response term would be successfully recalled of "NO" if they thought recall would be unsuccessful. These YES-NO predictions were then compared with recall performance. It was found that subjects could predict recall at greater than chance levels. In an attempt to discover how these judgments were made, the authors asked an additional group of subjects to rate the "ease" with which each pair could be learned. It was found that the perceived difficulty of the pairs was inversely related to the probability with which correct recall was predicted. Arbuckle and Cuddy suggested that subjects were assessing the associability of the pair members at the time of presentation and were using this information as a basis for their JK responses. Also, the authors suggested that this kind of stimulus

assessment may occur covertly in standard PA learning situations.

A second example of the use of the JK task was reported by Zechmeister and Shaughnessy (Note 4) in an examination of the "spacing effect" (or MP-DP effect). By way of background, if items in a free recall task are repeated in a distributed fashion (items intervening between repetitions) recall is generally superior to recall of items repeated in a massed fashion (contiguous repetition; see Hintzman, 1974, for a complete discussion). One explanation for this phenomenon is that while the nominal presentation time is equivalent for both massed and distributed repetitions, the functional study time is less for the second presentation of a massed item than for the second presentation of a distributed item. This "attenuation of attention" hypothesis was supported by Shaughnessy, Zimmerman, and Underwood (1972) who allowed subjects to pace their own presentation of an MP-DP list. Study times alloted to the second presentation of distributed items were greater than study times alloted to the second presentation of massed items. Given this shift in attention, an explanation for why this occurred was needed.

Zechmeister and Shaughnessy reasoned that if subjects made erroneous estimations of the likelihood of recall for massed items, a "rationale" for the incomplete processing of massed items could be offered. They presented a lengthly free recall list containing once-presented items and twice-presented items under both massed and distributed conditions. Following some of the items, subjects were required to make JKs. The usual spacing effect was obtained, and relatively accurate JK responses were observed. Once-presented items

were given lower JK ratings (less likelihood of recall) than twice-presented items; and, in fact, once-presented items were not recalled as well as twice-presented items.

More important to the concerns of this experiment were the JK ratings assigned to the second presentation of massed and distributed items. While distributed items were recalled better than massed items, similar JK ratings were assigned. The implication of this result is that subjects were overestimating their memory for massed items. Given this overestimation, a reason for the shift in attention during the processing of an MP-DP list can be claimed. The JK results of this experiment supported the "attenuation of attention" hypothesis.

A third study employing the JK task was designed to understand the possible sources of information upon which predictions can be made (King, Note 1). Specifically, King examined the effects of prior testing on JK accuracy. Since much of the remaining discussion refers to this study, it will be reviewed in detail.

In the King study, four groups of subjects learned three 24-item PA lists. Two of the groups learned the first two lists under an alternating study-test trial procedure. For the other two groups, intervening test trials were omitted for the first two lists, and a single test trial was given after five study trials. A second distinction between groups was the presence of absence or a JK rating trial after learning. For the first two PA lists, the two JK groups were shown the pairs after learning and were asked to rate the likelihood of recall of the response term when shown only the stimulus term. A six-point scale ranging from "sure to recall the item" to "sure not to

recall the item" was provided for the ratings. The two groups not making JKs were given an additional study trial in place of the JK trial for each of the first two lists in order to equate for total exposure time to the items. In the third list, all subjects made JKs after learning the pairs without intervening test trials preceeding the JKs.

When test trials preceeded the JK ratings prediction accuracy was substantially greater than when no test trials were given during learning. Further, the superiority shown by the group receiving test trials during the learning of the first two lists completely vanished when test trials were omitted on the third list. In an attempt to explain these findings, it was concluded that feedback information relevant to the JK was made available through the test trials. From the preceeding series of test trials, the subject could remember past performance and thereby have a basis on which to make the JK rating. In other words, the subject could infer that "since the item was recalled earlier, it is known". In support of this conclusion, consistently high correlations were observed between the JK rating assigned to an item and the number of trials on which that item was successfully retrieved.

King also examined the relationship between JKs and perceived "ease-of-learning" (EL). The PAs were shown to an additional group of subjects, and mean EL ratings were obtained for each pair. It was found that the items' EL ratings were highly correlated with the items' probability of recall ($\underline{r}$ = .63) and with the mean JK assigned to the item ($\underline{r}$ = .73). This pattern of correlations offers some

support to Arbuckle and Cuddy's claim that subjects assess the "ease" of pairs in order to make JKs. The effects of the preceeding test trials, however, suggests that the assessment of EL is not the sole source of information relevant to the JK.

While relatively few experiments have been reported which utilize the JK task, the above examples provide a working definition of the JK and suggest some direction to future research efforts. All three studies can be used as evidence for a person's ability to accurately make JKs. However, some caution is needed because each study used a different statistical technique to evaluate JK accuracy. (In a later section of this paper, the optimal method of scoring JK performance will be discussed.) It is impossible to determine whether erroneous conclusions were drawn as a result of the method of scoring JK performance. It should be noted that all three studies reported increases in probability of recall as a function of increasing rated likelihood of recall. Thus, it can be argued that under certain circumstances JKs can be accurately made.

Finally, the above experiments suggest an avenue of investigation which may lead to an understanding of how learners make JKs. Briefly stated, the JK task can be seen as a discrimination task. The subject must differentiate those items which can be recalled from those items which cannot be recalled. The King study demonstrated that this discrimination may be made on the basis of an item's "retrieval history". If the learner can accurately say "I got this item correct before", then correct recall will be predicted. Similarly, the Arbuckle and Cuddy experiment demonstrated that the discrimination

between items which will and will not be recalled may be made on the basis of perceived "ease" or "associability". In the following sections, the ability to monitor past performance and the ability to judge the "ease" with which items can be learned will be examined.

Memory for Remembered Events

In a multi-trial learning task, subjects have the opportunity to direct attention or encoding efforts on the basis of previous test trial performance. Zacks (1969) allowed subjects to pace their own presentation of a multi-trial PA list. Across a series of study and test trials, she monitored the study time assigned to each item. If a subject failed to retrieve an item correctly, Zacks found that on the following trial, that item was studied for a longer period of time than if the item was correctly retrieved on the preceeding attempt. Zacks suggested that the differential allocation of study time as a consequence of test trial performance is performed covertly under experimenter-paced PA learning conditions.

In a similar demonstration, Masur, McIntyre and Flavell (1973) required elementary school and college subjects to learn a list of items which was 50% longer than their immediate memory span. After presentation of the list for 45 sec., a recall test was administered. Then, for subsequent study trials, the subjects were told that they could study only one half of the items, and that they were to indicate which items they wanted to study. For the older subjects, the authors found that if an item was previously recalled, it was much less likely to be selected for further study than if previous recall attempts were unsuccessful.

These two studies demonstrate that past performance can direct

learning efforts. However, in order for a subject to benefit from past performance, past output must be accurately remembered. The learner must remember the "event" or occurrence of a successful act of retrieval. Gardiner and Klee (1976) refer to this ability as "memory for remembered events" (MRE) and have reported several experiments concerning output monitoring in free recall.

In a free recall task the subject is required to reproduce events from memory. During output, the nature of the task demands that the subject "keep track" of which items have and have not been reported in order to avoid repetition errors. To test this ability, Gardiner and Klee presented subjects with 10 lists of 15 items for free recall. Following the series of free recall tasks, all the items were presented and the subjects were required to indicate which items they had recalled on the earlier tests. This was referred to as a "recall-recognition" test. The usual serial position curve was obtained for the free recall tasks. However, a much different serial position curve was obtained for the recall-recognition task. Output monitoring, or MRE, was much less accurate for items which occupied recency positions during input than for items which occupied pre-recency positions. The same results were obtained when the initial task required serial recall rather than free recall. In a further experiment, the initial study lists were tested for recognition memory. Under these conditions, MRE was generally lower than following recall, and no differences in MRE were observed as a function of input position.

The authors concluded that the act or retrieval is an experience

which is encoded in episodic memory in much the same way that a to-be-learned stimulus is encoded. The act of retrieval is accompanied by certain articulatory or motor responses, and the saliency of these "performance features" can be influenced by the type of test used (recall or recognition) or by the mode of output. To support this claim, a series of short free recall lists were presented, and output was either oral, written, or oral plus written. Furthermore, during some of the test trials, "feedback" was impaired. That is, white noise and special writing paper prevented the subjects from knowing what they had recalled. When MRE performance was examined, oral plus written recall resulted in greater MRE accuracy than written recall. The oral output condition resulted in the lowest MRE accuracy. Furthermore, regardless of output condition, when feedback was impaired, MRE was less accurate than when it was not impaired. Presumably, the saliency of the experience of retrieval was decreased when feedback was impaired.

For the purposes of the present discussion, the Gardiner studies have demonstrated that intra-trial output monitoring ability can be empirically measured. Furthermore, and more relevant to the present discussion, Gardiner has suggested that MRE is likely to have inter-trial relevance as well.

> Here the subjects' knowledge of his previous performance can provide feedback information which may lead to decisions with respect to the regulation of a variety of control processes. For instance, the subject may modify his coding strategies, rehearsal patterns, and output priorities as a result of his performance on previous test trials. (Gardiner, Passmore, Herriot & Klee, 1977; pp. 53).

The conclusion is that subjects can (and do) modify study behaviors on

the basis of their memory for what they have remembered. It can be suggested that the JK performance observed in the King study (under conditions where test trials preceeded the JK) depended on the discrimination between previously recalled and previously unrecalled items. Therefore, a potential "cue" or "attribute" which allows a JK discrimination to be made is "retrieval history".

It should be pointed out that MRE ability has not been completely explained. It is likely that the recall-recognition task is a test of situational frequency discriminations. That is, in the Gardiner task, each study item is presented once and each item may or may not be recalled. If the item is recalled, the presentation frequency is incremented. (The subject's output can be seen as a "presentation".) During the recall-recognition test, the subject then must discriminate between items presented once (nonrecalled) and items presented twice (recalled). Numerous demonstrations of the ability to make such frequency discriminations have been reported (Hintzman, 1969; Hintzman & Block, 1971; Underwood, Zimmerman & Freund, 1971).

The same ability to make situational frequency discriminations could have been involved in the monitoring of "retrieval histories" in the King study. That is, three test trials were administered prior to the JK rating task. If subjects could accurately judge the frequency with which each item was successfully recalled, this cue could be used to make the JK. This point will be discussed in greater detail later in this paper; but first, an additional source of information relevant to the JK will be discussed.

Ease of Learning

As was pointed out earlier in this discussion, the ability to make JKs is likely to be dependent on the ability to judge the "ease" with which verbal items will be learned. Several experiments have been reported which suggest that adult learners have some understanding of the characteristics which do, in fact, determine learning ease.

An early attempt to study the relationship between item characteristics and perceived ease of learning was reported by Underwood and Schulz (1960, pp. 19-21). Subjects were shown a sample list of 10 items varying in meaningfulness. Then for each of 96 disyllables, subjects were asked to rate the ease with which the item could be learned relative to the sample list. The correlation between the EL ratings and meaningfulness values was .90. In a second study, the authors found a correlation of .86 between rated EL and ratings of association value for 90 nonsense syllables. After completing the ratings, Underwood and Schulz asked their subjects to indicate what factors they had used in making EL ratings. Among the dimensions suggested were familiarity, pronunciability and the association an item suggested.

Richardson and Erlebacher (1958) performed a similar study examining the perceived ease of pairs of stimuli. While some subjects were given the EL instructions described in the previous paragraph, other subjects were asked to rate the association or connection between pair members. For pairs of words, nonsense syllables and CCC trigrams, the rated EL of the pairs was highly correlated with the rated association between pair members. Also using pairs of nonsense

syllables Battig (1959, 1960) found that rated EL was correlated with the average association value for the members of each pair. These studies suggest that subjects may make EL ratings along dimensions which have been shown to influence learning.

Actual learning performance was compared with EL ratings by Underwood (1966). He first instructed subjects to imagine that they were participating in a free recall experiment. Next, a list of trigrams was presented and the subjects were asked to "rate the speed with which you would learn each trigram in the imagined task". Following the ratings, an incidental recall trial was requested, and then six study-test learning trials were administered. Other groups of subjects made pronunciability or meaningfulness ratings of the trigrams or simply learned the trigrams via the multi-trial procedure. Underwood presented an extensive analysis of the EL ratings and pointed to several problems identified with the "correlational" techniques employed, and so the results will be considered in detail.

First, there was a strong relationship between a trigram's perceived EL and its rated pronunciability (r = .94) and rated meaningfulness (r = .91). Again, the suggestion is that subjects are aware of how item characteristics such as pronunciability and meaningfulness influence learning.

Secondly, there was a strong relationship between perceived EL and actual learning. For the group that made EL ratings and learned the items, a very high correlation was observed between the mean EL rating for an item and the number of times an item was recalled across the six learning trials (r = .92). The same relationship was

observed when learning scores produced by the group that did not make EL ratings were entered into the correlation. This second group was necessary to assure that subjects were not "biasing" their learning as a consequence of the EL ratings they had made. That is, perhaps subjects could choose not to rehearse items that they had rated as difficult to learn.

A further question Underwood asked was whether or not subjects could assign EL ratings consistent with their own learning. That is, can subjects "predict" their own learning. To this end, the correlation between an item's EL rating and the number of correct recalls across the six learning trials was computed for each subject. The correlations across subjects ranged from -.32 to .78 with a mean of .48. Underwood interpreted these correlations with caution because of a possible statistical artifact. That is, suppose two learners produced identical EL ratings across the set of trigrams. If one subject learned all of the items by the fourth of six trials, the range in his learning scores would be limited. Furthermore, if the other subject learned only some of the items across the six trials, the range in learning scores would be relatively great. Thus, in the latter case a greater correlation coefficient is likely to be observed as compared with the former case. To examine this possibility, Underwood divided the subjects into six learning "ability" groups. The mean correlation between individual EL ratings and individual learning decreased systematically as learning ability increased. Thus, for fast learners, the correlation computed in this manner results in an underestimation of the relationship between EL ratings with group learning scores. No

systematic relationship between the magnitude of these correlations and learning ability was observed. The mean individual-group correlations across the six learning ability groups ranged between .67 and .72. (It should be noted that Underwood did not report the range of correlations within learning ability groups, and thus nothing can be claimed about the range in ability to judge EL.)

In summary, Underwood's analysis revealed that subjects could estimate the ease with which verbal items could be learned in the absence of instructions to learn. Furthermore, the results indicated that "slow" learners are just as adept at making EL ratings as "fast" learners when group learning scores are used as the criterion. While this finding is intriguing, it should be interpreted with caution. It is not clear that a correlation coefficient is an appropriate index of individual ability. Also, Lippman and Kintz (1968) pointed to another weakness in the Underwood experiment.

Lippman and Kintz suggested that the selection of trigrams in the Underwood study may have lead to artificially high correlations between EL and learning. That is, among Underwood's trigrams were three letter words (e.g., BUG, LOT, KIT) and CCC trigrams (e.g., XFH, PKF, VXK). Since the items were quite heterogeneous along dimensions which determine learning ease, perhaps the task was made artificially "easy". Lippman and Kintz (1968) replicated the Underwood study using only nonword CVCs. Four groups of subjects participated in the experiment. Two of the groups rated the trigrams for pronunciability, and two of the group rated EL. Also, within each rating condition, one half of the subjects rated the items before learning (one incidental recall

trial) and the remaining subjects rated the items after learning (10 study-test trials). Thus, the design resulted in two measures of EL, two measures of pronunciability, and measures of both intentional and incidental learning. Furthermore, it should be noted that Lippman and Kintz made a slight change in EL instructions. They told the subjects to rank the items according to how easy or difficult "a person" would find the trigrams to learn. It will be recalled that Underwood's instructions included the phrase, "which you would recall".

In general, Lippman and Kintz replicated the Underwood findings. However, the magnitude of the EL-learning correlations tended to be slightly less than the magnitude of Underwood's correlations. Unfortunately, it cannot be determined whether the change in instructions or the relative homogeneity of the trigrams was responsible for the decrease. Furthermore, Lippman and Kintz reported that the pronunciability ratings were more reliable than the EL ratings. The two group measures of pronunciability were correlated .95; while the two group measures of EL were correlated .85. Also, when the EL ratings were performed after learning, the correlation between EL and learning was greater than when the EL ratings were made before learning. Thus, it is possible that subjects were monitoring their recall performance while making EL ratings in the former case. Also, although the authors offered no explanation, EL ratings were more highly correlated with intentional learning than with incidental learning.

Pasko (Note 2) studied the relationship between EL ratings and JKs. Subjects were asked to imagine that they were participating in a PA experiment and to rate the ease with which they could learn the PAs.

Four lists of 16 items were presented during this phase of the experiment. Later, the same lists were presented for learning; and JKs were requested before each test. A second group of subjects made JKs during the learning of the four lists but did not make EL ratings for the items. Pasko examined the relationship between individual EL ratings and individual learning. The mean point-biserial correlation between EL ratings and recall-nonrecall was .26 (an individual r of this magnitude would be significantly different from zero). The correlations ranged from -.16 to .59. Thus, on the average, individual learning could be predicted by individual EL ratings. Next, Pasko obtained mean EL ratings and group learning scores by collapsing across individuals. There was a significant correlation between group learning scores and mean EL ratings for the 64 items (r = .45). When the mean EL ratings were correlated with the learning scores of the group that did not make the EL ratings, the coefficient was slightly larger (r = .53). The relationship between EL and learning in the Pasko study patterns that found by Underwood (1966) and Lippman and Kintz. However, since the magnitude of the EL-learning correlations tended to decrease when PAs were employed, it can be suggested that the perception of learning ease may be more difficult for PAs than for trigrams. Pasko was also interested in the relationship between EL ratings and JKs. Some of the subjects made EL ratings and JKs for the same items. Pasko argued that if JKs are based on the perception of an item's relative ease, then the EL rating assigned to an item should be similar to the JK assigned. Pasko obtained the correlation between EL ratings and JKs for each subject for each of the four lists. Across the four lists, the mean

EL-JK correlation was .32 (range = .06 to .60; an individual correlation coefficient of .32 is significantly different from zero). Thus, it was claimed that JKs may depend on the perception of ease.

The above conclusion should be interpreted with caution because of the correlational technique employed. It can be argued that the major problem with this technique is that it would not be possible to observe a correlation of zero between EL ratings and JKs for the same set of items. Consider the following hypothetical experiment. A 10-item free recall list is constructed. One half of the items are very common nouns and the other half of the items are very rare adjectives. First, the items are presented and EL ratings are obtained. If the ability to make EL ratings exists, then one would expect that the nouns would be judged easier than the adjectives. Next, the items are presented for learning and JKs are requested. If accurate JKs can be made, one would expect that the nouns would be judged more likely to be recalled than the adjectives. In this case, a correlation between JKs and EL ratings would be obtained. However, it can be suggested that the correlation must be obtained if EL ratings and JKs are each made accurately. If a zero correlation was obtained between EL ratings and JKs, one would immediately suspect that one of the judgments was inaccurate. Given "perfect" EL rating ability and "perfect" JK ability, the correlation would have to be quite high.

The above criticism can also be applied to the Arbuckle and Cuddy (1969) and King (Note 1) studies. In each of these studies, EL ratings were found to be highly correlated with JKs. However, in both cases, the same items were rated for EL and assigned JKs; and thus, by

definition, a correlation had to be observed. The point of this criticism is not to suggest that EL ratings are unrelated to JKs. Rather, it must be concluded that because of the correlational techniques employed, an empirical demonstration of the relation between the ability to perceive differences in learning ease and the ability to make JKs has not been reported. In the present research, these methodological problems will be overcome.

In summary, there is evidence that individuals can accurately judge what is "easy" or "difficult" to learn. One can speculate that an understanding of the relationship between item characteristics and learning ease can be acquired with learning experience. Also, it is likely that a variety of dimensions are employed in making EL ratings. Verbal items have been scaled for familiarity, meaningfulness, pronunciability, imagery, and orthographic distinctiveness; and many of these item characteristics are correlated. Perhaps a frequency or familiarity judgment is an integral part of an EL rating. Furthermore, as task conditions change, the relevant dimension may also change. For example, if all items are very common, perhaps EL ratings are based on differences in imagery. In general, the studies of EL ratings demonstrate that subjects can discriminate between easy and difficult items; and it is likely that a variety of item characteristics mediate the EL ratings.

Given these considerations, a strategy for making JKs on the basis of the perceived ease or difficulty of the items can be suggested. Earlier in this discussion, the JK task was described as a discrimination between items which can be learned and items which cannot be

learned on a given trial. The subjects must search for a "cue" which allows this discrimination to be made. Item differences may serve as a "cue" in the JK task. The ability to make EL judgments may allow subjects to make JK discriminations.

Methodological Issues in JK Studies

Before the proposed research is considered, the experimental method employed in the JK studies should be closely and critically examined. Two general issues will be discussed. First, it will be argued that in the JK task, subjects may adopt strategies which artificially inflate JK accuracy. A method for the prevention of these strategies will be described. Second, several different methods of measuring and statistically evaluating JK accuracy have been reported. These methods will be examined and the preferred scoring technique will be outlined.

Instructions to encourage accurate recall. The first methodological issue concerns strategies subjects may adopt which lead to artificially low JK error rates. For example, Arbuckle and Cuddy (1969) suggested that low JK error rates could result if the subjects selectively rehearsed items for which "yes" JKs were given and selectively ignored items for which "no" JKs were assigned. The intent of the JK task is not to influence later study behaviors. A second strategy which may lead to artificially low JK error rates is the deliberate withholding of known items at the time of test. If a subject remembered that a "no" JK was given for an item, the response could be withheld in order to achieve a correct prediction. By design, rehearsal patterns and retrieval strategies should be independent of the JK ratings.

To discourage the use of these strategies, a special set of instructions was designed by Pasko (Note 2). The instructions emphasized the importance of recalling as many items as possible. Specifically, at the beginning of the task, subjects were told that they were to participate in a "game" and that the game points would be assigned on the basis of correct recall and correct predictions. The rationale of the game is as follows. A six-point JK scale was provided for the ratings. The scale and the rules of the game were explained to the subjects before learning began. (Table 1 contains the JK scale and the description of the rules that were shown to the subjects.) The subjects were told that they would receive +5 points for each word that is correctly recalled and -5 points for each word not recalled. Next, subjects were told that additional bonus or penalty points would be assigned on the basis of the specific JK responses. Briefly, subjects were told that if their predictions matched their recall, they would receive bonus points corresponding to their degree of confidence in the JK rating. Similarly, penalty points corresponding to the level of confidence were assigned when recall did not match predictions. The maximum bonus or penalty was 3 points. Thus, in terms of game points, correct recall was more "valuable" than correct prediction. It can be seen that the rules of the game encourage maximum recall. For example, suppose that a subject was sure that recall would not occur (e.g., JK = 1). If, in fact, recall was unsuccessful the subject would lose 5 points for nonrecall but would gain 3 points back for making a correct prediction. However, if recall was successful for this item, the

Table 1

JK Scale and Rules for Assignment of Game Points

1 2 3	4 5 6
NO	Yes
Will not recall	will recall

1. For each response term recalled, you will get +5 points.
2. For each response term not recalled, you will get -5 points.
3. If you recall an item, and you made a "yes" prediction (i.e., 4, 5, or 6) then you will get bonus points. If you recall an item for which you made a "no" prediction, then you will lose points.

1	2	3	4	5	6
-3	-2	-1	+1	+2	+3

4. If you do not recall an item, you always lose 5 points. But you may gain points back if your prediction was "no" (i.e., 1, 2, or 3) for that item. If you predicted "yes" for a missed item, again you will lose points.

1	2	3	4	5	6
+3	+2	+1	-1	-2	-3

5. Note that the rules are designed such that you can maximize your points by recalling as many items as possible.

subject would lose 3 points for the incorrect prediction and would gain 5 points for correct recall. Under these conditions, recall of an item, regardless of the JK, would always result in an increase in game points and nonrecall of an item, regardless of the JK, would always result in the loss of points. These instructions are quite complicated and it is possible that many of the subjects did not fully understand the game. However, throughout the instructions subjects were encouraged to recall as many items as possible. It has not been determined if the game instructions do, in fact, prevent the use of selective rehearsal or selective withholding strategies. However, the use of these instructions is a necessary precaution in JK studies.

Measurement of prediction accuracy. A second important methodological issue concerns the measures employed to reflect accuracy of predictions. Several methods have been employed. First, all JK studies that have been reported have shown that the probability of recall increases as the judged likelihood of recall increases. While this result must be obtained if JKs are accurate, the technique cannot be used to measure an individual's JK performance.

Many of the studies of JK ability have viewed the JK paradigm as analogous to an absolute judgment recognition task (Arbuckle & Cuddy, 1969; King, Note 1). The subjects must respond "yes" if they believe that a recallable memory trace is present or "no" if they believe that no recallable memory trace is present. Then, after the ratings are made, recall or nonrecall follows. Given this framework, four JK response outcomes are possible. As is illustrated in Table 2, a "yes" JK prediction followed by correct recall is termed a "Hit", and a "no"

Table 2

JK Response Matrix

			Recall	
			Correct	Incorrect
JK	"Yes"	(4,5,6)	Hit	False Alarm
	"No"	(1,2,3)	Miss	Correct Rejection

Probability of a Hit = Hits / # recalled

Probability of a False Alarm = False Alarms / # not recalled

JK prediction followed by incorrect recall is termed a "Correct Rejection". These two outcomes represent correct predictions. "Misses" ("no" predictions followed by correct recall) and "False Alarms" ("yes" predictions followed by nonrecall) represent incorrect JKs. This terminology will be used throughout the remainder of this paper.

Once the JK response matrix is constructed, several statistical techniques for deriving an accuracy measure can be suggested. First, Arbuckle and Cuddy (1969) performed Chi-square tests on each subject's response matrix. A statistically significant Chi-square value indicated that the distribution of JK responses was "different" than would be expected if only chance were operating. This technique will not be used in the present research for the following reasons. First, one of the assumptions of the Chi-square test is that the observations are independent. It cannot be assumed that one JK in a list will not be influenced by performance on other items. Second, while the purpose of the Chi-square test may reveal that the response distribution is different from chance, it does not indicate just how the responses are distributed. That is, if this technique were to be informative, additional measures which reflect the type of error (i.e., Misses or False Alarms) would be necessary to fully understand JK performance. Finally, the Chi-square technique will not be employed in the present research because it is relatively untested in the memory literature.

As was mentioned above, the JK task can be seen as a recognition task, and various performance measures have been reported in the literature (Kintsch, 1970). Many of the dependent measures are an algebraic combination of the probability of a Hit and the probability

of a False Alarm. Formulas for the computation of these probabilities are contained in Table 2. Both these probabilities are necessarily involved if the measure is to be independent of guessing. The reason is as follows. A subject could easily identify all those items that will be recalled by simply responding "Yes" on the JK scale for every item. Conversely, a subject could be sure of never making a False Alarm by responding "No" on the JK scale for each item. In these two instances a response strategy or criterion is established by the subject. Indeed, these two strategies represent the extreme cases and the actual criteria used by subjects are likely to fall between these two extremes. The point is that if only Hits were examined, one would not know the extent to which guessing was responsible for achieving a given score. When both the probability of a Hit and the probability of a False Alarm are combined, guessing is said to be controlled or removed from the performance measure. Also, the use of both probabilities allows for the possibility of subjects adopting widely varying guessing strategies and yet achieving the same accuracy scores. The debate is over just how the probability of a Hit and the probability of a False Alarm should be combined to produce a performance measure.

Two general theoretical viewpoints concerning recognition performance have been reviewed by Egan (Note 5). First, the high threshold models of recognition postulate that there is some absolute memory state or degree of memory "strength" above which an item will be judged as "old" and below which an item will be judged as "new". For the present purposes, the subject would establish some absolute

criterion which would be used to discriminate between items that will and will not be recalled. Two measures of performance are derived from this viewpoint, and the appropriate formulas are listed below.

$$\text{Prob. Correct} = \text{Prob. (Hit)} - \text{Prob. (False Alarm)} \qquad (1)$$

$$\text{Prob. Correct} = \frac{\text{Prob. (Hit)} - \text{Prob. (False Alarm)}}{1 - \text{Prob. (False Alarm)}} \qquad (2)$$

The reasoning behind these two Prob. Correct measures differs. According to Formula 2 the ability to judge a new item as new is non-existent. That is, a new item is correctly classified as new on the basis of chance. Formula 1, on the other hand, is based on the assumption that "true" recognition performance is a combination of the ability to judge what is "new" as well as the ability to judge what is "old". The derivation of these formulas has been presented by Egan (Note 5). Since a measure of JK performance should reflect both the ability to judge what is known and the ability to judge what is not known, Formula 1 is preferred for application to the JK paradigm.

A second general framework for the analysis of recognition performance that Egan (Note 5) discusses is the theory of signal detection. According to this viewpoint, no absolute threshold of memory strength is used to discriminate old from new items. Rather, it is assumed that the memory strength or familiarity values of old and new recognition test items are each distributed normally. Although the mean of the distribution of familiarity values for old items is greater than that for new items, the two distributions overlap. Since the two distributions overlap, perfect responding is impossible. During recognition testing, some decision rule or cut-off point is

established such that Misses and False Alarms are kept to a minimum. If it is assumed familiarity values for old and new items are normally distributed with equal variance, the probability of a Hit and the probability of a False Alarm can be said to correspond to areas under the normal curve. These two areas can then be used to compute the differences between the mean of the "old" and "new" item familiarity distributions. This difference is the measure d' and is independent of guessing. Tables of d' values for given combinations of the probability of Hits and False Alarms are provided by Green and Swets (1966) and Hochhaus (1972). A measure of criterion, beta, can also be derived under the theory of signal detection. Beta can be used to indicate whether the subject established a relatively "strict" or "lax" criterion.

While the signal detection measures have been rather popular throughout the recognition memory literature, some degree of caution is in order. As was mentioned above, application of the theory of signal detection requires the assumption that the underlying familiarity distributions are normal. Moreover, a large amount of data from a single subject is necessary to validate this assumption mathematically. Some researchers have elected to employ measures which do not require this rather elaborate assumption about the underlying recognition decision processes (Underwood, 1974).

One favorable aspect of the recognition accuracy measures considered above is that they have been employed and accepted in the memory literature. These measures could be adopted quite easily for use in the JK paradigm. However, with regard to the JK task, one

weakness of these scoring techniques is that the responses are seen as strictly dichotomous. In the present research, a six-point JK scale was presented and even though subjects were told that the purpose of the task was to make a recallable--nonrecallable discrimination, they were instructed to use points all along the scale in order to reflect the confidence they have in their judgments. By collapsing the six-point scale into two categories ("yes" and "no") some information is lost. Furthermore, it is not likely that all subjects used the scale in the same manner. That is, some subjects may have clustered their responses around the center of the scale and other subjects may have used extreme points of the scale quite freely. Thus a desirable measure of JK performance under the present conditions would be derived from scale values actually used.

Shaughnessy (Note 6) has suggested a relatively straightforward technique for measuring JK accuracy that does take into account the subjects' use of the six-point JK scale. The measure was taken from a study of confidence judgments by Zimmerman, Broder, Shaughnessy and Underwood (1977) and is called the Confidence Accuracy Quotient (C.A.Q.). The formula is presented below.

$$C.A.Q. = \frac{\bar{X}_{JK_R} - \bar{X}_{JK_{\bar{R}}}}{\sqrt{s_R^2 + s_{\bar{R}}^2}}$$

The mean of the JK scale values assigned to the nonrecalled items is subtracted from the mean of the JK scale values assigned to the recalled items. In order to control for subjects' varying tendency to use extreme scale values, the difference between the means is divided

by the square root of the pooled variances of the recalled and non-recalled JK responses. Accurate JK performance would result in a positive C.A.Q. value. The magnitude of this accuracy measure is dependent on the relative difference of the JK values assigned to recalled and nonrecalled items. In theory, a subject using only the middle two or three scale values could be just as accurate as a subject who freely used all six JK scale values. Of course, the measure is undefined if recall is perfect or if there is no variance in the JK responses. Appendix A contains an illustration of how the formula is computed and the conditions under which the formula can be used.

The C.A.Q. measure is preferred for the present study because it captures a maximum amount of information from the JK response protocols. Furthermore, the measure fits well into the theoretical discussions of the JK task presented earlier in this paper. The JK task is seen as involving a discrimination between items that will be recalled and items that will not be recalled. The accuracy of both "Yes" and "No" JK scale values enters into the computation of the C.A.Q. Since the C.A.Q. measure is new and relatively untested, the Probability Correct measure and the signal detection measures were also employed in the results to be reported. These latter measures will provide an indication of the validity of the C.A.Q. measure.

The Present Research

In the preceding discussion, JKs have been formally defined as the subjectively rated likelihood of the later retrieval of presently studied information. The JK task requires a discrimination between

items that are likely to be recalled and items that are not likely to be recalled. Given this framework, it can be argued that the JK is a relative judgment and that some dimension exists along which the discriminations can be made. The purpose of the present research is to characterize this underlying dimension.

An initial way of viewing the JK discrimination is that, across a variety of learning situations, one "universal" dimension is employed in the JK process. This uni-dimensional view would suggest that something like "memory strength" is used to differentiate between known and unknown items. However, this viewpoint can be shown to be inadequate in light of previous JK research. According to a uni-dimensional view, in any learning task for which learning is at a less than 100% criterion, items will differ in terms of memory strength. That is, certain items will be recallable and other items will be unrecallable. Given this qualification, JKs should be consistently accurate regardless of changes in task conditions. Consider the King study. Correct recall was about 50%, and thus items can be said to have differed in terms of "strength". However, when no test trials preceded the JK, JK accuracy was substantially lower than when test trials were administered prior to the JK. Thus, changes in task conditions did lead to differences in JK accuracy. The uni-dimensional viewpoint would also predict that JK performance across a variety of situations would be highly correlated. That is, if a subject is "good" at judging memory strength when test trials are present, he or she should also be "good" at judging memory strength when no test trials are present. In the King study, no such correlation

was found. Thus the uni-dimensional viewpoint is not consistent with the available data.

The present research will attempt to support an alternative viewpoint. It can be suggested that a variety of dimensions exist along which JK discriminations can be made. According to this multi-dimensional viewpoint, the particular decision axis or dimension is a function of task and stimulus conditions. It will be argued that JKs will be accurate to the extent that items within a to-be-learned list differ along some perceptible dimension which the learner believes to be related to learning. From the studies reviewed earlier, two general classes of dimensions can be suggested. First, task-specific manipulations may influence the accuracy of JKs. As was seen in the King (Note 1) study, changes in presentation-test conditions influenced JK accuracy. When test trials were present, a "frequency of past success" dimension was available, and JKs were more accurate than when this dimension was removed from the situation. The research of Gardiner and Klee (1976) suggests that subjects can accurately monitor past retrieval performance and thus the dimension was perceptible. It is argued here that past retrieval performance is one of a group of dimensions that is related to task or presentation conditions which provides an index of discriminability between items likely to be recalled and items not likely to be recalled.

A second class of cues can be referred to as item-specific dimensions. Although the evidence is weak, differences in the perceived ease or difficulty of list items may also provide a decision axis along which JK discriminations can be made (Arbuckle & Cuddy,

1969). That is, irrespective of task or presentation conditions, differences in item characteristics which are believed to be related to likelihood of recall may allow subjects to judge which items can and cannot be recalled. According to a multi-dimensional viewpoint, both task-specific and item-specific cues can influence the level of JK accuracy. The research to be reported examined the effects of both task and stimulus manipulations on the accuracy of JKs. It was anticipated that this avenue of investigation would broaden the understanding of how JKs can be made under a variety of conditions.

In the present experiment, three groups of subjects learned three lists of paired-associates. For each list, JKs were requested before the test trial. Across the three lists, the presence of two "cues" or dimensions that are related to the likelihood of recall was manipulated.

For one group, a task-specific dimension was emphasized. Specifically, each of the first two lists contained items presented either once or three times. This group is referred to as the Varied Frequency group. In order to demonstrate the effectiveness of presentation frequency as a cue for making accurate JKs, this dimension was "removed" for the third list. All items were presented twice for learning of the third list.

For the second group, an item-specific dimension was made salient. Each of the first two lists was composed of items which varied in terms of "ease" of learning. As was stated earlier in this paper, perceived "ease" is not a unitary dimension. Studies of EL rating ability have demonstrated that familiarity, meaningfulness, and

pronunciability of verbal items may contribute to "rated" ease. Furthermore, many item characteristics covary. Familiarity, meaningfulness, and pronunciability are intercorrelated (Hall, 1971; Underwood & Schulz, 1960). For the present purposes, these characteristics were allowed to covary. Stimulus-response pairs making up the first two lists were constructed such that the likelihood of recall of the response terms varied widely within a list. Again, for the third list, this dimension was "removed". All third list items were of comparable ease. This group will be referred to as the Varied Ease group.

The third group was the Control group. All three lists were composed of items of constant ease, and each item was presented twice. Thus, the specific dimensions which could aid the JK discriminations for the Varied Frequency and the Varied Ease groups were not available to the Control group.

It was expected that the Varied Frequency and the Varied Ease groups would show greater JK accuracy on the first two lists than the Control group. Also, it was expected that the JK performance shown by the Varied Frequency and the Varied Ease groups would be greater on the second list than on the first list. Experience with the learning conditions may be required before the effects of the manipulated dimensions become apparent. Finally, on the third list, when the variation in presentation frequency and the variation in item difficulty are removed for the Varied Frequency and the Varied Ease groups, JK performance was expected to equal that of the Control group.

In addition to the between-group comparisons described above,

individual differences in JK ability were also examined. If the multi-dimensional viewpoint of the JK processes is valid, then the ability to make accurate JKs under given conditions should be correlated with the ability to perceive differences along the available dimensions. In the present case, it is argued that the learning conditions for the Varied Frequency group allow the JK discriminations to be made on the basis of perceived differences in presentation frequency. If this reasoning is correct, then those individuals who are relatively adept at making situational frequency discriminations should also be adept at making JKs under these conditions. Similarly, in the preceding discussion it was suggested that the Varied Ease group could make accurate JKs by disciminating between items along the dimension of perceived ease of learning of the list items. In this case, those subjects who are accurate judges of learning ease should also be adept at making JKs. In order to evaluate this reasoning, a battery of tests designed to measure specific memory abilities was administered after the JK tasks. The tests are briefly described in the following paragraphs.

The series of tests can be divided into two general categories. First, two tests measured the ability to make situational frequency discriminations. Second, four tests were created to measure the ability to assess characteristics of verbal stimuli that are related to learning ease.

To measure the ability to discriminate situational frequency, a long list of items was presented. The list consisted of items presented at each of several frequencies. Immediately afterwards, pairs

of list items were presented and subjects were instructed to select the member of each pair that had occurred more frequently in the list. A memory-for-remembered-events test patterned after Gardiner and Klee (1976) provided a secondary measure of frequency discrimination ability. It was hypothesized that performance on these two tests would be correlated with the JK accuracy scores of the Varied Frequency group.

Since the ability to judge learning ease is not well understood, several tests of this ability were designed. First, an EL rating method used by Underwood (1966) was adapted for use with paired-associates. Pairs which varied in ease were presented and subjects were asked to rate the pairs on a six-point EL scale. The same pairs were presented to an additional group of subjects in order to obtain actual ease of learning scores. The subjects' ability to rate ease was defined as the correlation between their ratings and the actual learning scores.

A second attempt to measure the ability to judge EL was a two-alternative forced choice test. The EL scale values reported by Richardson and Erlebacher (1958) were used to construct a list of 30 pairs of paired associates. Within each pair, the learning ease of each item was varied. The subjects were instructed to select the member of each pair of items that was easier to learn.

Underwood (1966) and Lippman and Kintz (1968) reported that perceived ease of learning was highly correlated with the meaningfulness or association value of the rated items. Also, as was mentioned earlier, perceived ease of verbal material is likely to be related to

the frequency of occurrence in the language. Thus, it was felt that the ability to perceive differences in ease of learning could be measured indirectly by asking subjects to judge background frequency and association value of English words. For the background frequency discrimination task, words representing the entire range of frequencies of occurrence in the English language were non-systematically paired. Subjects were asked to select the member of each pair that occurred more frequently in printed English. The same technique was employed for the meaningfulness discrimination task. Pairs of words were presented and subjects were instructed to select the member of each pair for which associates were more easily generated.

Since so little is known about how ease of learning is perceived, the intercorrelations among these four tests are of interest. Generally, it was expected that performance on each of these four stimulus assessment tasks would be correlated with the JK performance of the Varied Ease group.

In order to assure the validity of the individual differences analyses, the relative magnitude of the correlations between JK performance for each group and the various ability measures must be examined quite carefully. The logic of this design demands that the ability which correlates with JK performance for one group should be uncorrelated with the JK scores of the other group. Specifically, the ability to judge situational frequency should be more highly correlated with the JK performance of the Varied Frequency group than with the JK scores of the Varied Ease group. Conversely, the ability to judge

learning ease should be more clearly related to the JK performance of the Varied Ease group than the JK performance of the Varied Frequency group. If the manipulated task and stimulus conditions were actually involved in the JK process, then the memory ability tasks should differentially predict JK performance. This should be kept in mind when the individual differences analyses are discussed later in this paper.

METHOD

JK Tasks

Design. Three groups of subjects learned three lists of paired-associates. All subjects made JKs after studying each list. A transfer design was employed and the construction of the first two lists defined the major independent variable. For one group, the first two lists were composed of items which differed widely in terms of the learning ease of the pairs. This group will be referred to as the Varied Ease group. For the second group, list items did not differ in ease-of-learning. However, items within each of the first two study lists were presented either once or three times. This group will be referred to as the Varied Frequency group. Finally, the Control group learned three lists that were composed of items which did not differ in learning ease and which did not differ in presentation frequency. The third list learned by the Varied Ease and Varied Frequency groups was identical to the third list learned by the Control group.

Materials. Five 20 item paired-associate lists were constructed for the JK tasks. Stimulus terms were CVC trigrams selected from the Archer (1960) norms, and response terms were two-syllable nouns taken from the Paivio, Yuille, and Madigan (1968) norms. For three of the paired-associate lists, stimulus and response terms were selected from the middle ranges of meaningfulness values reported in

the respective norms. Since the range of meaningfulness across pair members was rather limited, these lists will be referred to as the Homogeneous lists. Stimulus terms association values ranged from 65 to 85 on the 100 point scale (e.g., JOL, YAC). Response term meaningfulness ranged from 4.9 to 6.4 on the 10 point scale (e.g., patent, welfare). Background frequency of these response terms ranged from 9 to 49 occurrences per million (Thorndike & Lorge, 1944). Stimulus and response terms were randomly paired.

The two additional lists were composed of items which differed widely in terms of learning ease. These two lists will be referred to as the Heterogenous items sets. One half of the items within each 20-item list were formed by pairing a high-meaningful stimulus with a high-meaningful response. Association values of these stimuli ranged from 35 to 90. Response term meaningfulness of these items ranged from 6.5 to 9.1 and response term background frequency ranged from 50 to more than 100 occurrences per million words (Thorndike & Lorge, 1944). The remaining 10 items within each list were formed by pairing a low-meaningful stimulus with a low-meaningful response. For these difficult items, association values of the stimulus terms were all less that 24 on the 100 point scale. Meaningfulness values for these response terms ranged below 4.5, and background frequencies were less than 22 occurrences per million words (Thorndike & Lorge, 1944).

The Varied Frequency and Control groups learned the three homogeneous lists. The study lists presented to these two groups differed in terms of presentation frequency of the pairs. For each of the

three lists learned by the Control group, each item occurred twice in the study series, and the average lag between repetitions was about 20 items. The first two study lists for the Varied Frequency group contained items presented at each of two situational frequencies. Specifically, one half of the items were presented once (1-p) and one half of the items were presented three times (3-p). The items were ordered such that within each tenth of the study list one 1-p item and three 3-p items occurred. Otherwise, the order was random, and the lag between repetitions was about 10 items. The third study list for the Varied Frequency group contained only twice-presented items.

The Varied Ease group learned the two heterogeneous lists followed by one of the homogeneous lists. Each of the three study lists contained two repetitions of each pair.

Stimulus and response pairs were typed on index cards for study trial presentation. A blank card was placed on the top and a card reading "STOP HERE" was placed on the bottom of each deck of study cards.

For the JK trial presentation, study pairs were ordered randomly with the restriction that items of each type (i.e., 1-p, 3-p, "easy", "hard") were interspersed throughout the entire JK list. Pairs were printed in a single column, and next to each pair was a blank line on which the JK response was to be written. The test lists were constructed by ordering the items in a different random sequence. Stimulus terms were printed in a single column and a blank line was provided

for the written response next to each CVC. The JK and test lists were inserted in envelopes that were designed to allow exposure of one item at a time.

Procedure. Subjects were seen in pairs and were assigned to groups by a blocked-randomization procedure upon appearance at the laboratory. All subjects were told that they were to participate in a study of memory and that their ability to predict what was known would be of concern. Participants were instructed that pairs of items would be presented and production of the two-syllable word would be required when shown only the CVC as a cue. An example was given if further clarification was needed.

The JK task instructions were presented before the first study trial began. Subjects were shown the JK rating scale (see Table 1) and then told that after studying the items the list would be shown again and they would be required to predict which response terms they would recall and which response terms they would not recall.

The JK scale and the "game" concept discussed in an earlier section of this paper were explained in detail at this point. Subjects were told that the six-point scale was designed to allow a YES-NO prediction and to measure the confidence of the prediction. A high number (i.e., 5 or 6) meant that they were relatively sure that recall would follow, and a low number (i.e., 1 or 2) meant that they were relatively sure that recall would not follow. Next, the rules for allotting JK game points were explained. Specifically, subjects were told that regardless of their prediction, correct recall would always result in more game points than incorrect recall (+5 versus -5).

The instructions also mentioned the assignment of bonus points for correct prediction and penalty points for incorrect prediction. The magnitude of the bonus or penalty was determined by the degree of confidence they expressed in their JK ratings. If the experimenter felt that the subjects did not understand these somewhat complicated instructions, the specific rules of the game were not belabored. For all subjects, the notion that scores were most heavily influenced by correct recall was strongly emphasized.

Following these instructions, subjects were told that tape recorded tones were to pace them through the study deck. Tones occurred at a 3 sec. rate. After the study trial, a copy of the JK scale and an envelope containing the JK list was placed in front of the subject. They were then instructed to uncover an item and write their prediction on the list whenever they heard a tone on the tape. For the JK task list, tones occurring at a 5 sec. rate paced the subjects through the list.

A similar procedure was employed for the test trial that followed immediately. An envelope containing the test list was handed to the subjects, and they were told to attempt to write the appropriate two-syllable word next to each stimulus. They were instructed to work on one item at a time and tones occurring at a 5 sec. rate paced them through the test list. Subjects were encouraged to guess if they were unsure.

A 2 min. interval separated each of the three JK tasks. During this interval, any procedural questions were answered.

After each session, study list items were arranged in a new

random order. Also, the order in which the various lists were learned was counterbalanced. Specifically, for the Varied Frequency and Control groups, each of the three homogeneous item sets served as the first, second or third list approximately equally often. For the Varied Ease group, each of the two heterogeneous lists served as the first and second list for approximately one half of the subjects. Furthermore, each of the three homogeneous lists served as the third list for about one third of the subjects in the Varied Ease group.

Ability Tests

MRE task. Immediately following the series of JK lists, a memory-for-remembered-events (MRE) task was administered. The 60 pairs that were just learned were presented and the subjects were instructed to indicate which items they had correctly recalled on the preceeding tests.

Four different MRE test forms were constructed. The Varied Frequency and Control groups each learned the same 60 pairs on the JK tasks (i.e., the homogeneous item sets). These 60 pairs were organized randomly with the restriction that within each fifth of the MRE list, four items from each homogeneous item set occurred. Therefore, regardless of the order in which the item sets were learned, pairs from the first, second, and third lists were interspersed throughout the entire MRE list.

Since subjects in the Varied Ease group learned two heterogeneous item sets and one of three homogeneous item sets, three additional MRE test forms were required. These test lists were constructed by ordering the paired associates such that within each fifth of the MRE list four

items from each of the two heterogeneous item sets and four items from one of the homogeneous item sets occurred. Thus, regardless of the order in which the heterogeneous item sets were learned, and regardless of which homogeneous item set served as the third list, the pairs were interspersed equally throughout the entire MRE list.

The instructions for the MRE test were simple. Subjects were told to place an "X" next to those pairs they thought they recalled correctly during the earlier test trials. Subjects paced themselves through the 60-item series and were encouraged to guess if they were unsure.

Situational frequency judgment task. The purpose of this task was to provide an indication of the subjects' ability to perceive differences in the frequency with which items were presented in a study list. Ten items from the Spreen and Schulz (1966) norms were presented at each of five frequency levels: 1, 2, 3, 4, and 5. The study list required 150 positions. Within each half of the study list, five items were presented at each situational frequency level. Also, the same item never occupied adjacent positions. The study list was presented orally at a 4 sec. rate.

A two-alternative-forced-choice frequency discrimination test was constructed. Given five presentation frequencies plus 10 "new" items (frequency of zero), six items types resulted. There are 15 possible pairings of these six item types and one instance of each pairing was included within each half of the 30-pair discrimination test. Test pairs were printed on sheets of paper. All subjects received the same study and test lists.

The Situational Frequency Discrimination task was the first of the series of ability tests administered during the second session. Subjects were told that they would hear a long series of words and that some of the items may be repeated in the series. Although subjects were told that their memory or the words would be tested, no specific mention of frequency discriminations was made. The test list was administered immediately after the study list presentation. Subjects paced themselves through the test list and were encouraged to guess if necessary.

Background frequency judgment task. In order to measure the ability to perceive differences with which words occur in print, it was first necessary to obtain a group of words which represented the entire range of frequencies of occurrence in the English language. Studies of perceived word frequency by Shapiro (1969) and Carroll (1971) provided such a pool of words. Using their scale values of perceived frequency as a guide, 35 pairs of words were formed for the two-alternative-forced-choice test. The difficulty of the discrimination between pair members differed nonsystematically across items (e.g., result--thud; veterinary--dill).

Subjects were instructed to simply circle the member of each pair of words that occurred more frequently in printed English. Subjects were told that all of the items were real English words even though some occurred very rarely. Guessing was encouraged, and subjects completed the task at their own pace.

Meaningfulness discrimination task. The purpose of this test was to measure the ability to perceive differences in the ease with

which associations may be generated for various words. A 20-pair two-alternative-forced-choice test was constructed by selecting words representing the complete range of meaningfulness values reported by Paivio, Yuille, and Madigan (1968). Pairing of the words was nonsystematic, and differences in the meaningfulness values between pair members varied across items (e.g., bird--decree, saloon--shotgun). No attempt was made to control for background frequency of the pair members.

Before the test was administered, the concept of "word assocation" was explained to the subjects. Subjects were told that one word may remind them of other words. For example, the word "apple" may remind them of "red", "tree", "worm" or "pie". Furthermore, it was explained that some words may remind them of more different words than others. The word "apple" was contrasted with the word "jealousy". Then, subjects were instructed to examine the words within each pair and to circle the word for which more associations could be readily generated. Subjects were encouraged to work slowly and to try to generate associations to each word. The test was self-paced and guessing was allowed.

Ease-of-Learning ratings. This task was intended to measure the ability to judge the ease with which paired-associates could be learned. Subjects were shown a list of 27 paired-associates that were similar to those learned during the JK task. Stimulus terms were selected from the Archer (1960) norms, and response terms were taken from the Paivio, Yuille, and Madigan (1968) norms. Pair members were selected from throughout the entire range of meaningfulness and

association values reported in the respective norms. None of these 27 pairs was among those learned during the JK tasks.

A nine-point scale was to be used for the ratings. End points of the scale were labeled "Very Difficult to Learn" (low numbers) and "Very Easy to Learn" (high numbers). To indicate the extremes of the range in learning ease across the 27 items, a very easy pair and a very difficult pair occupied the first two positions of the list. Ratings for these two pairs (KEY--locker 9 and XYB--inanity 1) were assigned to provide "anchors" for the remaining 25 EL judgments.

Subjects were told to imagine that the pairs were presented for learning and that after a study interval, production of the right-hand member of the pair would be required when the left-hand member of the pair was shown as a cue. Subjects were reminded that they had performed such a task earlier, but for the present purposes, they would not be required to recall the response terms. The instructions emphasized the importance of using the anchor pairs as an aid in making the ratings. Subjects proceeded through the list at their own rate.

In order to derive actual learning scores for the 25 rated pairs, an independent group of 30 individuals learned the pairs. The items were presented twice at a 5 sec. rate. Item repetitions were distributed throughout the list, and the pairs were presented in a different random order for each subject. Recall was tested immediately after presentation. As will be explained in more detail later, these actual learning scores were employed in the computation of an EL score for each subject.

Ease-of-Learning discrimination task. As a secondary measure of the ability to perceive differences in learning ease a two-alternative-forced-choice test was constructed. Richardson and Erlebacher (1958) reported EL ratings for a large pool of paired-associates. From these norms, 40 paired-associates representing very easy (e.g., first--new) and very difficult (e.g., guk--huq) items were selected for use. The 40 items were then grouped into 20 sets of two in order to form the two-alternative-forced-choice test. The grouping was nonsystematic and the magnitude of the difference in learning ease between set members differed widely across the list. Each of the two sets of paired-associates was printed in a numbered row on the test sheet.

It was explained that the purpose of this forced-choice task was very similar to the previous rating task. Subjects were told to circle the paired-associate in each row that was easier to learn. Guessing was encouraged and the test was self-paced.

General Procedure

As was mentioned earlier, the experiment was administered on two separate days. The JK tasks and the related MRE task were administered during the first session. Subjects in the Varied Ease and Varied Frequency groups were asked to return approximately 48 hours later. Every effort was made to accomodate the participants' schedules in order to assure maximum attendance for the second session. On the second day the five remaining ability tests were administered. For all returning subjects the ability tasks were presented in the order in which they were described in the preceding paragraphs. On the second day the subjects were seen in groups of two to four, and a

different laboratory room was employed for the two sessions.

No specific mention was made of the relationship between the JK tasks and the ability tests. The subjects were simply told that the experiment had "two parts". The specific task instructions were given before each test was distributed to the group, and the experimenter waited until all subjects had completed one task before going on to the next.

Subjects

Loyola University undergraduates participated in the experiment in order to fulfill a course requirement. Thirty-six subjects served in each of the three groups. Of those that were asked to return for the second session (n = 72), 45 complied (63%).

RESULTS

The results are considered in two separate sections. The JK tasks will be considered first. Then the relationships between JK performance and the ability tests will be examined.

JK Analyses

Before the recall and JK performance measures were analyzed, it was necessary to examine the quality of the obtained JK data. That is, in order to perform the various JK analyses to be discussed below the subjects' response protocols must meet several criteria. As will be seen later, the Confidence Accuracy Quotient (C.A.Q.) requires that recall be greater than zero percent and less than 100% correct. Also, some variability in the JK responses is required (i.e., the standard deviation of the JK ratings must be greater than zero). Data from four subjects in the Varied Ease group, four subjects in the Varied Frequency group, and three subjects in the Control group failed to meet these two requirements. Thus, these subjects' data were eliminated from further consideration. Consequently, in order to equate the number of subjects in each group, data from one randomly selected Control group subject were also discarded. The following analyses are based on the remaining 32 subjects in each of the three groups.

Recall. Analyses were first performed to determine if paired-associate recall differed between the three groups. Figure 1 displays the mean number of items correctly recalled on each list for each of the

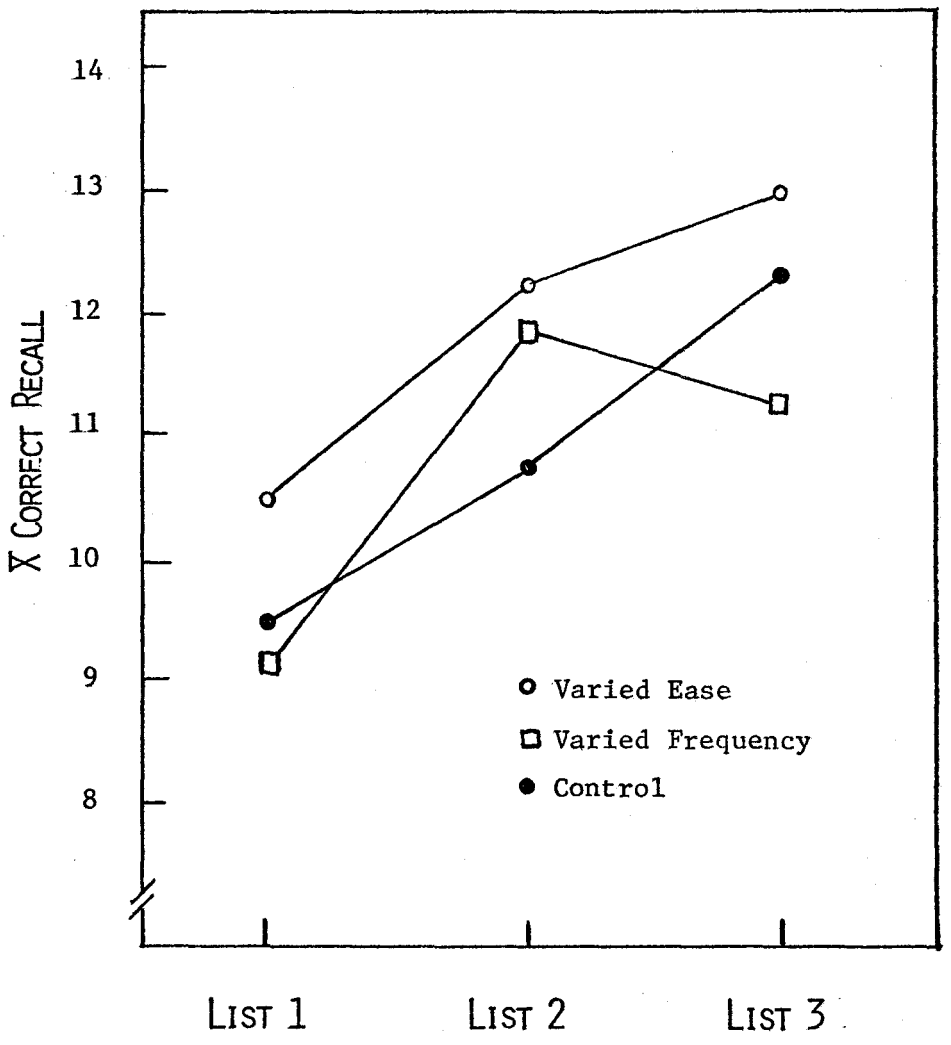

Figure 1. Mean correct recall as a function of groups and lists.

three groups. A 3 by 3 mixed analysis of variance was computed on these recall scores. The results of this analysis are contained in Table 3. Overall level of recall did not differ among the three groups, $F(1, 93) = 1.12$, $p < .10$. However, the main effect for lists reached significance, $F(1, 186) = 24.33$, $p < .001$. As can be seen in Figure 1 recall generally increased across lists. A linear trend analysis supported this conclusion, $F(1, 93) = 44.36$, $p < .001$. It can be seen from Figure 1 that only the Varied Frequency group recall decreased on the third list. Although the List by Group interaction was not significant, $F(4, 186) = 1.12$, $p > .10$, a simple effects analysis revealed that the difference among the recall means on the third list was marginally significant, $F(1, 186) = 5.31$, $p < .10$. Thus, this marginal difference on the third list was the only deviation in the pattern of recall scores shown by the three groups.

For the Varied Ease and Varied Frequency groups, each of the first two study lists contained two different types of items. Recall protocols for the first two lists were collapsed and the level of recall for each item type was examined. As expected, the Varied Ease group recalled more "easy" items than "difficult" items ($\bar{X}$s = 17.00 and 5.94, respectively, $t(31) = 24.50$, $p < .001$). Also, the Varied Frequency group recalled more 3-p items than 1-p items ($\bar{X}$s = 13.03 and 8.25, respectively, $t(31) = 6.01$, $p < .001$).

Probability of recall as a function of JK rating. One indication of the ability to predict correct and incorrect recall can be obtained by simply displaying the probability of correct recall for items given each of the six JK ratings. These probabilities were calculated by

Table 3

Analysis of Variance Summary Table

for Recall

Source	SS	df	MS	F
Groups	82.38	2	41.19	1.12
Error (Between)	3413.86	93	36.71	
Lists	345.76	2	172.88	24.34*
Lists by Groups	31.93	4	7.98	1.12
Error (Within)	1321.38	186	7.10	

* $p < .001$

collapsing across subjects within each group. The proportions correct are displayed in Figure 2. No statistical tests were performed on these proportions because subjects differed in their tendency to use all six JK ratings. However, several global statements can be made regarding this indication of JK accuracy. First, it appears as though the slopes of the curves for the Control group are less steep than the slopes of the curves for the other two groups. Also, within the Varied Ease group (top panel) there seems to be the greatest difference in slopes across the three lists. Also, overall, the slopes of the curves are slightly positive, and while this analysis does not allow precise statements about JK accuracy to be made, such curves must be obtained if the ability to make JK exists.

C.A.Q. scores. As was reviewed in an earlier section of this paper, several different statistical techniques for measuring JK accuracy have been reported in the literature. The Confidence Accuracy Quotient (C.A.Q.) developed by Zimmerman et al. (1977) was selected as the preferred measure. The C.A.Q. is best understood as an index of the subjects' sensitivity of discriminations between recallable and non-recallable items. The C.A.Q. is a ratio. The numerator is computed by substracting the mean JK rating assigned to nonrecalled items from the mean JK rating assigned to recalled items; and the denominator is the square root of the pooled variance of the JK ratings for recalled and nonrecalled items. The formula is as follows:

$$\text{C.A.Q.} = \frac{\bar{X}_{JK_R} - \bar{X}_{JK_{\bar{R}}}}{\sqrt{s_R^2 + s_{\bar{R}}^2}}$$

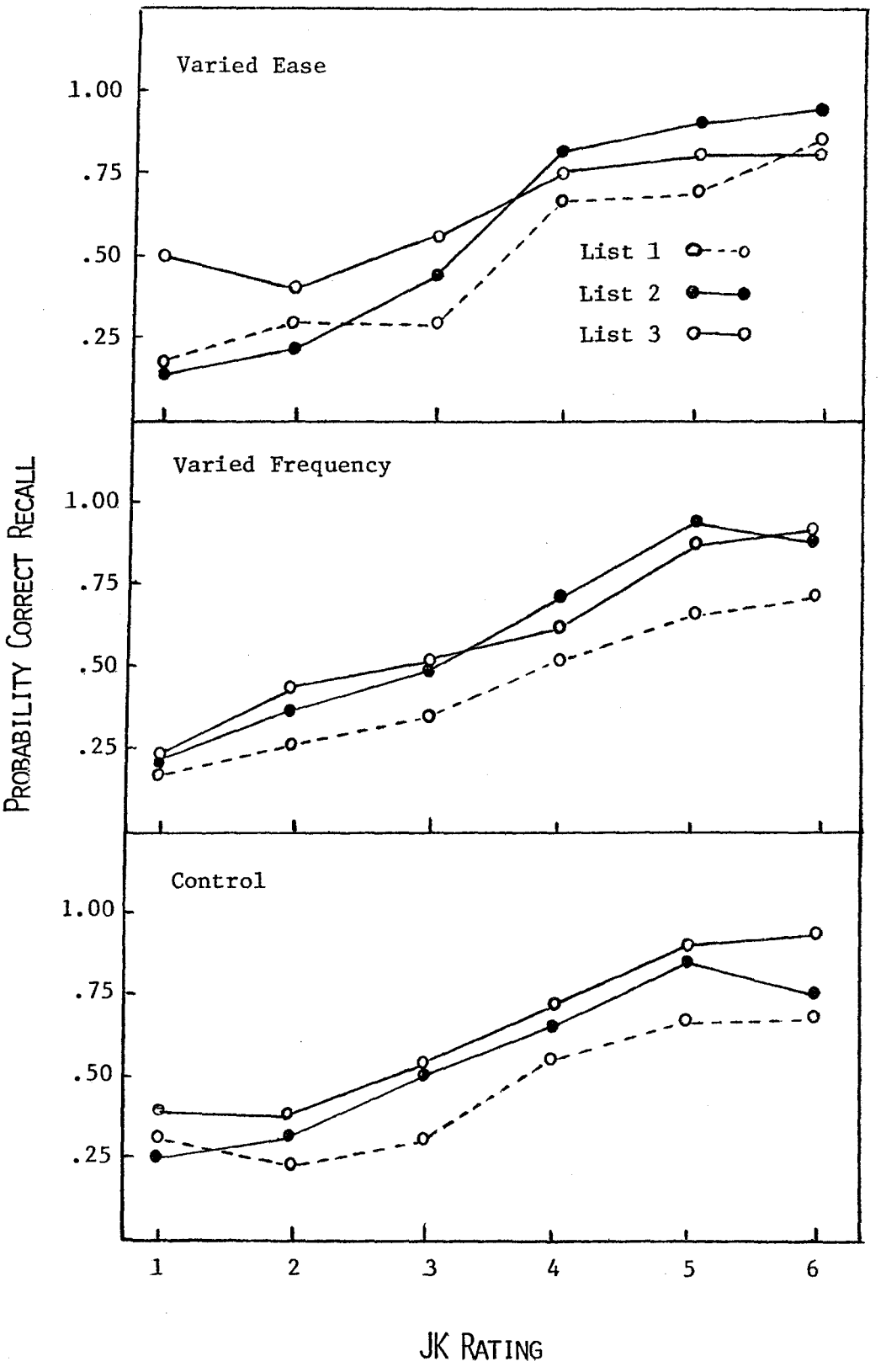

Figure 2. Probability of recall as a function of JK rating.

Conceptually, the measure is very similar to d' of signal detection theory. The advantage of the C.A.Q. over d' is that no assumptions are required as to the underlying distributions of probabilities along a decision axis. Also, since the difference between the two JK means is "weighted" by the variability of the JK ratings, the measure is theoretically independent of changes in the tendency to use extreme points along the JK scale. Subjects who tend to cluster JK ratings around the midpoint of the scale should not necessarily produce higher C.A.Q. scores than those subjects who freely use extreme scale points.

As was mentioned earlier, the selection of this dependent measure meant that data from several subjects had to be discarded. That is, if none (or all) of the items were recalled the measure could not be computed. Also, if the pooled variance of the JKs were zero the measure would clearly be undefined. Given that the purpose of the JK task is to examine the ability to differentially predict recall and nonrecall by assigning JK scale values, it is not unreasonable to exclude subjects' data that do not meet these two criteria.

The mean C.A.Q. scores on each list for each of the three groups is illustrated in Figure 3. A 3 by 3 mixed analysis of variance was performed on these data. The source table is contained in Table 4. A significant main effect for Groups was obtained, $F(2, 93) = 11.10$, $p < .001$; and a significant main effect for Lists was also observed, $F(2, 186) = 6.90$, $p < .005$. Furthermore, the Groups by List interaction reached significance, $F(4, 186) = 7.65$, $p < .001$. In order to describe this pattern of results more completely, the following internal analyses were performed.

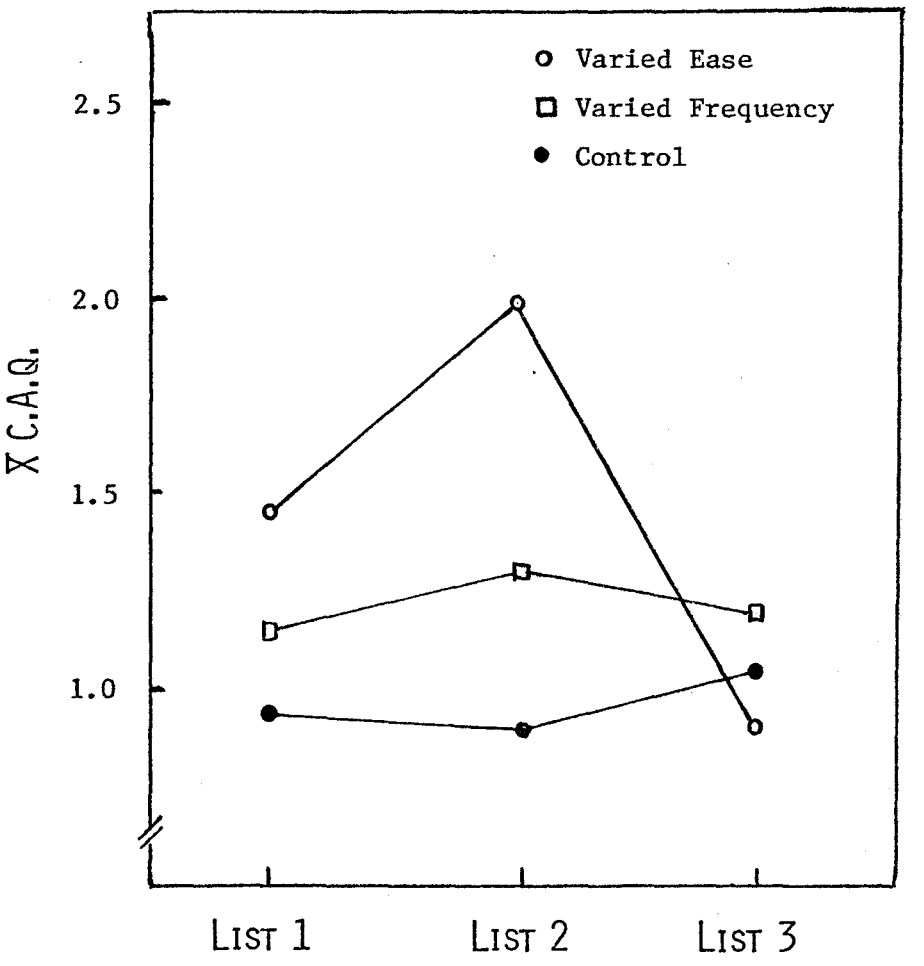

Figure 3. Mean C.A.Q. as a function of groups and lists.

Table 4

Analysis of Variance Summary Table

for C.A.Q. Scores

Source	SS	df	MS	F
Groups	13.80	2	6.90	11.10**
Error (Between)	57.79	93	.62	
Lists	7.11	2	3.55	6.90*
Lists by Groups	15.75	4	3.94	7.65**
Error (Within)	95.73	186	.51	

* $p < .005$

** $p < .001$

The change in JK accuracy across lists was of central concern. As can be seen in Figure 3, JK performance for the Varied Ease group increased across the first two lists and then decreased on the third list. A trend analysis revealed a significant quadratic component, $F(1, 93) = 43.40$, $p < .001$. Furthermore, planned comparisons revealed that JK accuracy for List 2 was significantly greated than accuracy for List 3, $F(1, 62) = 19.85$, $p < .001$. This pattern of JK accuracy scores was as expected. A similar trend analysis was performed on C.A.Q. scores for the Varied Frequency group. No quadratic component was revealed, $F < 1.0$. It was expected that both the Varied Ease and Varied Frequency groups would show increasing and then decreasing accuracy across lists. This expectation was fulfilled only for the Varied Ease group. Finally, the JK performance shown by the Control group was as expected. The means did not differ across lists, $F < 1.0$.

In addition to the trend analyses, differences between group means were examined at each list by computing planned orthogonal comparisons. On the first list, JK accuracy for the Varied Ease and Varied Frequency groups was greater than JK accuracy for the Control group, $F(1, 62) = 26.12$, $p < .001$. Also, the Varied Ease group produced more accurate JK scores than the Varied Frequency group, $F(1, 62) = 15.96$, $p < .001$. On the second list, JK accuracy for the Varied Frequency group was greater than that of the Control group, $F(1, 62) = 24.56$, $p < .001$; and JK accuracy for the Varied Ease group was greater than that of the Varied Frequency group, $F(1, 62) = 37.07$, $p < .001$. Finally, for the third list, JK accuracy for the Varied Frequency group was greater than that of the Varied Ease group, $F(1, 62) = 5.00$, $p < .01$.

The immediate interpretation of the changes in C.A.Q. scores across lists is that JK sensitivity, or the ability to discriminate between recallable and nonrecallable items was influenced by list composition. However, because of the derivation of the C.A.Q. formula, the observed changes could have occurred if the variability of the JK ratings decreased for those lists on which the C.A.Q. scores increased. That is, if for some reason the square root of the pooled variances (denominator of the C.A.Q.) of the JK ratings decreased while the differences between the JK ratings assigned to recalled and nonrecalled items (numerator of the C.A.Q.) remained constant, then the C.A.Q. scores would increase. If this pattern of results occurred, one could only conclude that some spurious changes in JK response tendencies were responsible for the observed changes in C.A.Q. scores. To examine this possibility, the numerator and denominator of the C.A.Q. scores were analyzed separately. Table 5 contains the mean of the difference between JK ratings assigned to recalled and nonrecalled items. It can readily be seen that the magnitude of these differences closely patterns the means of the C.A.Q. scores. An analysis of variance suggests that the means of the differences did differ across lists and groups, $F(4, 186) = 9.10$, $p < .01$. Table 6 contains the means of the denominators of the C.A.Q. scores. If only the magnitude of the JK response variability was responsible for the increases in C.A.Q. scores, then a decrease in the denominators would be observed for those conditions for which an increase in JK accuracy was shown. As can be seen in Table 6, the reverse was the case. Relatively high C.A.Q. scores were accompanied by relatively high variability. Thus, the suggested

Table 5

Mean Difference Between JKs Assigned to Recalled and Nonrecalled Items

	List 1	List 2	List 3
Varied Ease	1.75	2.29	.99
Varied Frequency	1.30	1.58	1.37
Control	.92	.92	1.10

Table 6

Mean Variability of the JK Ratings*

	List 1	List 2	List 3
Varied Ease	1.13	1.42	.79
Varied Frequency	.90	1.08	.98
Control	.78	.79	.86

* Variability is defined as the square root of the pooled variances of the JK ratings assigned to recalled and nonrecalled items.

"artifact" of the C.A.Q. measure was not operating in the present task. The increases in variability of the JK responses accompanying the increases in C.A.Q. means is likely to be a reflection of the tendency to use more extreme JK scale values as the confidence in the perceived difference between recallable and nonrecallable items increases.

In summary, the observed JK performance for two of the groups followed the expected pattern. First, the Control group showed no changes in JK accuracy across lists. Also, the overall JK accuracy for this group was generally lower than the performance for the other two groups. Since no dimension that was related to the JK discrimination was made salient for the Control group, this performance was expected. For the Varied Ease group, JK task conditions were such that a JK-relevant dimension was available for the learning of the first two lists but not for the third list. As expected, JK performance for the Varied Ease group increased across the first two lists and then declined on the third list. Learning conditions for the Varied Frequency group were also designed to make a JK-relevant dimension salient for the learning of the first two lists. It was expected that the JK performance for this group would match that of the Varied Ease group. Although the pattern of JK scores for the Varied Frequency group across the three lists was in the expected direction, the differences were not significant.

Before considering other measures of JK accuracy, the observed C.A.Q. scores can be compared with "chance" performance. Given the theoretical basis for the C.A.Q. measure, there would be no difference between the mean JK rating assigned to the recalled and nonrecalled items if no

ability to make accurate JKs were evident. That is, the C.A.Q. score would be zero if the performance were at chance levels. For each group, the mean C.A.Q. score across all three lists was found to be significantly greater than zero (Varied Ease--$t(31) = 13.74$, $p < .01$; Varied Frequency--$t(31) = 12.09$, $p < .01$; and Control--$t(31) = 7.93$, $p < .01$). Thus, JK accuracy was above chance levels for each group.

As was mentioned earlier, the C.A.Q. measure is relatively new and untested. Thus, although the theoretical basis for the C.A.Q. was preferred over other measures reported in the literature, several alternative measures were also examined. The signal detection sensitivity measure, d', the Probability of a Hit corrected for guessing (P(JK), see Formula 1), and the JK-Errors measure (King, Note 1) were computed and these means are displayed in Table 7. As can be seen, the pattern of JK performance as measured using these alternative techniques is essentially the same as that obtained with the C.A.Q. measure. Results of analyses of variance supported this conclusion. (The ANOVA tables are contained in Appendix D.)

The intercorrelations among the four JK measures and the number of correctly recalled items are contained in Table 8. For this analysis, correlations were computed separately for each group and for each list. In order to summarize the large number of correlations that resulted, Table 8 contains the mean of the nine individual correlations (3 groups times 3 lists). (The complete matrix of correlations is contained in Appendix E.) While all of the correlations between JK accuracy scores were different from zero, it can be seen that the correlations between JK-Errors and the other accuracy measures were generally lower than the

Table 7

Mean d', P(JK) and JK-Errors for

Each Group Across Lists

d'	List 1	List 2	List 3
Varied Ease	1.83	2.30	1.20
Varied Frequency	1.35	1.64	1.45
Control	1.06	1.08	1.05
P(JK)			
Varied Ease	.50	.65	.34
Varied Frequency	.37	.46	.39
Control	.31	.30	.32
JK-Errors			
Varied Ease	5.16	3.66	7.03
Varied Frequency	6.97	5.91	6.31
Control	6.81	6.72	6.13

Table 8

Summary of Correlations Among JK Accuracy Measures and Recall*

	d'	P(JK)	JK-Errors	Recall
C.A.Q.	.80	.80	-.59	-.09
d'		.93	-.63	-.01
P(JK)			-.62	.01
JK-Errors				-.04

* Entries are means of the correlations computed separately for each group and for each list within a group. Correlations are based on 32 pairs; a coefficient of .35 is different from zero, $p < .05$.

intercorrelations between the C.A.Q., d', and P(JK) scores. Since the JK-Errors measure is based on absolute numbers of JK Misses and False Alarms while the other measures take the subjects' distribution of YES and NO JK ratings into account, this pattern can be expected.

An appropriate measure of JK accuracy should not be related to the level of recall. That is, if the JK measure is an accurate reflection of the ability to predict nonrecall as well as recall, then subjects' accuracy scores should not be correlated with recall. As can be seen in Table 8, none of the measures of JK accuracy was correlated with recall. When the correlations are examined separately for each list, it can be seen that for the third list the recall-JK accuracy correlations were slightly more negative than for the first two lists. No explanation is readily available for this slight change in relationship across lists.

JK response bias. It has been proposed that the JK discrimination is made on the basis of the perceived differences between recallable and nonrecallable items along some available dimension. Presumably, the relative magnitude of the differences among items must reach some criterion level before a YES or NO JK rating is assigned. In order to describe the changes in response bias or criterion across the three lists, it was necessary to examine the relative distribution of YES and NO JK ratings. Specifically, the instances where JK ratings were incorrect were analysed. According to Underwood (1974) the relative number of Misses and False Alarms provides an indication of response bias in an absolute judgment recognition task. This approach was adapted to the JK paradigm. The formula for Response Bias is simply:

$$\text{Response Bias} = \frac{(\text{Misses} - \text{False Alarms})}{(\text{Misses} + \text{False Alarms})}$$

The values of this measure range from +1.0 to -1.0. A high positive value indicates a very strict criterion (i.e., relatively few False Alarms). That is, in this case, the subject would be quite sure that recall would follow before responding YES on the JK scale. A high negative value indicates a relatively lax criterion. In this instance, the subject would too readily report that recall would follow, and many False Alarms would result. A response bias of zero results when the number of Misses and False Alarms are about the same.

The mean Response Bias scores for each group across the three lists is contained in Table 9. An analysis of variance revealed that the Response Bias changed across the three lists, $F(2, 186) = 28.65$, $p < .001$. Planned comparisons revealed that the response bias scores for the first list were significantly lower than for the two subsequent lists, $F(1, 186) = 57.34$, $p < .001$. Thus, for each group, a relatively lax criterion was employed for the first list, and many False Alarms resulted. For the second and third lists, a much more strict criterion was observed, and Misses became more likely than False Alarms. The main effect for Groups and the Groups by Lists interaction were not significant, $Fs < 1.0$.

The correlations between Response Bias scores, C.A.Q. scores, and recall were also examined. The correlations for each list and the overall correlations are contained in Table 10. First, note that overall, there is no correlation between C.A.Q. scores and Response

Table 9

Mean JK Response Bias

	List 1	List 2	List 3
Varied Ease	-.23	.27	.19
Varied Frequency	-.24	.33	.34
Control	-.23	.10	.22
Mean	-.23	.23	.25

Table 10

Response Bias, JK Accuracy and Recall Correlations

Overall	Response Bias	Recall
C.A.Q.	-.01	-.02
Response Bias		.56*
List 1		
C.A.Q.	-.18*	-.01
Response Bias		.54*
List 2		
C.A.Q.	.20*	.19*
Response Bias		.62*
List 3		
C.A.Q.	-.05	-.23*
Response Bias		.44*

* A coefficient of this magnitude is greater than zero, $p < .05$.

Bias scores. This should be anticipated if the C.A.Q. measure of sensitivity is, in fact, independent of Response Bias. Also, note that overall (as well as for each list) Response Bias was positively correlated with recall. This was also as expected. As the relative number of Misses increases, so must the level of recall. By definition, a Miss is a correctly recalled item.

When the C.A.Q.-Response Bias correlations are examined for each list separately, it can be seen that for the first list, the C.A.Q.-Response Bias correlation was slightly negative. For the second list, however, the correlation was slightly positive. In other words, accurate JK performance was accompanied by the tendency to adopt a lax criterion on the first list; but on the second list, accurate JK performance was accompanied by the use of a relatively strict criterion. For the third list, the Response Bias was not related to JK accuracy.

From this pattern of results it is apparent that, first, there was a criterion shift (from lax to strict) between the first and second lists. Perhaps familiarity with task demands could account for this tendency to become more cautious in assigning JK ratings. Second, the correlations suggest that those subjects making accurate JKs were likely to adopt a more strict response criterion as task experience increased. On the third list, JK task conditions and JK performance changed and thus it is difficult to make any conclusions about the relationship between accuracy and Response Bias.

Ability Tests

In this section, each of the tests designed to measure individual differences in memory ability will be discussed. Then, the relation-

ships between these measures and JK performance will be reviewed.

MRE ability. The first task to be considered is the Memory-for-Remembered-Events (MRE) test that was administered to all three groups immediately after the JK lists were learned. The absolute judgment test was very similar to the JK task and was scored by constructing a four-fold response matrix for each subject. Since the learners' task was to report "Recalled" or "Not Recalled" for each of the pairs that were in fact either recalled or not recalled, four outcomes were possible. As in the JK task, a Hit and a Correct Rejection occurred whenever the MRE response matched actual recall performance. A Miss occurred when the subject reported that recall was not successful when in fact it was. A False Alarm resulted when the subject incorrectly reported that recall was successful. The formulas contained in Table 2 were employed to convert the absolute frequencies of Hits and False Alarms into probabilities. It was not possible to employ the C.A.Q. measure of accuracy for the MRE task because confidence judgments were not collected. Hence, both d' and the Probability Correct (corrected for guessing, Formula 1) were computed. In order to maintain consistency with the Gardiner and Klee (1976) MRE research d' was selected as the primary dependent measure in the analyses to be reported. It should be noted that the correlation between d' and the Probability Correct was quite high (r = .93).

Table 11 contains the mean d' for MRE performance as a function of the list within which the judged item was learned. Overall, the performance was quite accurate. The average d' across all subjects was 3.29. For comparison, none of the d' means observed for JK

Table 11

Mean d' for MRE Task

	List 1	List 2	List 3
Varied Ease	3.09	3.50	3.84
Varied Frequency	3.09	3.22	3.72
Control	3.08	2.85	3.23
Mean	3.08	3.19	3.60

performance exceeded 2.50. An analysis of variance revealed that MRE accuracy increased across lists, $F(2, 186) = 7.29$, $p < .001$. As would be expected, MRE accuracy increased as the interval between recall attempts and the MRE judgments decreased. Further, although the Lists by Groups interaction was only marginally significant, $F(4, 186) = 1.94$, $p < .10$. Newman-Keuls tests revealed that for both the second and third list items, the Varied Ease and Varied Frequency groups made more accurate MRE judgments than the Control group.

In order to more fully understand the MRE performance, the accuracy of the responses was computed as a function of the various item types for the Varied Ease and Varied Frequency groups. For the Varied Ease group, the mean number of subjects making correct MRE responses for each easy and difficult item was 30.95 and 29.35 respectively. For the Varied Frequency group, the mean number of correct MRE responses for 1-p items was 28.45, and for 3-p items the mean was 29.05. Thus, the accuracy of MRE was not influenced by the task or stimulus manipulations that were present during learning. Since the MRE task has been likened to a situational frequency judgment task, these results were unexpected. However, due to the very low error rates, perhaps a ceiling effect was masking the true influence of item differences and presentation frequencies on MRE performance.

Situational frequency discrimination ability. The first task administered during the second session was the Situational Frequency Discrimination (SFD) task. Only subjects from the Varied Ease and Varied Frequency groups completed this task and the remaining tasks to be considered. Table 12 contains the probability of a correct

Table 12

Probability Correct Situational Frequency Discrimination

		Frequency of Correct Pair Member				
		1	2	3	4	5
	0	.75	.88	.87	.99	.95
Frequency of	1		.71	.75	.86	.96
Incorrect	2			.79	.69	.80
Pair Member	3				.48	.69
	4					.52

judgment for each of the pairings of items presented at the six frequency levels. As can be seen, performance followed the expected pattern. As one moves from left to right across the table, a greater proportion of correct responses was observed. Further, as one glances down the table, the pairs become more difficult and the probability of a correct response decreases. Overall, the mean number of correct responses was 23.20 out of a possible 30.00. The SFD performance for the Varied Ease group and the Varied Frequency group did not differ, $t < 1.0$.

Background frequency discrimination ability. The mean number of correct responses on the Background Frequency Discrimination (BFD) test for the Varied Ease and Varied Frequency groups are contained in the second row of Table 13. Performance on the 35-item test did not differ between groups, $t < 1.0$. Overall, the mean number of correct discriminations equaled 25.15 and was greater than chance performance, $t(44) = 17.03$, $p < .01$. An item analysis was performed to further understand the ability to discriminate background frequencies. For each pair the difference in frequency of occurrence (Kucera & Francis, 1967) was computed. The probability of a correct discrimination was found to be unrelated to the absolute magnitude of the difference in background frequency between pair members, $r = .03$. It is very likely that the perceived level of frequency is not a linear function of the actual frequency of occurrence (Shapiro, 1969), and thus the lack of a simple relationship can be expected. The overall quality of test construction was examined by computing the Kuder-Richardson (1937) Formula 20 index of internal consistency. The test was moderately reliable, $r_{xx} = .48$.

Table 13

Performance on Ability Tests

	Varied Ease	Varied Frequency	t
Situational Frequency Discrimination	23.00 (.76)*	23.42 (.78)	< 1.00
Background Frequency Discrimination	25.06 (.72)	25.25 (.72)	< 1.00
Meaningfulness Discrimination	13.95 (.70)	12.46 (.62)	1.66
Ease-of-Learning Discrimination	15.29 (.76)	15.67 (.78)	< 1.00
Ease-of-Learning Correlations	.61	.65	< 1.00
Stimulus Assessment Score	.03	-.02	< 1.00

* Numbers in parentheses represent proportion correct.

Meaningfulness discrimination ability. The Meaningfulness Discrimination (MD) test has little or no precedent in the literature which would provide a basis for comparison. The mean number of correct responses on the 20-item test is displayed in the third row of Table 13. Although the Varied Ease group revealed slightly more accurate performance than the Varied Frequency group, the difference was not significant. Overall performance was greater than chance levels, $t(44) = 6.82$, $p < .001$. From an item analysis it was learned that the probability of a correct discrimination was moderately related to the magnitude of the difference in meaningfulness between pair members, $r = .27$. The Kuder-Richardson index of internal consistency was .63.

Ease-of-Learning performance. Two tests designed to measure the ability to perceive differences in the ease with which items could be learned was administered. The first task was modeled after Underwood's (1966) study of ease-of-learning. Paired associates were presented and subjects were to rate the items on a 9-point scale. Low scale values indicated difficult pairs and high scale values were assigned to easy pairs. An independent group of subjects learned these pairs and thereby provided actual learning scores. Overall, the probability of correct recall was correlated with the means of the subjects' EL ratings, $r = .73$. For comparison, the corresponding correlation observed by Underwood (1966) was somewhat higher, $r = .92$. For the purposes of the individual differences analysis, the dependent measure was defined as the correlation between the subject's EL ratings and the actual learning scores. The mean Ease-of-Learning Correlation (EL-r) for each group is contained in the fifth row of

Table 13. The overall mean (.62) was slightly lower than the corresponding mean correlation reported by Underwood (.71). To assure a normal distribution for the statistical tests to be discussed below, the correlation coefficients were adjusted according to Fisher's r-to-z transformation (Hays, 1973, pp. 662). The means of the correlations did not differ between groups, $t < 1.0$. Finally, it should be noted that the actual correlations were relatively closely clustered around the mean. That is, although the correlations ranged from .27 to .73, all but 17 percent of the coefficients ranged between .40 and .67.

The second test of the ability to perceive ease-of-learning was a two-alternative forced choice task. The mean number of correct Ease-of-Learning Discriminations (EL-D) for each group is contained in the fourth row of Table 13. Again, performance on the 20-item test did not differ between groups. Further, performance was greater than would be expected if subjects were selecting pairs randomly, $t(44) = 15.47$, $p < .01$. The Kuder-Richardson index of internal consistency equalled .49. (Appendix C contains each of the ability tests discussed thus far.)

Relationships between the ability measures--interim discussion. In order to evaluate the quality of the ability tests more adequately and in order to understand the actual abilities being measured, the correlations among the tests were computed. The correlation matrix is contained in Table 14. The following observations can be made.

First, the MRE test was included as a secondary measure of the ability to discriminate between situational frequencies. It can be

Table 14

Correlation Matrix for Memory Ability Tests and Recall

	SFD	BFD	MD	EL-D	EL-r	Recall
MRE	-.06	.10	.12	.09	.06	-.10
SFD		.09	-.04	.16	-.09	.04
BFD			.41*	.36*	-.03	.39*
MD				.38*	.08	.34*
EL-D					.20	.29*
EL-r						.17

---

* Coefficient significantly different from zero, $p < .05$.

Note: MRE = Memory-for-Remembered-Events; SFD = Situational Frequency Discrimination; BFD = Background Frequency Discrimination; MD = Meaningfulness Discrimination; EL-D = Ease-of-Learning Discrimination; EL-r = Ease-of-Learning Correlation.

suggested that the ability to judge which items were recalled involves a discrimination between retrieval event frequencies. Thus, according to prevailing theoretical notions concerning recognition memory performance (Underwood, Zimmerman, & Freund, 1971), MRE performance and SFD performance should be correlated and influenced by similar factors. No such relationship is indicated in the present research. The SFD scores were not correlated with the MRE scores. Also, it was mentioned earlier that item ease (or background frequency) and item presentation frequency had no influence on MRE accuracy. Thus, apparently the relationship between MRE and frequency judgment ability is not a simple one. However, as will be seen below, the SFD test may not have been adequately constructed, and caution is appropriate before any conclusions can be drawn.

The second observation from Table 14 concerns the lack of a statistical relationship between SFD performance and BFD performance. One might expect that performance on these two tests would be correlated because the perception of "frequency" or "familiarity" is common to both tasks. The observed results suggest that the perception of event frequency may involve different processes than the perception of lexical or semantic frequency. However, this suggestion should be qualified by the fact that first, the SFD test was the least reliable test according to the Kuder-Richardson values, and secondly, the SFD scores were not correlated with any other dependent measure. Perhaps the SFD test was not a sufficiently sensitive measure of the true ability to perceive differences in situational frequency.

Several of the tests were designed to measure the ability to

perceive stimulus characteristics. That is, the BFD, MD, EL-D and EL-r tests were constructed to tap various aspects of the general ability to perceive ease-of-learning. To a moderate degree, the correlations in Table 14 support the claim that these tests were measuring the same underlying ability. The strongest correlations in the matrix were observed among the BFD, MD, and EL-D scores. However, the performance on the EL-r task was statistically related to neither the EL-D nor the BFD and MD performance. Further, the EL-r scores were not correlated with recall while the BFD, MD, and EL-D scores were each correlated with recall. Indeed, the EL-r performance was not strongly related to any of the other measures. As was mentioned earlier, perhaps the somewhat limited range in observed EL-r values reduced the likelihood of observing a significant correlation. It is unlikely that these abilities are actually unrelated. The preferred interpretation of the lack of relationship between EL-r and the other measures is that the test was not sensitive to individual differences in the ability to perceive learning ease.

From these results it was concluded that three of the test scores could be statistically combined to produce a meaningful overall measure of Stimulus Assessment ability. The BFD, MD, and EL-D tests were interrelated in the expected manner, and each test exhibited a moderate degree of internal consistency according to the Kuder-Richardson values. To arrive at a combined score, each subject's score on each test was converted to a z-score, and the z-scores were then added together. This procedure assured that each test was equally weighted in the combined total (Brown, 1976, pp. 145). The

mean of these Stimulus Assessment scores for each group is contained in Table 13. As can be seen, performance did not differ between groups, $t < 1.0$. This combined score was employed in the individual differences analyses to be discussed below.

Finally, it should be recalled that the purpose of the ability tests was to provide measures of two general abilities--the perception of situational frequency and the perception of item characteristics. By examining the correlation matrix, it can be seen that these two general abilities appear to be statistically unrelated to one another. That is, the SFD performance was not related to the BFD, MD, EL-D, or EL-r performance. This pattern of results would be expected if the two general abilities were, in fact, orthogonal to one another. However, this conclusion should be made with caution because of the suspected poor quality of the SFD test. The notion of independence of these two abilities is crucial to the individual differences analyses to be discussed below.

Individual differences analysis. For the sake of clarity, the rationale behind the following correlational analyses will be briefly restated. The aim was to test the notion that the processes which contribute to the JK discrimination vary as a function of the particular task and stimulus conditions present during learning. The tests mentioned above were designed to measure the underlying processes which were believed to aid the discrimination between recallable and nonrecallable items. Specifically, the learning conditions for the Varied Ease group were such that the perception of item ease would facilitate accurate JK performance. If this were the case, then

those subjects who performed relatively well on the tests designed to assess ease-of-learning perception should also have made accurate JKs. Therefore, JK accuracy was expected to have been correlated with BFD, MD, EL-D, and EL-r performance. The Varied Frequency group, on the other hand, learned homogeneous item sets, and thus the perception of learning ease should not have been related to JK performance.

The intralist changes in presentation frequency were believed to provide the dimension along which JK discriminations could be made for the Varied Frequency group. Thus, for this group, JK performance should be correlated with SFD performance. Since no such variation in presentation frequency was present for the Varied Ease group, the correlation between SFD and JK accuracy should be zero for these subjects.

An important aspect of the present argument is the requirement that the two general abilities be independent. That is, the ability tasks were designed to isolate two separate memory processes. However, it should be acknowledged that the measured abilities could simply be manifestations of the same underlying ability (e.g., verbal processing ability or verbal intelligence). If performance on the various ability tests were correlated, statements about the specific abilities contributing to JK performance could not be made. To the extent that the ability tests were valid, there appeared to be no evidence that the SFD, BFD, MD, EL-D, and EL-r tests were all measuring one common verbal skill. This, it is appropriate to proceed with the analyses according to the stated plan.

The correlations between JK performance and the ability measures

are contained in Table 15. For the purposes of this analysis, JK accuracy was defined as the mean of the C.A.Q. scores on the first two lists for each subject. List 3 JK accuracy scores were omitted because the relevant task and stimulus manipulations were not present during learning. No changes or transformations were performed on the data with the exception of the EL-r measure. Here, the correlation coefficients were transformed into z-scores. Contrary to expectations, none of the correlations was substantially different from zero. Only the correlations for the EL-D measure followed the expected pattern. The relationship between JK accuracy and EL-D performance was slightly stronger for the Varied Ease group than for the Varied Frequency group. When the combined Stimulus Assessment scores were examined, no statistical relationship with JK accuracy was observed.

Essentially the same pattern of correlations was observed when JK performance on List 1 and List 2 were entered into the analysis separately. Also, Lists 1 and 2 were collapsed to arrive at an overall measure of JK accuracy, and no major differences in results were observed. Furthermore, the appropriate scatter-plots corresponding to these correlations revealed no evidence of curvilinear relationships between measures.

Given these unexpected results, the following additional analyses were performed to isolate the reasons for the lack of statistical relationships. First, as was stated in the preceding section, some doubt was expressed as to the statistical quality of the ability tests. In order to determine if the tests were statistically valid, and to determine if the observed range in test scores was sufficient to allow

Table 15

Correlations Between JK Accuracy and Ability Tests

	Varied Ease (n = 21)	Varied Frequency (n = 24)
Memory for Remembered Events	.09	.10
Situational Frequency Discrimination	-.33	-.17
Background Frequency Discrimination	-.11	.06
Meaningfulness Discrimination	.16	.08
Ease-of-Learning Discrimination	-.10	-.19
Ease-of-Learning Correlation	.08	-.04
Stimulus Assessment Score	.07	.10

correlations to be observed, the correlations between these measures and recall were examined. It could be argued that if the ability tests were shown to be related to recall then it is less likely that some statistical inadequacy of the ability measures was the reason for the nonsignificant results. Table 16 contains the correlations between the ability scores and the number of correctly recalled paired-associates across the first two lists. The correlations were computed for each group and for all subjects combined. In general, performance on three of the ability tests was related to recall. Accurate recall was accompanied by relatively good performance on the BFD, MD, and EL-D tests. The Stimulus Assessment scores were also correlated with recall since this measure is simply a combination of these three test scores. Thus, apparently three of the tests were of sufficient statistical validity to reveal correlations in an expected pattern.

A second possible reason for the ambiguous results could be that the C.A.Q. measures were not reliably reflecting JK ability. Or expressed in another way, regardless of the particular measure, perhaps JK ability as measured in the present research, was not consistent across the JK trials. To examine this possibility, the test-retest reliability of the C.A.Q. scores was computed. The correlations among the accuracy scores for each list are contained in Table 17. The analysis was performed on each group separately as well as for all subjects combined. While some of the correlations were statistically different from zero, there was only a moderate degree of reliability for the C.A.Q. scores. The alternative measure of JK accuracy also failed to reveal acceptable reliability. Therefore, it appears that

Table 16

Correlations Between Recall and Ability Tests

	Varied Ease (n = 21)	Varied Frequency (n = 24)	Overall (n = 45)
Memory for Remembered Events	.23	.20	.17
Situational Frequency Discrimination	.02	.06	.04
Background Frequency Discrimination	.58*	.35*	.39*
Meaningfulness Discrimination	.44*	.24	.34*
Ease-of-Learning Discrimination	.48*	.25	.29*
Ease-of-Learning Correlation	.44*	.13	.17
Stimulus Assessment Score	.46*	.29	.44*

* Coefficient significantly different from zero, $p < .05$.

Table 17

JK Accuracy Interlist Correlations

	List 2 C.A.Q.	List 3 C.A.Q.
Overall		
List 1 C.A.Q.	.27*	-.08
List 2 C.A.Q.		.04
Varied Ease		
List 1 C.A.Q.	.15	.16
List 2 C.A.Q.		.46*
Varied Frequency		
List 1 C.A.Q.	.42*	.08
List 2 C.A.Q.		.10

* Coefficient significantly different from zero, $p < .05$.

the JK accuracy measure may be responsible for the unexpected results for the individual differences analyses.

DISCUSSION

The purpose of the present research was to clarify the processes underlying the ability to judge what will be recalled on a later test of retrieval. The general premise was that there are a variety of cues or dimensions along which recallable items can be discriminated from non-recallable items. According to this proposed multi-dimensional hypothesis, the particular decision axis or cue is determined by the conditions under which the judged information is learned. In the following sections of the discussion, the results will be reviewed and examined in relation to this premise. The weaknesses of the present research as well as the implications for further research will be outlined.

It was predicted that the construction of the study lists in the present paradigm would influence the level of JK accuracy. For the Varied Ease group this prediction was upheld in that JK accuracy was greater when list items varied according to learning ease than when list items were of relatively constant ease. The predictions regarding JK performance for the Varied Frequency group were only tentatively supported. That is, although the statistical differences across lists were not significant, the trend in the data suggested that JK accuracy was slightly greater when presentation frequency of list items was varied than when presentation frequency was held constant. The observed performance for the Control group lent further support to these conclusions. For this group, learning conditions were constant across all lists

and no JK-related dimension was systematically varied. Again, as expected, JK performance under these conditions did not change across lists. Furthermore, the overall level of accuracy was somewhat lower for the Control group than for the other two groups. Thus, the general pattern of results followed the predictions. Before considering these results in relation to a theoretical description of the JK process, some specific aspects of the JK paradigm as presently defined should be examined.

JK accuracy and recall level. The design of the paired-associate lists used in the present experiment was intended to allow for opportunities to predict nonrecall as well as recall. The lists had to be of sufficient difficulty such that recall would not be perfect. Although data from several subjects were discarded because recall was either too high or too low, the resulting recall performance was near the expected 50% correct level. It can be suggested that this prevented the subjects from making JKs on a "list" basis. That is, it is not likely that learners found all the items in a list to be so readily recallable (or so extremely difficult) that a strategy of judging groups of items or all items as recallable would be adopted. The purpose of the task was not to inquire about list difficulty. Rather, items should have been judged in isolation. The aim of the JK was to force subjects to discriminate between recallable and nonrecallable items. In principle, the paired-associate task demands that the subjects process one item at a time. Also, given that recall was less than perfect, the learning task was appropriate for the present intent.

A second important issue involving recall level concerns the

relationship between recall level and JK accuracy. There was no apparent correlation between the pattern of recall and the pattern of JK accuracy across lists. Increases or decreases in correct recall were not accompanied by systematic increases or decreases in JK accuracy. Further, when JK accuracy remained constant, the recall level clearly changed (e.g., Control group). Again, given the intent of the JK paradigm, this is as expected. This is important in that one would be suspect of the validity of the JK accuracy measure if a strong correlation was observed with recall level. The purpose of the JK task is to allow correct predictions of nonrecall as well as correct predictions of recall. The lack of a correlation between JK accuracy and recall suggests that, to some degree, the accuracy measure is truly reflecting the ability to predict nonrecall.

A final observation concerning JK accuracy and recall level is of theoretical interest and should be the subject of future research. Throughout discussions of JK ability, it has been suggested that accurate predictions of recall would mean that learners could efficiently allocate study time and thus raise overall recall scores. Subjects would know that some part of the to-be-learned material was sufficiently learned and that other parts required more effort to assure later retrieval. Although the intent of the present research was not to show that learners could use accurate JKs to the benefit of later retention, one might have expected that superior JK performance would be followed by increasingly accurate recall on later lists. This would be a tentative demonstration that learners acquired a transferable skill based on the ability to efficiently allocate study time. Upon first glance,

the recall performance of the Varied Ease group followed this prediction. Recall increased on subsequent lists after relatively accurate JK performance on the first list. However, no such relationship was observed for the other two groups. Also, examination of the correlations between JK performance and subsequent recall performance revealed no evidence for such a relationship. Thus, the theoretically appealing notion that accurate prediction of recall leads to the use of efficient study behaviors awaits further support.

The C.A.Q. measure. As was mentioned earlier in this paper, there is some concern over the technique used to measure JK accuracy. The C.A.Q. measure is new, and from the observed results, there is little reason to doubt that it is a satisfactory measure. The correlations revealed strong relationships between the C.A.Q. and other JK measures. Also, the separate analyses of the variability in the JK ratings (denominator of the C.A.Q.) and the mean difference in ratings assigned to recalled and nonrecalled items (numerator of the C.A.Q.) lead to the conclusion that a true change in discriminability among list items was responsible for the changes in C.A.Q. scores. According to these analyses, it was not likely that a tendency to artificially restrict or increase the range in ratings was causing the changes in C.A.Q. scores. These observations lead one to accept the C.A.Q. measure of JK accuracy without reservation.

Theoretical Implications

The purpose of the following sections is to provide a more critical examination of the theoretical implications of the JK results. Special attention will be directed toward specific group differences

and the unexpected findings of the individual differences analysis.

JKs and the perception of learning ease. There have been several references in the JK literature to the link between JKs and the ability to judge the ease with which the given material may be learned. Arbuckle and Cuddy (1969) reported that the probability of predicting correct recall decreased as the judged difficulty of the rated item increased. King (Note 1) and Pasko (Note 2) observed a similar relationship between ease-of-learning ratings and JK ratings. Although these suggestions are theoretically appealing, the above mentioned studies were not designed to provide a direct test of the link between JKs and ease-of-learning perception. Whenever JKs and ease-of-learning ratings are made on the same items a statistical relationship must be observed if either set of ratings is said to be accurate. An easy item will have a high probability of being correctly recalled, and an accurate JK will, by definition, indicate prediction of correct recall. The present study was designed to provide an alternative technique for examining this link. The focus was not on the similarity of assigned ratings, but rather accuracy of the ratings was of central concern. The logic was as follows. If the assessment of learning ease is central to the JK process, then by emphasizing a priori variations in learning ease within a list, the JK task should become easier than if no such cue is present. Hence, JK accuracy should be greater when this dimension is made salient than when all list items are of relatively constant ease. The logic of the design was extended in order to demonstrate that the perception of learning ease is merely one of several processes by which JKs are made. That is, it was expected that other dimensions related to the probability of

recall could also influence the level of JK accuracy. Therefore, it may be said that under certain conditions, processes such as the perception of presentation frequency may play a central role in the JK process.

In general, the manipulations of list construction had the expected effect on JK accuracy. However, one unexpected aspect of the results deserves close attention in light of the JK-ease-of-learning link. That is, why was the JK performance of the Varied Ease group superior to that of the Varied Frequency group? Why did the manipulation of learning ease have a greater impact on JK performance than the manipulation of presentation frequency? Both dimensions are related to the probability of recall and it was expected that both would have an (equal) effect on JK accuracy.

Given the preceding discussion, one may be tempted to immediately conclude that the results of the present study support the notion that the JK process is closely dependent on the ability to perceive learning ease. Perhaps the demands of the JK task are such that the learners' attention is drawn to the characteristics of the judged item more readily than it is drawn to contextual factors such as presentation frequency.

Before this conclusion is accepted, a second explanation must be considered. As was briefly mentioned in the results section of this paper, the proportion correct recall for each item type was computed for the Varied Ease and Varied Frequency groups. The easy items were recalled more frequently than the difficult items, and the 3-p items were recalled more frequently than the 1-p items. These results are

not at all surprising. However, what is noteworthy is that the relative difference in recall between easy and difficult items was greater than the difference between 3-p and 1-p items. Thus, although in principle, both dimensions did influence recall levels, the variability in learning ease may have been a more extreme or salient cue than the variation in presentation frequency. Caution is in order when discussing the relative effect of the two cues because, statistically, both dimensions had a clear influence on recall. The relative difference may suggest that the perceptibility of the two dimensions was not equal under the present learning conditions. Furthermore, it is difficult to determine whether these two distinct dimensions could ever be made "equivalent". Thus, any unqualified claim that ease-of-learning is inherently more closely related to the JK process than contextual factors is not warranted. Given this point, can any statement be made about the relative importance of the two experimental variables under discussion? The question remains; why was the JK performance of the Varied Ease group greater than that of the Varied Frequency group? To arrive at an answer, it is valuable to refer to the body of available evidence concerning JKs.

JKs as mediated decisions. Throughout this paper, the aim has been to demonstrate that as learning conditions change, various cues may serve as aids to the JK decision. From the available evidence, it can be stated that two very general classes of dimensions can be outlined. First, the characteristics of the to-be-learned material can influence the magnitude of the JK ratings and the accuracy of the JK performance. Arbuckle and Cuddy (1969), King (Note 1), and Pasko (Note 2)

have shown that assigned JK ratings are directly related to the perceived ease of the information. While not providing a direct test of the dependence of JK ability on the ability to make ease-of-learning ratings, these authors have suggested the importance of this link. The present study provided a direct test of the influence of variations in learning ease on JK accuracy. These studies point to the importance of what King (Note 1) and Pasko (Note 2) have termed "stimulus knowledge". Experienced learners bring to the JK task some understanding of the item characteristics which determine learning ease.

A second line of evidence also emerges from the JK studies. Task-specific cues have been shown to affect JK performance. Zechmeister and Shaughnessy (Note 4) demonstrated that item presentation frequency and the spacing of repetitions can influence the absolute magnitude of the JK ratings. From the present study it was seen that variations in presentation frequency can lead to slightly improved JK performance. Also, the presence of test trials has been shown to have a positive influence on JK accuracy. From these findings, it is apparent that the learning context in which the JKs are made, regardless of a priori item differences, can provide useful cues in making JK discriminations.

The demonstration that a variety of cues under a variety of learning circumstances can affect the JK suggests a framework within which the process may be further analysed. That is, the JK is best seen as a judgment which is dependent on the perception of cues which serve to mediate the discrimination. No one dimension has been isolated that can account for all the observed JK results. It is likely that further research will demonstrate the role of additional

cues such as rehearsal patterns (Rundus, 1971) or partial attribute recall (Blake, 1973). The theoretical thinking about the JK should not be limited to only those specific aspects of the learning task that have been shown to be related to JK accuracy.

Given this conclusion, can any explanation be given as to why the Varied Ease group made more accurate JKs than the Varied Frequency group in the present experiment? Since both types of cues, item-specific and task-specific, have been implicated in the JK process, the observed difference in JK accuracy between the two groups may suggest a "hierarchy" of cues. Perhaps the learners' attention is focused on item characteristics initially, and only if these cues are unavailable will attention be paid to task-specific variations. Consider the following real-world analogy. Suppose a student is asked to judge on which of two upcoming classroom tests he will do better--English Literature or Physics. Regardless of the amount of time each was studied or the conditions under which each was studied, the student may respond that he will do better on the English Literature test because "English is easier than Physics". Perhaps only if the discrimination cannot be made on the basis of "ease" will other factors be considered. The secondary cues (such as study time) may be just as informative to the student, but these factors may not be immediately considered. It is likely that further research will demonstrate this hierarchy of cues useful to the JK discrimination.

JKs and MRE ability. A secondary concern of the present research was directed at the ability to monitor past performance. The King (Note 1) study had clearly pointed to the importance of this ability.

As a more sophisticated understanding of the JK process is attained, it is likely that JKs and MRE judgments will be shown to be manifestations of similar underlying processes. In the present study, the intra-list manipulation of presentation frequency and the intra-list variation in learning ease influenced MRE performance and JK accuracy in much the same manner. For second list items, the Varied Ease group and the Varied Frequency group produced more accurate MRE scores than the Control group. Furthermore, the hypothesized commonality between situational frequency discrimination and MRE performance received no support. Presentation frequency and background frequency (i.e., ease) did not influence MRE accuracy. Also, MRE performance was not correlated with situational frequency discrimination ability. Although considerable caution is in order because of the suspected statistical insensitivity of the frequency discrimination test, it is likely that MRE ability cannot be simply likened to recognition memory ability. From the between-groups comparison, it might be suggested that the variation within lists may have provided a cue for MRE discriminations in much the same way that the variations aided the JK. For example, perhaps the MRE judgments were accomplished by learners reasoning that "it was easy therefore I probably got it right". Again, further research may lead to the conclusion that JK ability and MRE ability have much in common.

JK and memory abilities. One intention of the present research was to demonstrate the relationship between various memory abilities and JK ability. The individual differences tests were designed to isolate and measure selective memory abilities. A few encouraging results

emerged from this effort. First, it had been suggested by Underwood (1966) and Lippman and Kintz (1968) that ease-of-learning perception was strongly related to the perception of item pronunciability, meaningfulness and frequency. The present study was designed to avoid the logical flaw that occurs whenever the same items are subjected to several different types of ratings. The design focused on the accuracy of background frequency judgments, meaningfulness discriminations, and ease-of-learning ratings as measured independently. Although the quality of these tests is not beyond criticism, the general pattern in the correlations suggested that those individuals who perform well on ease-of-learning rating tasks can also accurately judge the relative frequency with which an item occurs in the language and can accurately assess the ease with which associations can be generated to words. Thus, there appears to be further evidence that a common element is present in these tasks. It is likely that Underwood's (1966) speculation was correct concerning the relation of perceived ease to other verbal characteristics.

A second encouraging finding in the present research was that recall performance was moderately related to "stimulus assessment" ability. Those subjects who made accurate background frequency, meaningfulness, and ease-of-learning discriminations also tended to produce superior recall scores. Because of this finding, the use of "stimulus assessment" ability as an explanatory tool in future research can be anticipated. Apparently, the ability has some construct validity.

Finally, no convincing evidence was presented which linked specific

memory abilities with JK performance under the conditions of the present study. Although the experimental manipulations had the desired effect on JK accuracy, the individual differences analyses produced ambiguous results. The design was such that the Varied Ease group's performance was expected to be correlated with performance on the stimulus assessment tests. The Varied Frequency group's JK performance was expected to be related to situational frequency discrimination ability. It was concluded that the JK scores did not reflect sufficient statistical reliability to reveal the desired correlational pattern. Furthermore, it should be acknowledged that the design of the present study may have been too "optimistic". The JK task is relatively new and apparently it is not amenable to such specific analysis given our current level of understanding. These disappointing correlational results should not be interpreted to mean that the direction of the thinking was inappropriate. Rather, it is likely that the necessary psychometric control has not been achieved for the JK paradigm. Also, in light of recent individual differences analyses reported by Hunt et al. (1975) and Hogaboam and Pellegrino (1978), the technique itself should not be judged as inappropriate for the examination of cognitive processes involved in experimental tasks such as the JK task.

General Conclusions

Overall, the present research lends support to the so-called "multi-dimensional" view of the JK process. The observations point to the conclusion that the ability to judge what will or will not be recalled is dependent on or mediated by the ability to perceive various cues present in the learning task. Some of the cues may be item-specific

and thus dependent on previous experience with various types of verbal material. Other cues may be task-specific and are dependent on familiarity with certain task demands. Taken together, the learners' understanding of and ability to use cues for this purpose can be said to be part of what has been called "metamemory" (Flavell & Wellman, 1977), or the general knowledge of one's memory ability that is a sign of a well-developed memory system.

The emphasis throughout this paper has been on the inter-dependence of memory abilities. The study of learners' monitoring of their memories has made reference to processes which have been the subject of considerable research efforts. The perception of situational frequency and the perception of item characteristics can assume a new role as processes closely related to the monitoring of one's memory. It is likely that future research will demonstrate how other memory abilities can be called upon to achieve accurate JKs under different conditions. This general framework emphasized the interplay of verbal abilities and is consistent with the current belief that the learner is an active processor of information. Memory monitoring ability must be seen as an integral part of the entire system.

REFERENCE NOTES

1. King, J. F. Judgments of knowing: a learning-to-learn analysis. Unpublished master's thesis, Loyola University, 1976.

2. Pasko, S. J. Judgments of knowing relative to stimulus knowledge, practice, and individual differences. Unpublished doctoral dissertation, Loyola University, 1977.

3. Lovelace, E. A. Prediction during learning of later retrieveability. Paper presented at the 15th Annual Meeting of the Psychonomic Society, Boston, 1974.

4. Zechmeister, E. B., & Shaughnessy, J. J. When you know that you know and when you think that you know but you don't. Paper presented at the 15th Annual Meeting of the Psychonomic Society, Boston, 1974.

5. Egan, J. P. Recognition memory and the operating characteristic. Indiana University Hearing and Communication Laboratory Technical Report, 1958.

6. Shaughnessy, J. J. Confidence-judgment accuracy as a predictor of test performance. Manuscript in preparation.

REFERENCES

Arbuckle, Y. Y., & Cuddy, L. L. Discrimination of item strength at time of presentation. Journal of Experimental Psychology, 1969, 81, 126-131.

Archer, E. J. A re-evaluation of the meaningfulness of all possible CVC trigrams. Psychological Monographs, 1960, 60, 216-221.

Atkinson, R. C., & Shiffrin, R. M. Human memory: a proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.) Advances in the psychology of learning and motivation. Vol 2, New York: Academic Press, 1968.

Battig, W. F. Scaled difficulty of nonsense-syllable pairs consisting of syllables of equal association value. Psychological Reports, 1959, 5, 126.

Battig, W. F. Comparison of two methods of scaling nonsense-syllable pairs for ease-of-learning. Psychological Reports, 1960, 6, 363-366.

Blake, M. Prediction of recognition when recall fails: Exploring the feeling-of-knowing phenomenon. Journal of Verbal Learning and Verbal Behavior, 1973, 12, 311-319.

Brown, F. G. Principles of educational and psychological testing. New York: Holt, Rinehart and Winston, 1976.

Carrol, J. B. Measurement properties of subjective magnitude estimates of word frequency. Journal of Verbal Learning and Verbal Behavior,

1971, 10, 722-729.

Craik, F. I. M., & Lockhart, R. A. Levels of processing: a framework for memory research. Journal of Verbal Learning and Verbal Behavior, 1972, 11, 761-784.

Craik, F. I. M., & Tulving, E. Depth of processing and the retention of words in episodic memory. Journal of Experimental Psychology: General, 1975, 104, 268-294.

Flavell, J. H., & Wellman, H. M. Metamemory. In R. V. Kail, & J. W. Hagan (Eds.), Perspectives on the development of memory and cognition. Hillsdale, NJ: Lawrence Erlbaum Associates, 1977.

Gardiner, J. M., & Klee, H. Memory-for-remembered-events: as assessment of output monitoring in free recall. Journal of Verbal Learning and Verbal Behavior, 1977, 16, 227-233.

Gardiner, J. M., Passmore, C., Herriot, P., & Klee, H. Memory-for-remembered-events: effects of response mode and response-produced feedback. Journal of Verbal Learning and Verbal Behavior, 1977, 16, 45-54.

Green, D. M., & Swets, J. A. Signal detection theory and psychophysics. New York: Wiley, 1966.

Groninger, L. D. Predicting recognition during storage: the capacity of the memory system to evaluate itself. Bulletin of the Psychonomic Society, 1976, 7, 425-428.

Hall, J. T. Verbal learning and retention. New York: J. B. Lippincott Co., 1971.

Hart, J. F. Memory and the feeling-of-knowing experience. Journal of Educational Psychology, 1965, 57, 347-349.

Hays, W. L. Statistics for the social sciences. New York: Holt, Rinehart and Winston, 1973.

Hintzman, D. L. Apparent frequency as a function of frequency and the spacing of repetitions. Journal of Experimental Psychology, 1969, 80, 139-145.

Hintzman, D. L. Theoretical implications of the spacing effect. In R. L. Solso (Ed.), Theories of cognitive psychology: the Loyola symposium. Hillsdale, NJ: Lawrence Erlbaum Associates, 1974.

Hintzman, D. L., & Block, R. A. Repetition and memory: Evidence for a multiple trace hypothesis. Journal of Experimental Psychology, 1971, 88, 297-306.

Hochhaus, L. A table for the calculation of d' and beta. Psychological Bulletin, 1972, 77, 375-376.

Hogaboam, T. W., & Pelligrino, J. W. Hunting for individual differences in cognitive processes: Verbal ability and semantic processing of pictures and words. Memory & Cognition, 1978, 6, 189-193.

Hunt, E., Lunneborg, C., & Lewis, J. What does it mean to be a high verbal? Cognitive Psychology, 1975, 7, 194-227.

Jenkins, J. J. Can we have a meaningful theory of memory? In R. L. Solso (Ed.), Theories of cognitive psychology: the Loyola symposium. Hillsdale, NJ: Lawrence Erlbaum Associates, 1974.

Kintsch, W. Learning, memory and conceptual processes. New York: Wiley, 1970.

Kreutzer, M. A., Leonard, C., & Flavell, J. H. An interview study of children's knowledge about memory. Monographs of the Society for Research in Child Development, 1975, 40(No. 1).

Kucera, H., & Francis, W. N. Computational analysis of present-day American English. Providence, RI: Brown University Press, 1967.

LaPorte, R. E., & Nath, R. Role of performance goals in prose learning. Journal of Educational Psychology, 1976, 68. 220-264.

Lippman, L. G., & Kintz, B. L. Group predictions of item differences of CVC trigrams. Psychonomic Science, 1968, 12, 265-266.

Masur, E. F., McIntyre, C. W., & Flavell, J. H. Developmental changes in apportionment of study time among items in a multi-trial free recall task. Journal of Experimental Child Psychology, 1973, 15, 237-246.

Paivio, A., Yuille, J. C., & Madigan, S. A. Concreteness, imagery, and meaningfulness values for 925 nouns. Journal of Experimental Psychology Monographs, 1968, 76(1, Pt. 2).

Richardson, J., & Erlebacher, A. Associative connection between paired verbal items. Journal of Experimental Psychology, 1958, 56, 62-69.

Schulman, A. I. Memory for words recently classified. Memory & Cognition, 1974, 2, 47-52.

Shapiro, B. J. The subjective estimation of relative word frequency. Journal of Verbal Learning and Verbal Behavior, 1969, 8, 248-251.

Shaughnessy, J. J., Zimmerman, J., & Underwood, B. J. Further evidence of the MP-DP effect in free recall learning. Journal of Verbal Learning and Verbal Behavior, 1972, 11, 1-12.

Spreen, O., & Schulz, R. W. Parameters of abstraction, meaningfulness, and pronunciability for 329 nouns. Journal of Verbal Learning and Verbal Behavior, 1966, 5, 459-468.

Thorndike, E. L., & Lorge, I. The teachers' word book of 30,000 words.

New York: Columbia University Press, 1944.

Tulving, E., & Madigan, S. A. Memory and verbal learning. Annual Review of Psychology, 1970, 21, 437-484.

Underwood, B. J. Individual and group predictions of item difficulty for free learning. Journal of Experimental Psychology, 1966, 71, 673-679.

Underwood, B. J. The role of the association in recognition memory. Journal of Experimental Psychology Monograph, 1974, 102(No. 5).

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

Underwood, B. J., Zimmerman, J., & Freund, J. S. Retention of frequency information with observations on recognition and recall. Journal of Experimental Psychology, 1971, 87, 149-162.

Zacks, R. T. Invariance of total learning time under different conditions of practice. Journal of Experimental Psychology, 1969, 82, 441-447.

Zimmerman, J., Broder, P. K., Shaughnessy, J. J., & Underwood, B. J. A recognition test of vocabulary using signal detection measures and some correlates of word and nonword recognition. Intelligence, 1977, 1, 5-31.

APPENDIX A

APPENDIX A

Computation of C.A.Q.

The purpose of this additional comment about measures of JK accuracy is to clarify the method used to compute the C.A.Q. scores and to discuss briefly some alternative methods. The C.A.Q. formula is:

$$\text{C.A.Q.} = \frac{\bar{X}_{JK_R} - \bar{X}_{JK_{\bar{R}}}}{\sqrt{s_R^2 + s_{\bar{R}}^2}}$$

JK_R refers to the JK ratings assigned to recalled items. $JK_{\bar{R}}$ refers to JK ratings assigned to nonrecalled items. The numerator is straightforward and requires no special justification. The denominator, on the other hand, may require special attention. The computational formula for the pooled variance is:

$$\text{Pooled Variance} = \frac{\left[JK_R^2 - \frac{(JK_R)^2}{\#\text{ recalled}}\right] + \left[JK_{\bar{R}}^2 - \frac{(JK_{\bar{R}})^2}{\#\text{ not recalled}}\right]}{\#\text{ items} - 2}$$

The denominator is simply the square root of the above value.

It can be suggested that the standard deviation of the JK ratings, considered as one group, may provide an equally appropriate measure of the variability of the JKs. Several reasons can be given for not using the

simpler standard deviation. First, the purpose of the accuracy measure is to reflect the "distance" between the means of the two distributions. The mean of the ratings assigned to recalled items is to be compared with the mean of the ratings given to nonrecalled items. So, in theory, the pooled estimate of variability is more closely tied with the intent of the measure than is the standard deviation. Second, the two distributions of JK ratings may not be of equal variance or skewness. That is, perhaps the ratings given to recalled items would be negatively skewed while the ratings given to nonrecalled items would be positively skewed. Furthermore, because of varying opportunities for recall and nonrecall, the shapes and variances of the two distributions may change independent of one another.

The actual distributions from the present JK task can be used to illustrate. Collapsing across all subjects and all lists, the following distribution of JK ratings was observed.

	1	2	3	4	5	6
Frequency	605	925	1133	1247	963	887

As can be seen, the overall distribution is roughly normal and each of the six ratings was utilized a substantial number of times. Now, notice how the distributions change when each group is considered separately, and when the JKs given to recalled items are distinguished from the JKs given to nonrecalled items.

		1	2	3	4	5	6
Control	Recalled	42	98	184	282	231	220
	Nonrecalled	99	228	231	172	68	65
Varied Frequency							
	Recalled	48	112	180	208	244	210
	Nonrecalled	174	213	213	155	64	55
Varied Ease							
	Recalled	67	85	142	281	289	296
	Nonrecalled	175	189	183	105	67	41

The larger JK ratings were more frequently assigned to recalled items than to nonrecalled items. Furthermore, this is most clearly seen for the Varied Ease group. This should be expected since this group made the most accurate predictions. Also, note how the distribution for nonrecalled items is positively skewed. This is also what should be expected.

Although the above distributions are for grouped data, the same shift in the shape of the distributions should be observed for individuals' JK responses. For this reason, it is preferable to employ a measure of variability which does not ignore this difference in the distributions for recalled and nonrecalled items. It can be argued that the simple standard deviation of all the JKs would not be sensitive to this subtle difference.

In order to reach a complete understanding of the behavior of this dependent measure, the C.A.Q. was computed using the simple standard deviation. In fact, there was very little difference between the results using this technique and the results using the pooled variance estimate. The correlations between the two dependent measures for each group and for

each list are presented below.

	List 1	List 2	List 3
Control	.98	.99	.97
Varied Frequency	.92	.93	.97
Varied Ease	.97	.97	.96

As can readily be seen, the selection of one measure of variability over the other makes very little difference in the present study. However, it is likely that in future studies more extreme changes in the distributions of recalled and nonrecalled items may be observed. If, for example, subjects restricted their use of the JK scale to one or two values, there might be a greater disparity between the two methods of computation. This was not the case in the present experiment, but it can be argued that the dependent measure selected should be as widely applicable as possible. The use of the pooled variance is preferred for this reason.

APPENDIX B

APPENDIX B

Study Lists

Homogeneous Paired-Associate Lists

SEN - slipper	BOS - unit	NIC - capture
LUK - deceit	CUD - speaker	FOD - potent
TAK - pepper	MAS - damsel	TOL - salute
FET - northwest	GEN - portrait	LIS - conquest
NAW - chaos	SIG - painter	DER - scarlet
BEK - welfare	DIS - revolt	SOY - steerage
CUZ - sulphur	LOP - comrade	GIP - mercy
MOR - circus	TUN - hatred	MUF - daylight
GOB - mantle	FUP - odor	CAD - session
YAC - kindness	JEF - humor	BAM - vapor
PIB - baron	NEL - builder	ROS - tower
WIS - panic	REM - vigor	JUT - friction
KUP - decree	XAP - buffoon	XEN - malice
VEL - monarch	QIK - boredom	QAD - hindrance
DAR - reflex	HOL - madness	HAZ - lecture
HUR - item	VAC - sickness	VEX - menace
QIL - fatigue	KAP - forehead	KED - elbow
XIT - nephew	WAT - moisture	WEL - rosin
JOL - hardship	POX - background	PAM - assault
ROG - limelight	YAH - folly	YUM - instance

Heterogeneous Paired-Associate Lists

FIZ - sugar	HAP - disease
PIJ - surtax	KYV - garret
GOV - table	FEL - forrest
MEJ - gadfly	GEX - foible
DEM - ticket	VAL - market
FIQ - essence	XUR - excuse
KIX - weapon	DOL - palace
RYW - preview	NIJ - savant
PED - apple	ROL - paper
VOF - fatigue	XOV - concept
BOR - animal	WIM - season
JIQ - abbess	QUJ - adage
PAS - bottle	BAW - river
QIH - debacle	JYK - henchman
YEL - baby	SUP - prison
ZOJ - blandness	SOJ - namesake
KAN - college	TUX - party
GYQ - outcome	XEJ - forethought
SAK - potato	PER - cottage
NYJ - context	VUF - array

APPENDIX C

APPENDIX C

Memory Ability Tests

Background Frequency Discrimination Test*

				Proportion Correct
stride	(16)	couple	(122)	.98
victim	(27)	final	(156)	.68
skirmish	(4)	modulate	(1)	.66
torpor	(2)	drivel	(1)	.60
convert	(12)	ignite	(2)	.64
switch	(43)	list	(133)	.82
address	(77)	early	(366)	.58
night	(411)	price	(108)	.54
anchor	(15)	dissent	(5)	.42
swift	(32)	music	(216)	.84
veterinary	(4)	dill	(3)	.48
ocular	(1)	straggle	(3)	.76
other	(1702)	again	(578)	.56
room	(383)	until	(461)	.50
sunshine	(8)	transfer	(38)	.48
can	(1772)	time	(1599)	.70
many	(1030)	make	(794)	.70
cameo	(1)	juror	(4)	.76
janitor	(4)	idol	(7)	.46
day	(686)	little	(831)	.42

down	(895)	end	(410)	.58
result	(244)	thud	(3)	.98
scale	(60)	each	(877)	.96
suit	(48)	superb	(14)	.78
base	(91)	heritage	(21)	.78
volcano	(2)	humor	(47)	.92
spread	(83)	charter	(33)	.90
after	(1070)	half	(275)	.96
case	(362)	nature	(191)	.54
insect	(14)	easel	(5)	.98
plateau	(3)	infant	(11)	.90
name	(294)	clear	(219)	.88
law	(299)	world	(787)	.48
frost	(6)	jump	(24)	.88
gator	(2)	kneel	(5)	.94

---

* Numbers in parentheses are frequencies of occurrence reported by Kucera and Francis (1967).

Meaningfulness Discrimination Test*

				Proportion Correct
bird	(7.89)	decree	(5.16)	.94
baby	(7.04)	bacteria	(6.12)	.86
determination	(4.64)	fault	(4.80)	.76
flask	(6.28)	hospital	(7.44)	.96
grass	(7.54)	hope	(5.52)	.76
idea	(4.88)	gentleman	(5.80)	.88
morgue	(6.56)	plain	(5.20)	.72
saloon	(7.12)	shotgun	(7.88)	.20
strawberry	(6.71)	semester	(5.48)	.58
python	(5.88)	pudding	(7.31)	.72
tool	(6.88)	wine	(7.54)	.56
advice	(5.39)	betrayal	(5.00)	.54
style	(5.84)	yacht	(7.20)	.76
wheat	(7.96)	thief	(6.50)	.30
arrow	(6.80)	expression	(6.13)	.60
revolt	(5.60)	spinach	(7.08)	.52
mosquito	(7.84)	medallion	(6.32)	.74
jelly	(6.00)	forrest	(9.12)	.72
deluge	(5.32)	causality	(4.38)	.38
author	(5.24)	clock	(7.08)	.60

* Numbers in parentheses are meaningfulness values reported by Paivio, Yuille, and Madigan (1968).

Ease-of-Learning Rating Task

Item	Actual Probability Correct	$\bar{X}$ Rated Ease-of-Learning
FIG - poster	.48	4.80
QAZ - cuisine	.22	3.72
DIP - energy	.59	4.36
BUK - library	.74	6.96
LET - salad	.59	7.42
NYZ - nymph	.70	4.80
WOK - pacificism	.48	2.32
TYN - microscope	.52	3.18
PIC - gallery	.41	6.70
PAK - lawn	.52	4.20
SIC - doctor	.82	8.28
SAV - barrel	.19	4.12
JAX - distance	.33	3.70
HAF - domicile	.26	2.70
YEG - loquacity	.22	1.56
GIT - musician	.37	5.06
WOR - army	.77	7.58
FAN - dynasty	.48	4.54
NAT - inhabitant	.48	5.66
PEP - candy	.74	7.50
PUN - dome	.33	3.96
PAG - newspaper	.59	6.82
FYQ - flash	.44	2.68
PAW - storeroom	.41	3.26
XEZ - discrete	.04	1.80

Ease-of-Learning Discrimination Task*

				Proportion Correct
first - new	(13.96)	major - various	(12.38)	.98
fine - warm	(13.48)	many - keen	(12.73)	.82
tarsal - hard	(12.11)	rabbinical - pretty	(10.69)	.92
human - recuperative	(12.02	dark - nutty	(13.12)	.82
whilom - ritualistic	(9.09)	tonal - wobbly	(10.87)	.88
nudist - waxy	(12.67)	nosy - vulpine	(11.30)	.82
past - zestful	(12.38)	sneaky - diluvial	(11.05)	.71
happy - late	(13.51)	styptic - yellow	(12.20)	.92
jellied - white	(12.61)	vast - less	(13.16)	.65
close - tenpenny	(12.18)	besprent - daily	(10.75)	.78
lorn - top	(11.14)	gray - pivotal	(10.49)	.61
next - loamy	(11.50)	tangy - waste	(12.62)	.86
fit - visceral	(11.62)	long - towery	(12.93)	.90
daq - cov	(5.77)	laj - vux	(4.77)	.90
fem - hos	(8.22)	fal - tex	(9.50)	.41
fev - mir	(6.48)	pav - kof	(7.59)	.51
rus - kip	(6.82)	xej - fon	(4.33)	.94
wi- - sec	(6.99)	sic - jil	(8.23)	.53
kng - nsh	(4.80)	bes - ceh	(5.91)	.80
guk - huq	(5.96)	sav - poh	(7.07)	.76

* Numbers in parentheses are EL values reported by Richardson and Erlebacher (1958).

APPENDIX D

APPENDIX D

Analyses of Variance for Alternative JK Measures

Analysis of Variance Summary Table

for JK d'

Source	SS	df	MS	F
Groups	24.71	2	12.35	9.37***
Error (Between)	122.57	93	1.31	
Lists	9.14	2	4.57	3.89**
Lists by Groups	11.72	4	2.93	2.49*
Error (Within)	218.52	186	1.17	

---

* $p < .05$

** $p < .01$

*** $p < .001$

Analysis of Variance Summary Table

for Probability Correct JK

Source	SS	df	MS	F
Groups	1.65	2	.82	9.66***
Error (Between)	7.97	93	.08	
Lists	.70	2	.35	5.25**
Lists by Groups	.95	4	.23	3.56*
Error (Within)	12.42	186	.06	

---

* $p < .05$

** $p < .005$

** $p < .001$

Analysis of Variance Summary Table

for JK-Errors

Source	SS	df	MS	F
Groups	92.22	2	46.11	3.60*
Error (Between)	1190.76	93	12.80	
Lists	62.22	2	31.11	4.87**
Lists by Groups	148.08	4	37.02	5.80***
Error (Within)	1186.99	186	6.38	

---

* $p < .05$

** $p < .01$

*** $p < .005$

APPENDIX E

APPENDIX E

Correlations Between JK Accuracy Measures

Correlations Between JK Accuracy Measures
for the Varied Ease Group

	d'	P(JK)	JK-Errors	Recall
List 1				
C.A.Q.	.85	.84	-.68	-.43
d'		.92	-.50	-.28
P(JK)			-.62	-.36
JK-Errors				-.09
List 2				
C.A.Q.	.84	.73	-.64	.09
d'		.87	-.70	.18
P(JK)			-.60	.05
JK-Errors				-.04
List 3				
C.A.Q.	.84	.89	-.56	-.18
d'		.96	-.68	-.19
P(JK)			-.49	-.21
JK-Errors				.02

A coefficient of .35 is significantly different from zero, $p < .05$.

Correlations Between JK Accuracy Measures

for the Varied Frequency Group

	d'	P(JK)	JK-Errors	Recall
List 1				
C.A.Q.	.84	.85	-.58	-.03
d'		.93	-.60	-.02
P(JK)			-.57	.09
JK-Errors				.01
List 2				
C.A.Q.	.85	.83	-.49	.22
d'		.96	-.64	.38
P(JK)			-.60	.31
JK-Errors				-.10
List 3				
C.A.Q.	.64	.70	-.59	-.29
d'		.91	-.64	.19
P(JK)			-.63	.12
JK-Errors				-.05

A coefficient of .35 is significantly different from zero, $p < .05$.

Correlations Between JK Accuracy Measures

for the Control Group

	d'	P(JK)	JK-Errors	Recall
List 1				
C.A.Q.	.81	.82	-.64	.07
d'		.94	-.62	.03
P(JK)			-.70	.13
JK-Errors				-.04
List 2				
C.A.Q.	.78	.81	-.60	-.02
d'		.96	-.68	.21
P(JK)			-.74	.25
JK-Errors				.06
List 3				
C.A.Q.	.77	.73	-.59	-.31
d'		.96	-.60	-.19
P(JK)			-.62	-.18
JK-Errors				-.02

A coefficient of .35 is significantly different from zero, $p < .05$.

APPROVAL SHEET

The dissertation submitted by Joseph F. King has been read and approved by the following committee:

Dr. Eugene B. Zechmeister, Director
Associate Professor, Psychology, Loyola

Dr. Deborah L. Holmes
Assistant Professor, Psychology, Loyola

Dr. Emil J. Posavac
Associate Professor, Psychology, Loyola

The final copies have been examined by the director of the dissertation and the signature which appears below verifies the fact that any necessary changes have been incorporated and that the dissertation is now given final approval by the Committee with reference to content and form.

The dissertation is therefore accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

4-2-77
Date

Eugene B Zechmeister
Director's Signature