Journal of Experimental Psychology: General
1994, Vol. 123, No. 3, 297-315

0096-3445/94/$3.00

Memory in Naturalistic and Laboratory Contexts: Distinguishing the Accuracy-Oriented and Quantity-Oriented Approaches to Memory Assessment

Asher Koriat and Morris Goldsmith

A distinction is drawn between the quantity-oriented approach to memory that has dominated traditional laboratory research, and the accuracy-oriented approach that is emerging in the study of everyday memory. This distinction is shown to underlie some troubling confusions in the interpretation of empirical findings. In particular, the *recall–recognition paradox,* which involves the claimed superiority of recall over recognition memory in naturalistic settings, is shown to stem from the common confounding between memory property (quantity vs. accuracy) and 2 other variables that have not generally been distinguished—test format (production vs. selection) and report option (free vs. forced reporting). Three laboratory experiments reveal the fundamentally different roles played by report option and test format in accuracy-based and quantity-based memory research. Implications for memory assessment, metamemory, and the everyday–laboratory controversy are discussed.

The impetus for this article derived from an examination of recent claims that memory in real-life situations differs in significant ways from the kind of memory that has traditionally been investigated in the laboratory (see, e.g., Cohen, 1989; Conway, 1991; Gruneberg, Morris, & Sykes, 1991; Neisser, 1988b). Empirical evidence cited in this context reveals some intriguing puzzles that seem to call for a closer scrutiny of the pretheoretic assumptions underlying memory assessment. For example, there is evidence to suggest that memory for real-life events is considerably better than would be expected from laboratory research (e.g., Wagenaar, 1986, indicating about 96% retention after 2 years!). Other evidence, however, has led authors such as Barclay (1986) to argue that "memories for most everyday life events are ... transformed, distorted, or forgotten" (p. 89). Also, in seeming defiance of the standard laboratory observation that multiple-choice recognition memory performance is superior to recall, the established wisdom in eyewitness research holds that free-narrative recall testing is actually preferable to recognition, because of the contamination introduced by directed questioning (see, e.g., Hilgard & Loftus, 1979; Neisser, 1988b).

A critical analysis of the evidence and claims surrounding the everyday–laboratory controversy (see, e.g., the January 1991 issue of *American Psychologist.*) led us to distinguish two fundamentally different treatments of memory implicit in current memory research. The first treatment, which seems to have dominated traditional laboratory research, views memory as a storage place and thus evaluates memory primarily in terms of the amount of information retained or lost. The second treatment, implicit in much of the more recent work on everyday memory, views memory as a representation of past events and thus evaluates memory in terms of its faithfulness or correspondence to these events.

We begin by outlining the basic characteristics of these alternative conceptions of memory in terms of two memory metaphors, the *storehouse* and the *correspondence* metaphors (for a detailed analysis of the metatheoretical basis of this distinction, see Koriat & Goldsmith, 1994). Each is shown to imply a distinct approach to memory assessment—a quantity-oriented and an accuracy-oriented approach, respectively. We then propose a three-variable classification of assessment methods, in terms of memory property (quantity vs. accuracy), test format (production vs. selection), and report option (free vs. forced), which provides a useful framework for analyzing some of the troubling inconsistencies that emerge when one compares naturalistic and laboratory findings. Finally, we report several experiments designed to clarify these inconsistencies by exposing the interactions among the three variables, further demonstrating the general utility of the proposed framework.

Storehouse Versus Correspondence Conceptions of Memory

The pervasive influence of conceptual metaphors in the study of memory is well documented (e.g., Kolers &

Asher Koriat and Morris Goldsmith, Department of Psychology, University of Haifa, Haifa, Israel.

This research was supported by Grant 032–1106 from the Israel Foundations Trustees. It was conducted while Morris Goldsmith was supported by a National Science Foundation Graduate Fellowship and completed while Asher Koriat was the Irv Acenberg visiting scientist at the Rotman Research Institute of Baycrest Centre, Toronto, Ontario, Canada. The work was carried out at the Institute of Information Processing and Decision Making, University of Haifa. We are grateful to Ian Begg, Ron Fisher, Maryanne Garry, Larry Jacoby, Colleen Kelly, and Ulric Neisser for their helpful comments on earlier drafts of the article. We thank Michal Sion for her help in setting up the experiments and in the preparation of the manuscript of this article.

Correspondence concerning this article should be addressed to Asher Koriat, Department of Psychology, University of Haifa, Haifa, Israel.

Roediger, 1984; Malcolm, 1977; Marshall & Fryer, 1978; Roediger, 1980; see also Gentner & Grudin, 1985). In this article we focus on the contrast between the *storehouse* and the *correspondence* metaphors of memory and their specific implications for memory assessment. We illustrate the distinction between these conceptions by comparing two representative memory situations—list learning and eyewitness testimony.

List Learning and the Storehouse Metaphor

Consider first the typical laboratory paradigm of list learning: A subject is presented with a list of words that he or she is asked to memorize. After a certain retention interval, the subject is asked to recall as many words from the list as possible. Memory performance is measured by the percentage of words recalled out of the total number of words presented.

The list learning paradigm embodies a particular way of thinking about memory that is intrinsic to the *storehouse metaphor.* In this conception, memory is seen as an information-storage place (see Roediger, 1980), and indeed, the list learning paradigm essentially simulates the course of events presumed to take place when memory items are initially *deposited* and subsequently *retrieved.* It is also assumed that the contents of memory consist of discrete, elementary units. Thus, as in many memory experiments, the list learning paradigm makes use of discrete stimuli, typically referred to as *items,* whose essential characteristic is their countability, allowing measures of memory effectiveness based on the number of recovered elements. Moreover, memory is assessed in an *input-bound* manner: One begins with the input and asks how much of it was recovered in the output, that is, how much was retained and how much was lost. In scoring free-recall performance, for instance, incorrect responses (i.e., commission errors) are often simply ignored. Forgetting, then, is basically conceived as *information loss,* indicated by the proportion of input items that cannot be recovered. This treatment assumes that the items are interchangeable, that is, equivalent as far as the total memory score is concerned. Thus, it makes no difference whether *HAT* was remembered and *GUN* was forgotten, or vice versa. In general, the content of the recollected items is immaterial. What matters is not *what* is remembered but rather, *how much.*

These attributes of the storehouse conception characterize a *quantity-oriented* approach to memory, in which memory is conceived primarily in terms of its amount (see Schacter, 1989). This approach is reflected in the traditional experimental paradigms used to study memory (e.g., list learning, paired associates; see, e.g., Puff, 1982), the type of phenomena investigated (e.g., the effects of list length, retention interval, spacing, serial order), and the memory measures used (e.g., percentage of items recalled or recognized). Until recently, the dominance of the storehouse metaphor in guiding memory research was virtually unrivaled (Roediger, 1980). Thus, even though perhaps no investigator today would completely endorse a strict storehouse conception, it is important nonetheless to face its underlying logic, which still pervades much of the way that memory is treated in contemporary memory research.

Event Memory and the Correspondence Metaphor

Let us next consider, in contrast, a second memory situation, one that is perhaps more representative of everyday memory; for example, an eyewitness is asked to report whatever he or she can regarding the circumstances of a crime. This type of situation is also represented in the experimental paradigms that attempt to simulate eyewitness testimony, for example, a subject is presented with a staged event and later is asked to recount the event or is questioned about specific details (see, e.g., Loftus, 1979a). Such situations, as well as many other real-life memory phenomena, motivate a different way of thinking about memory, one in which the basic criterion is not the quantity of items remaining in store but rather the *correspondence* between what the person reports and what actually happened (see Winograd, in press). Although there appears to be no single concrete metaphor (like the storehouse) that alone can provide the essential features for such an alternative conception, this view can nevertheless be conveyed in terms of a more abstract *correspondence metaphor,* consisting of the following interrelated attributes:

First, memory is considered to be *about* some past event, to constitute a representation or description of a past episode (see, e.g., Conway, 1991). Thus, memory reports are seen as consisting of propositions that have truth value, that is, that can be judged as right or wrong or as being more or less true to aspects of an actual event.

Second, in the eyewitness situation, as in many real-life situations, interest lies primarily in the extent to which the memory report is *reliable, trustworthy, accurate* (see, e.g., Deffenbacher, 1988, 1991; Hilgard & Loftus, 1979; Loftus, 1979a). Thus, memory is evaluated in terms of its *fit* with previous events—the extent to which it accords with reality—and forgetting is conceived as a *loss of correspondence* between the memory report and the actual event, as a deviation from veridicality rather than as a mere loss of items. This leads to a focus on the many different types of qualitative memory distortions—fabrication, confabulation, simplification, and the like (see, e.g., Alba & Hasher, 1983; Bahrick, Hall, & Dunlosky, 1993; Bartlett, 1932; Brewer & Nakamura, 1984; Dawes, 1966; Goldmeier, 1982; Loftus, 1979a, 1979b, 1982; Neisser, 1981, 1988c; Wells & Loftus, 1984; Riley, 1962).

Third, the content of the memory report is important: Unlike in the quantity-oriented approach, where interest focuses almost exclusively on how much is remembered, in the correspondence-oriented approach (and virtually all real-life memory situations), *what* is remembered matters a great deal (Conway, 1991). In the courtroom, for instance, whether the witness remembered that the assailant had a gun but forgot that he wore a hat, rather than vice versa, might make a crucial difference.

Fourth, the assessment of memory correspondence is inherently *output bound.* Unlike the storehouse approach, which begins with the input and asks how much of it is recovered in the output, in a correspondence view of memory it is more natural to begin with the output (e.g., the eyewitness report) and examine to what extent it accords with the input. In general, accuracy can be measured only for what a person reports (e.g., the height of the assailant, the color of his hair), not for what is omitted. Thus, whereas in the storehouse view subjects are held accountable primarily for what they *fail* to report, under the correspondence view subjects are accountable primarily for what they *do* report. This treatment has much in common with the way we think about perception. In perception, the question of interest is not how much of the information impinging on our senses we perceive, but rather, how well what we perceive corresponds to what is out there. Thus, in a sense, the correspondence view treats memory as the perception of past events and asks to what extent is this perception dependable (cf. "memory psychophysics," Algom, 1992).

Taken together, these aspects of the correspondence metaphor characterize an *accuracy-oriented* approach to memory. This way of treating real-life memory may be clearly seen in both psycholegal and autobiographical memory research, in which the study of memory accuracy and distortion is given a high priority (see, e.g., Barclay, 1988; Barclay & Wellman, 1986; Deffenbacher, 1988; 1991; Loftus, 1979a, 1979b, 1982; Neisser, 1988b; Neisser & Winograd, 1988; Wells & Loftus, 1984; Winograd, in press; Winograd & Neisser, 1992). Of course, the focus on accuracy does not preclude consideration of the completeness of the memory report as well (see Hilgard & Loftus, 1979).

In sum, when the two situations illustrated above—list learning and eyewitness testimony—are contrasted, they seem to imply two fundamentally different ways of thinking about memory. Each conception entails a different criterion for evaluating memory performance: In a word, whereas the storehouse conception treats memory as something that can be *counted,* the correspondence conception treats memory as something that can be *counted on.*

Distinguishing Quantity-Based and Accuracy-Based Methods of Memory Assessment

Although many discussions of everyday memory phenomena imply a departure from the storehouse conception toward a correspondence view, it is not generally realized that the latter view implies very different methods of assessing memory than those commonly used in the traditional assessment of memory quantity. We shall now examine some of the more specific implications of the two conceptions for the assessment of memory performance.

The correspondence metaphor perhaps finds its most unique expression in the *wholistic* assessment of memory performance, that is, in efforts to derive an overall measure of faithfulness for the memory output considered as a whole. Such efforts have been made in certain circumscribed domains, primarily in the area of memory for spatial information (see, e.g., Allen, Siegel, & Rosinski, 1978; Hart, 1979, 1981; Pick & Lockman, 1981; Siegel, 1981; Siegel & Schadler, 1977; Waterman & Gordon, 1984). These measures are generally based on pattern-matching techniques to compute the goodness of fit between a particular target stimulus and its reconstruction from memory. (For a more detailed discussion of this approach, see Koriat & Goldsmith, 1994.) Also, when interest is focused on memory for particular attributes, memory correspondence can be evaluated along continuous dimensions, such as height, angle, speed, or time (e.g., Algom, Wolf, & Bergman, 1985; Baddeley, Lewis, & Ninno-Smith, 1978; Bahrick et al., 1993; Bartlett, 1932; Byrne, 1979; Goldmeier, 1982; Huttenlocher, Hedges, & Bradburn, 1990; Huttenlocher, Hedges, & Duncan, 1991; Huttenlocher, Hedges, & Prohaska, 1988; Linton, 1975; Loftus & Marburger, 1983; McNamara, 1986; Nelson & Chaiklin, 1980; Riley, 1962; Tversky, 1981; Tversky & Schiano, 1989; White, 1982).

Wholistic memory assessment, however, presents many difficulties (see Koriat & Goldsmith, 1994; Neisser & Harsch, 1992), which is perhaps one reason why most researchers in need of an overall memory measure, regardless of the underlying metaphor, have used an *item-based* approach. In this approach, the memory report is segmented into discrete items or propositions that can be dichotomously evaluated as either right or wrong and generally are given equal weight in computing the overall memory score. Such an approach, which follows from the storehouse metaphor, is less well suited to the focus on memory correspondence, and indeed, Neisser (1988c) has chided everyday memory researchers for treating memory "as if it were just a set of remembered concrete experiences" (p. 356). For practical reasons, however, it often does seem fruitful to slice memory (or at least memory reports) into individual items of information that may be considered independently. When this is done, quantity-based and accuracy-based memory assessment may still be distinguished in terms of the contrast between *input-bound* and *output-bound* evaluation. Despite its limitations, then, the item-based approach has the advantage of allowing both types of assessment to be compared within a common framework. This is the approach taken in our research reported below.

Input-Bound Versus Output-Bound Measures

To illustrate the distinction between input-bound and output-bound measures, let us again consider an eyewitness situation in which both the witnessed event (input) and the memory report (output) can be segmented into sets of propositions (items). We may then distinguish a quantity-based and an accuracy-based measure of memory as follows: The quantity measure is equivalent to the likelihood of correctly remembering an input item, whereas the accuracy measure is the likelihood that a reported item is correct (i.e., corresponds to the input). Thus, to compute accuracy, we simply begin with the list of statements made by the eyewitness and calculate what percentage of these are true. Because this

measure is *output bound,* it directly reflects the extent to which each reported item may be depended on to be correct. In contrast, the quantity measure, being *input bound,* requires an itemized description of the entire event to determine the percentage of items remaining in store (such a description, of course, could never capture *all* of the input information; see McCauley, 1988).

Despite their different definitions, in practice there are cases in which the quantity and accuracy measures cannot be distinguished operationally. Thus, when memory is tested through a *forced-choice* procedure, the likelihood of remembering each input item (quantity) is necessarily equivalent to the likelihood that each reported item is correct (accuracy). For example, assuming that a person's memory for 20 items of information is tested using a forced-choice recognition test, and 12 out of the 20 items are answered correctly, then the quantity and accuracy scores will both be 0.60.

When, then, will item-based quantity and accuracy measures differ operationally? This will usually occur under *free-report* conditions, that is, when subjects are free to volunteer or withhold information. Consider, for example, the following situation: An eyewitness is asked to remember the people he saw entering a bar and reports that he saw A, B, and C. If A, B, and C did indeed enter the bar, then this testimony is entirely accurate. The fact that other people, D and E, also entered the bar but were not reported by the witness will not detract from the (output-bound) accuracy of the information that was provided. In contrast, construed as a free-recall task intended to tap the (input-bound) amount of information that can be reproduced, reporting only three people out of five will obviously count against the reporter.

More generally, input-bound and output-bound measures are necessarily equivalent when the output list (e.g., people reported as entering the bar) is the same length as the input list (people actually entering the bar). Such is invariably the case in forced-choice testing methods. In free-report conditions, however, the option to reply is controlled by the subject, that is, he or she is allowed to say "I don't know" (cf. Neisser, 1988b). Thus, the operational distinction between output-bound accuracy (i.e., "dependability") and input-bound quantity is applicable only to free-report testing methods.

The role of report option in differentiating accuracy-based and quantity-based memory measures illustrates how a concern with memory correspondence may bring to the fore issues that are less intrinsic to a storehouse framework—in this case, the active role of the rememberer in controlling his or her memory output. As we shall see, a more careful consideration of this variable may help to resolve some current puzzles in the memory literature.

Disentangling Item-Based Assessment Methods: A Proposed Three-Variable Classification

We introduced this article by noting some confusion surrounding claims of differences between memory performance in real-life and traditional laboratory contexts. We now examine some of the seemingly incongruous findings emerging from the everyday–laboratory controversy and show how they can be clarified in terms of the distinction between the quantity-oriented and accuracy-oriented approaches to memory assessment. In addition to *memory property* (quantity vs. accuracy), we focus on two other assessment variables that previous work has failed to distinguish clearly—*test format* (production vs. selection) and *report option* (free vs. forced). We address each variable in turn and attempt to clarify some important empirical issues in which all three are implicated.

Memory Property

As noted earlier, there is considerable ambivalence concerning the quality of everyday memory performance in comparison with memory in a laboratory context. Although a comprehensive review is beyond the scope of this article, we agree with Barclay and Wellman (1986), who, after considering the inconsistent evidence, concluded that "the general accuracy of autobiographical memories is thus unclear" (p. 94; see also Neisser, 1988b). We propose, however, that this lack of clarity may be due in part to a failure to distinguish between the two memory properties, quantity and accuracy.

The most impressive memory for autobiographical events, for example, has been found using recognition tests that do not include foils (e.g., Brewer, 1988; Linton, 1975, 1978; Wagenaar, 1986; see Wallace, 1980, for a discussion advocating this procedure). Strictly speaking, such tests allow no possibility for error; the subject merely rates the strength or amount of memory remaining for each stimulus item. In contrast, when foil alternatives are included, subjects have been found to exhibit a large proportion of "false memories," identifying foil items as true events. For example, in Barclay and Wellman's (1986) study (see also Barclay, 1988), correct acceptance of original records was very high even after 1 year (94%); however, false-alarm responses were also relatively frequent, increasing to a level of 59% after 10–12 months. Barclay and Wellman concluded that "recognition memory for autobiographical events is both strikingly accurate and inaccurate, when viewed from two different perspectives" (1986, p. 99). This conclusion might be profitably recast in terms of memory property: Their subjects appear to have exhibited good quantity but poor accuracy performance. Indeed, a similar pattern was reported as early as 1904 by Stern (1904, reproduced in Neisser, 1982). Noting that "testimony can be evaluated in terms of two principal criteria: The amount of recall and its accuracy," Stern observed that "the average amount, in terms of the absolute number of correct responses is quite substantial. . . .Accuracy, in contrast, is poor indeed" (p. 101 in Neisser, 1982).

Test Format

The second variable to be considered is *test format,* which refers to the nature of the procedure used to test memory.

Testing procedures may be distinguished along a continuum that represents the extent to which the possible response alternatives are constrained, ranging from *production* tests, in which subjects produce answers with little or no intervention (e.g., free-narrative memory reports[1] and free recall), to *selection* tests, in which the experimenter or interrogator provides one or more memory stimuli as response alternatives (e.g., lineup identification and multiple-choice recognition). Between the two extremes there are procedures, such as cued recall or directed questioning, that exert intermediate levels of constraint (e.g., "Did the assailant have a beard?" or "What was the color of the car?"). Although there seems to be general agreement in both everyday and laboratory research that test format is an important variable affecting memory performance, there is a lack of clarity regarding not only the mechanisms involved but even the direction of the effects.

In the study of eyewitness testimony, for instance, there is a widely held belief that testing procedures involving recognition or directed questioning can have contaminating effects on memory (see, e.g., Hilgard & Loftus, 1979). This idea has received a great deal of support from the work of Loftus and her associates on the effects of misleading postevent information, demonstrating that memory for an episode can sometimes be distorted by information contained in subsequent questioning (see Hall, Loftus, & Tousignant, 1984; Loftus, 1975, 1979a, 1979b; Loftus & Hoffman, 1989; Loftus, Miller, & Burns, 1978; Wagenaar & Boer, 1987). Other work has also supported this possibility. For instance, the viewing of mug shots has been shown to impair subjects' ability to recognize faces that they had seen earlier (e.g., Brown, Deffenbacher, & Sturgill, 1977; Gorenstein & Ellsworth, 1980). Also, Lipton (1977) compared memory for a filmed murder under four testing procedures and found that memory accuracy decreased from unstructured free-narrative questions through open-ended questions to short-answer, leading questions and—worst of all—multiple-choice questions.

Thus, it is well established in eyewitness research that "the form in which a question is put to a witness exerts a strong influence on the quality of the answer" (Hilgard & Loftus, 1979, p. 348). In fact, the general recommendation is to elicit information initially in a free-narrative format before moving on to directed questioning, and even then, to put more trust in the former (see Fisher, Geiselman, & Raymond, 1987; Flanagan, 1981; Hilgard & Loftus, 1979; Timm, 1983).

This body of evidence, however, stands in sharp contrast to the well-established superiority of recognition over recall memory in traditional list-learning laboratory experiments (e.g., Shepard, 1967; see also Brown, 1976). In fact, this discrepancy could be taken to suggest the operation of variables specific to the context of inquiry—real life versus laboratory. For instance, Neisser (1988b), in reporting findings from a naturalistic study, stressed that "practitioners of the traditional psychology of memory might be surprised" by the finding that "recognition produces errors where recall does not" (p. 553). However, because laboratory list-learning experiments have focused exclusively on memory quantity, this discrepancy may simply be due to differences in the memory property studied. In that case, the findings would actually suggest the possibility of an *interaction between memory property and test format:* Whereas recognition testing may be superior to recall in terms of memory quantity performance, recall testing may yield better accuracy performance (see Hilgard & Loftus, 1979; Lipton, 1977; Neisser, 1988b). Because such an interactive pattern might still seem to be at odds with traditional laboratory wisdom, we call it the *recall–recognition paradox.*

The clearest support for this pattern comes from Lipton's (1977) above-mentioned study, in which the results for memory quantity were precisely the opposite of those reported earlier for accuracy: Quantity performance was worst for unstructured free-narrative and best for multiple-choice questioning. Neisser (1988b) also found such divergent effects on quantity and accuracy. In Neisser's study, recall testing yielded far better accuracy than did multiple-choice recognition, but the reverse was true for quantity performance. As Neisser put it, the recall subjects generally produced "no errors but also not much recall" (p. 553). The recall–recognition paradox is also implied in Hilgard and Loftus's (1979) discussion; they concluded from the available evidence that free-narrative reports "are consistently more accurate but less complete than reports obtained through specifically directed inquiry" (p. 342).

Report Option

The foregoing discussion implies that the recall–recognition paradox may be construed as an interaction between test format and memory property. There is a serious problem with this interpretation, however, because questioning methods that differ in test format often vary in *report option* (free vs. forced) as well. Consider, for example, the following quote from Neisser (1988b) regarding some determinants of memory accuracy:

> The important distinction seems to be between what might be called "open retrieval"—unconstrained free or cued recall—and "forced retrieval," which includes the situation of most witnesses as well as all multiple-choice tests. When people are simply asked what they remember without being compelled or constrained or motivated to make an impressive reply, then what they say is generally not wrong. Why should it be? They can always say "I don't remember" if they don't remember. . . . But it's quite another story when an interrogator presses a witness to give a specific answer, or when a subject must choose one of several response alternatives whether he wants to or not . . . (pp. 548–549).

The confounding between report option and test format implicit in Neisser's discussion reflects the common practice in both naturalistic and traditional memory research. In

[1] Strictly speaking, free-narrative questioning procedures are not item based. Nevertheless, the memory reports are most often analyzed in terms of discrete items of information that are scored as correct or incorrect (e.g., Fisher, Geiselman, & Amador, 1989), in which case performance may be compared with that obtained using other, item-based procedures.

questioning witnesses, for example, free-narrative reporting not only guards against the potential presentation of contaminating or leading information in the question (a test-format variable), it also allows the witness the freedom to choose which items of information to report (report option). Directed questioning or recognition types of interrogation, however, often involve either explicit or implicit demands that an answer be provided (Neisser, 1988b). Similarly, in traditional item-based laboratory research, recall testing typically allows subjects the freedom to report only what they feel they actually remember, whereas recognition testing is almost invariably implemented as *forced* recognition in two distinct respects: Not only are subjects confined to the alternatives presented by the experimenter (selection format), they also are forced to choose an answer for each and every item (forced report). These two types of constraint are, however, logically independent.

Despite the confounding between test format and report option, the contrast between recall and recognition testing has traditionally been considered in terms of test format alone. Thus, there has been little concern for the fact that subjects in a recall test typically have the option to volunteer or withhold responses and that recall performance may therefore reflect both memory and metamemory processes operating at the time of retrieval (but see, e.g., Bousfield & Rosner, 1970; Erdelyi, Finks, & Feigin-Pfau, 1989; Klatzky & Erdelyi, 1985; Koriat & Goldsmith, 1993; Nelson & Narens, 1990, in press; Roediger & Payne, 1985). Clearly, however, subject-controlled *metamemory* processes exert an effect in most everyday free-reporting situations, and such processes may also be operative, perhaps to a lesser extent, in more sterile, laboratory-based test situations as well (see e.g., Bousfield & Rosner, 1970; Erdelyi et al., 1989; Gruneberg, Monks, & Sykes, 1977; Koriat, 1993; Koriat & Goldsmith, 1993; Nelson & Narens, 1990). We therefore use the term *free production* (or free recall) to denote the feature characteristic of most recall tasks in both laboratory and everyday memory research, namely, that the subject is free to report an answer or to withhold it.[2]

Certainly, the confounding between test format and report option further complicates the interpretation of the recall–recognition paradox. For example, in Neisser's (1988b) study, in which recall was more accurate than recognition, the recall procedure differed from the recognition procedure in both test format and report option. Thus, the recall subjects may have used the option of free report to refrain from giving answers that were likely to be wrong. This would tend to boost their memory accuracy but reduce memory quantity, suggesting that the recall–recognition paradox might in fact reflect an *interaction between report option and memory property.*

Indeed, Neisser (1988b) pointed out that his recall subjects seemed to achieve greater accuracy by providing fewer answers. In addition, however, the subjects might also have used a further aspect of report option to boost their accuracy—control over the "grain size," or generality, of their responses (cf. Yaniv & Foster, 1990, 1993). Clearly, the correspondence between memory reports and past events can improve when the answers are more general and less detailed. Thus, Neisser observed that his recall subjects tended to choose "a level of generality at which they were not mistaken" (p. 553). Also, Barclay (1986, 1988) pointed out that memory reports may be truthful in reconstructing the gist of an event yet quite inaccurate in reproducing the details (see also Neisser, 1981, 1988c; Spence, 1982).

Proposed Conceptual Framework

To sum the discussion so far, there appears to be some confusion in the literature regarding different methods of assessing memory performance, particularly in connection with memory accuracy. Overall, our analysis points to three variables that must be considered conjointly—memory property, test format, and report option. These may exhibit interactive patterns whose interpretation is complicated by the fact that in previous studies the contributing factors were generally confounded, both among themselves and with different research settings. As a framework for unraveling their individual and combined effects, Figure 1 presents a proposed classification of item-based assessment methods in terms of the three variables. Note that although both test format and report option might perhaps be represented as continuous dimensions, for simplicity they are treated here as binary attributes. Also, because of the operational equivalence between forced-report quantity and accuracy measures discussed earlier, this classification distinguishes six (rather than eight) different types of assessment methods.

Using this classification as a guide, we shall now report some experimental work in which the three implicated factors are empirically disentangled.

Experimental Evidence: Dissociating Accuracy and Quantity Measures

The experimental work reported below is designed to help resolve the recall–recognition paradox and, more generally, to demonstrate the utility of distinguishing between the quantity-oriented and accuracy-oriented approaches to memory assessment. Our general strategy was as follows: First, using item-based assessment procedures, we orthogonally manipulated test format and report option and derived both quantity and accuracy measures. Second, we avoided the grain-size problem inherent in many naturalistic free-report testing procedures by taking control over grain size away from the subject, as is commonly done in list-learning experiments. This allowed recall and recognition accuracy and quantity measures to be meaningfully compared. Third, we conducted all of the experiments in a traditional laboratory context, which permitted us to explain the obtained pattern of effects without recourse to presumed

[2] In fact, the term *free recall* might have been expected to carry this connotation generally. However, in traditional usage this term has been reserved for the contrast with serial recall, denoting merely that the subject is free to choose the order in which items are to be recalled.

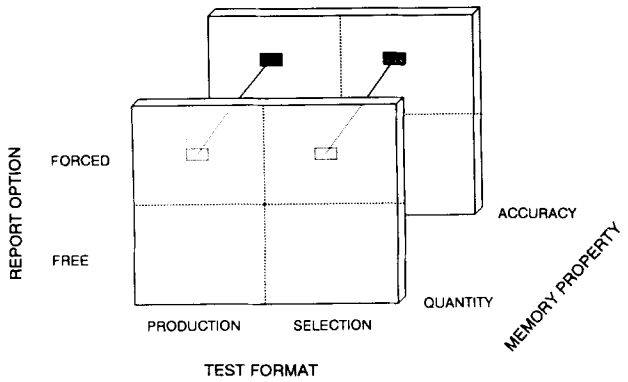

Figure 1. A proposed three-way classification of item-based memory assessment methods. The connecting lines indicate that the forced-report quantity and accuracy methods are operationally equivalent.

differences between the dynamics of memory in naturalistic and laboratory contexts.

Experiment 1

In Experiment 1 we examined the effects of test format and report option on accuracy-based and quantity-based measures of memory performance. A general-knowledge test was administered in either a production (recall) or a selection (multiple-choice) test format, and performance was scored for both quantity (input bound) and accuracy (output bound). In addition, report option was orthogonally manipulated: In the forced-report conditions, subjects were required to answer all items, whereas in the free-report conditions subjects had the option of volunteering or withholding answers. Thus, the design made use of two relatively uncommon testing procedures—forced recall and free recognition. The forced-recall procedure required subjects to respond to each question, even if they felt they were just guessing. Conversely, the free-recognition procedure allowed subjects the option to refrain from choosing any of the alternatives for a given item.

The experiment also included a payoff schedule that was designed to provide subjects in all conditions with a common performance incentive. Essentially, subjects were rewarded for each correct answer but were penalized by an equal amount for each incorrect answer.

When comparing the standard tests of memory—free recall and forced recognition—we expect to replicate the recall–recognition paradox: Recognition memory should yield superior quantity performance, whereas recall should yield superior accuracy. By orthogonally manipulating test format and report option, however, the design will allow the source of the paradox to be clarified.

Method

Subjects. Eighty-nine Hebrew-speaking undergraduate psychology students at the University of Haifa (31 men and 58 women) participated in the experiment for course credit. Subjects were randomly assigned to four experimental conditions, with 22–23 subjects in each group.

Stimulus materials. A 60-item general-knowledge test that covered a broad range of topics was developed in Hebrew. Two versions of the test were prepared, a production (recall) version and a 5-alternative multiple-choice selection (recognition) version. The questions for the two tests were identical, but in the recall version a blank line was provided next to each question for recording the response, whereas in the recognition version five possible answers were listed, one of which was correct (the foils were designed to be as plausible as possible). The questions were formulated such that the correct answer was always a single word or a proper name (cf. Brown & McNeill, 1966; Nelson & Narens, 1990). This was designed to minimize the problem of subject control over the grain size of responses, discussed earlier. Examples of the questions used include: What was the name of the composer who wrote the "Moonlight Sonata" (selection alternatives: Beethoven, Bach, Tchaikovsky, Schumann, Brahms)? What is the chemical process responsible for the formation of glucose in the plant cell (selection alternatives: electrolysis, glucolysis, photosynthesis, dialysis, unitization)? All instructions were provided together with the test in a self-contained booklet.

Procedure. The experiment was administered in group sessions that lasted about 45 min each. After some preliminary instructions, the performance payoff schedule, the same in all conditions, was explained: Subjects were paid New Israeli Shekel (NIS) 1 (about $0.50) for each correct answer but were penalized the same amount for each incorrect answer, for a randomly selected sample of 15 items. Subjects were assured, however, that they would be exempt from paying losses. Thus, the possible bonus could range from NIS 0 to NIS 15. Subjects were nevertheless urged to think of each and every answer as potentially yielding an NIS 1 gain or loss.

The instructions then differed for the two report-option conditions. Subjects in the forced-report conditions were required to answer all the questions, even if they had to guess, but were encouraged to provide or mark their best answer in any case. Subjects in the free-report conditions, in contrast, were told that they could choose whether or not to answer any given question and that they would not be penalized (but neither would they receive any bonus) for omitted items.

After completing the test, subjects in the free-report conditions participated in a second phase, in which they were required to answer the questions that they had initially omitted. (Three subjects did not participate in this phase—1 recall subject who was unable to conform to the instructions and 2 recognition subjects who answered all items in Phase 1). There was no additional payoff associated with Phase 2 performance.

Results

Two memory indices were calculated for each subject; an input-bound quantity score, defined as the percentage of correct answers out of the total number of questions (i.e., 60); and an output-bound accuracy score, defined as the percentage of correct answers out of the total number of questions actually answered by the subject. Because in the forced-report conditions the number of output answers necessarily equals the number of input questions, the quantity and accuracy scores for these conditions are equivalent, as noted earlier (see Figure 1).

The means of the quantity and accuracy scores for each condition are presented in Table 1. Because most previous discussions of recall and recognition performance imply a comparison between free recall and forced recognition, we first compare the results for these two conditions. The "paradoxical" pattern we have referred to previously is immediately apparent: On the one hand, better memory quantity performance was observed for forced recognition than for free recall, $t(42) = 4.35$, $p < .0001$. On the other hand, better memory accuracy performance was found for free recall than for forced recognition, $t(42) = 2.56$, $p < .05$.

The source of this paradox can be clarified, however, by considering the joint effects of test format and report option. First, in a two-way analysis of variance (ANOVA) for the quantity scores, test format was the critical variable: Recognition yielded significantly more correct answers than did recall, $F(1, 85) = 30.07$, $p < .0001$, but report option had no effect and there was no interaction ($F < 1$, for both).

Second, a similar ANOVA on the accuracy scores yielded $F(1, 85) = 13.72$, $p < .0005$, for test format; $F(1, 85) = 53.93$, $p < .0001$, for report option; and $F(1, 85) = 13.03$, $p < .0005$, for the interaction. Here, report option was the critical variable: Allowing subjects the option of free report clearly improved their accuracy performance, and this effect was more pronounced for the recall than for the recognition test. Despite the overall effect of test format (evidenced when the free- and forced-report conditions are collapsed), there was no difference between recall and recognition accuracy in the free-report condition. Note that in the forced-report condition, the accuracy and quantity measures are equivalent by definition and in effect indicate better recognition quantity performance. For the free-report condition, however, the two measures are independent, and here the means of the memory accuracy scores are virtually identical for the recall and recognition tests ($t = 0.07$ for the difference).

Some insight into the mechanism underlying these findings can be gained by considering the Phase 2 results. In Phase 1, free-report subjects refrained from answering an average of 23.1 and 11.8 questions on the recall and recognition tests, respectively. When required to answer these items in Phase 2, correct answers averaged 14.4% for recall and 31.1% for recognition. These percentages are considerably lower than the respective accuracy scores for Phase 1, $t(20) = 25.5$, $p < .0001$, for recall and $t(20) = 9.33$, $p < .0001$, for recognition. This suggests that the increased accuracy of the free-report condition stemmed from the screening of answers that were less likely to be correct than those that were volunteered. The screening process, however, was apparently not perfect, because some correct answers also were withheld.

Table 1
Means and Standard Errors of Quantity and Accuracy Memory Scores (Percentage Correct) in Experiment 1 for Each Test Format × Report Option Condition

	Quantity		Accuracy	
Test format	Free	Forced	Free	Forced
Recall				
M	47.8	47.6	76.6	47.6
SD	3.2	2.8	2.3	2.8
Recognition				
M	61.5	67.0	76.9	67.0
SD	3.0	3.0	2.4	3.0

Discussion

What are the implications of these findings? First and foremost, the results demonstrate the importance of distinguishing between memory quantity and memory accuracy. These two memory properties were found to be dissociable: Test format affected quantity performance but not accuracy, whereas report option affected accuracy but not quantity. Thus, when comparing memory performance across different testing situations, it is critical to consider the memory property being evaluated as well as the potentially divergent contributions of test format and report option.

Second, the experiment replicated the recall–recognition paradox and in particular the superior accuracy of free recall over forced recognition found in everyday memory research. Importantly, this pattern was obtained within a sterile laboratory setting, indicating that it is not confined to real-life situations and to the various social and functional variables that may be specific to them. Furthermore, unlike previous, naturalistic studies, in which the recall superiority may have resulted from a greater control over the grain size of the responses (e.g., Neisser, 1988b), this superiority was observed even when grain size was equated across recall and recognition formats. Thus, although such control may perhaps be useful in improving the accuracy of memory reports (particularly in naturalistic settings), clearly this is not the only variable operating.

Third, when subjects were allowed equal opportunities to screen their responses, *free* recall was no more accurate than was *free* recognition. This finding may have important practical and theoretical implications, because it runs counter to the common belief that memory recognition tests are inherently damaging to memory accuracy. Note that it was free recognition rather than free recall that produced the best overall performance, because it was as accurate as recall but at the same time yielded better quantity performance.

Finally, the results help illustrate the utility of distinguishing between the traditional quantity-oriented approach to memory assessment, guided by the storehouse metaphor, and the accuracy-oriented approach emerging in everyday memory research. In particular, the failure to acknowledge some of the unique emphases of the accuracy-oriented approach (e.g., subject control, output-bound measures) may be responsible for some confusion in the current memory literature.

Experiment 2

We designed Experiment 2 to generalize the results of Experiment 1 to the typical laboratory paradigm of list learning. This procedure is commonly assumed to tap episodic memory (Tulving, 1972, 1983) rather than the semantic memory tested in Experiment 1. If the results of Experiment 1 are found to hold with respect to the list-learning procedure of Experiment 2, this should further substantiate the conclusion that the superior recall accuracy noted by everyday memory researchers does not stem from special characteristics of naturalistic contexts but instead reflects metamemory processes that have probably been operating (whether relevant or not) in much of the memory research conducted to date.

Memory for a list of words was tested under three conditions, similar to the respective conditions of Experiment 1—free recall, free recognition, and forced recognition. Forced recall was not included because it is technically difficult to implement in a list-learning paradigm, and its inferior accuracy should be sufficiently obvious (see, e.g., Roediger & Payne, 1985). We expected to obtain the same pattern of results as in Experiment 1, except that because of the lack of a factorial design, the analyses would rely on specific comparisons.

Method

Subjects. Sixty Hebrew-speaking subjects (24 men and 36 women) received NIS 4 (about $2) for their participation. They were randomly and equally divided among the three experimental conditions.

Stimulus materials. A list of 35 common Hebrew words was tape recorded at a rate of 4 s/word. The multiple-choice recognition test consisted of 35 rows of 5 words each—the correct word and 4 distractors. All distractors were of the same grammatical class as the target and were semantically related to it. Example: stop, continue, complete, finish, start. The order of the rows was random.

Procedure. The experiment was conducted either individually or in groups of up to 3 subjects. Subjects were first played the recorded list, which they were told to try and remember for future testing (the type of test was not specified). There were two filler tasks, each lasting 2 min, between the study and test phases. For the test phase, subjects were told that they would win 1 point for each correct word but would lose 1 point for each incorrect word and were urged to try to maximize their scores by answering accurately (monetary incentives were not used here). The instructions differed according to the experimental condition: Free-recall subjects listed as many words as possible on a blank sheet. In the recognition conditions, forced-recognition subjects were required to respond to all test items by circling one word in each row, whereas free-recognition subjects were given the option to refrain from answering a particular item in order to maximize their scores. As in Experiment 1, subjects in the free-recognition condition also participated in a second phase in which they answered the items that they had initially omitted (except for 1 subject, who answered all items in Phase 1).

Results

We derived input-bound quantity scores and output-bound accuracy scores as in Experiment 1. The results are presented in Table 2.

Consider first the comparison between the free-recall and forced-recognition conditions. On the one hand, forced recognition produced better quantity performance than did free recall, $t(38) = 7.75$, $p < .0001$. On the other hand, free recall produced better accuracy than did forced recognition, $t(38) = 3.56$, $p < .001$. Thus, the recall–recognition paradox was also exhibited using the standard tests of episodic memory.

The results for the third condition, however, indicate that the critical variable affecting accuracy is report option, not test format: Although free recognition was substantially more accurate than was forced recognition, $t(38) = 4.16$, $p < .0005$, free recall was no more accurate than was free recognition, $t(38) = 0.42$, *ns*. Thus, the superior accuracy observed when comparing the standard recall and recognition tests apparently stems from the option given recall subjects to volunteer a response or not. As in Experiment 1, allowing recognition subjects this same option enhanced their accuracy but did not reduce their quantity scores compared with forced recognition, $t(38) = 0.72$.

The free-recognition subjects withheld an average of 9.3 items in Phase 1. Examination of the Phase 2 results reveals that the accuracy of the answers to these withheld items (31.3%) was substantially lower than that of the volunteered answers in Phase 1 (85.1%), $t(18) = 14.58$, $p < .0001$. Thus, as in Experiment 1, the increased accuracy of free recognition apparently stemmed from an effective (though imperfect) screening process.

Discussion

Experiment 2 replicated the pattern of results obtained in Experiment 1, but this time in the context of a more traditional, list-learning paradigm. Thus, the same relationships among test format, report option, and memory property have been established in two laboratory situations that involved quite different memory tasks—a general-knowledge

Table 2
Means and Standard Errors of Quantity and Accuracy Memory Scores (Percentage Correct) for Each Test Format × Report Option Condition in Experiment 2

	Quantity		Accuracy	
Test format	Free	Forced	Free	Forced
Recall				
M	27.7	—	83.2	—
SD	3.3	—	3.4	—
Recognition				
M	62.0	65.6	85.1	65.6
SD	3.4	3.6	3.0	3.6

Note. Dashes signify that data were not collected for that measure.

test of long-term semantic memory and a list-learning test of episodic memory. This is, of course, in addition to the original naturalistic context in which the recall–recognition paradox was originally observed (e.g., Lipton, 1977; Neisser, 1988b). The generality of these relationships across tasks and contexts testifies to the broad applicability of the distinction between quantity-based and accuracy-based memory assessment and also illustrates how at least some of the dynamics that emerge in real-life memory situations may be operationalized and studied fruitfully in the laboratory.

Experiment 3

The results of Experiments 1 and 2 indicate that memory quantity performance and memory accuracy performance are dissociable. Whereas quantity performance varies primarily with test format, memory accuracy varies with report option. In addition, the findings suggest that, unlike memory quantity performance, memory accuracy is under subject control: When given the option of free report, subjects were able to boost their accuracy scores, apparently by screening out incorrect answers.

In the present experiment, we examine whether the option of free report can be used to improve memory accuracy even further when a high premium is placed on accuracy. We investigated this question by using the same general-knowledge test as in Experiment 1 but now with a stronger incentive for accuracy: Whereas in Experiment 1 the bonus for a correct answer was offset by an equal penalty for an incorrect answer, in Experiment 3 subjects forfeited all winnings if they volunteered even a single incorrect answer. Here we included only the free-report conditions, both recall and recognition. This allowed us to compare memory performance in these high-incentive conditions with the respective moderate-incentive conditions of Experiment 1.

If subjects are able to regulate their accuracy performance in accordance with accuracy motivation, this will add one more dimension in terms of which accuracy-based results may be expected to differ from traditional, quantity-based findings. With regard to the latter, the evidence indicates that subjects cannot be induced through monetary incentives to improve their memory quantity performance beyond what they attain under normal conditions (e.g., Nilsson, 1987; Weiner, 1966a, 1966b). However, we expect that when subjects are strongly motivated for accuracy, they will be able to achieve high levels of accuracy by using a more stringent screening policy.

Moreover, if the screening process is effective but not perfect, as was hinted by the Phase 2 results of Experiments 1 and 2, then the higher accuracy performance could come at the expense of lower quantity performance, that is, a *quantity–accuracy trade-off* (see Klatzky & Erdelyi, 1985; Koriat & Goldsmith, 1993; Lipton, 1977; Neisser, 1988b). Although such a trade-off was not evidenced in Experiments 1 and 2, perhaps it will emerge when there is a stronger motivation for accuracy. Thus, a comparison of both memory accuracy and memory quantity performance across Experiments 1 and 3 may shed further light on the dynamics underlying free-report memory performance.

The results of Experiment 3 also will allow us to reexamine the hypothesis that recall is inherently more accurate than recognition, but this time under conditions in which there is a strong motivation for accuracy. The type of payoff asymmetry used in this experiment may better characterize the situation of a person on the witness stand, for instance, where accuracy is given a high premium. Thus, although not supported in the previous experiments, perhaps the claim for inferior recognition accuracy will find support in a situation more characteristic of legal testimony.

Method

Design and procedure. In Experiment 3 we used two free-report conditions, differing only in test format (recall vs. recognition). The stimuli and procedure were the same as those of the respective free-report conditions in Experiment 1, except for the incentive manipulation: Subjects were instructed to provide as many correct answers as they could, but *only* correct answers. Specifically, they would win NIS 1 (about $0.50) for each correct answer volunteered, but would forfeit all winnings if they volunteered even a single incorrect answer. Thus, the two conditions of Experiment 3 (high incentive) together with the respective free-report conditions of Experiment 1 (moderate incentive) conform to a 2 × 2, Incentive × Test Format factorial. As in Experiment 1, a second phase was included in which subjects provided answers to all items initially omitted.

Subjects. Fifty-six Hebrew-speaking psychology students at the University of Haifa (18 men and 38 women) participated in the experiment for course credit. They were randomly and equally divided between the two experimental conditions.

Results

In view of the strict accuracy incentive, a question of immediate interest is whether there were any "winners" at all. Surprisingly, fully 25% of the subjects (14) succeeded in achieving 100% accuracy. These were equally distributed between the recall and recognition conditions. As one might expect, recognition winners provided more correct answers on average than did recall winners (28.9 vs. 19.6); $t(12) = 2.15$, $p < .05$, one-tailed. Overall, the 14 winners provided from 12 to 38 (correct) answers, for an average bonus of NIS 24.2. Non-winners provided from 13 to 48 answers, with an average of 28.3. Of these answers, however, an average of 3.2 were incorrect.

The effects of accuracy incentive can be seen in Table 3, in which the results of Experiments 1 and 3 are compared. Consider the accuracy scores first. As predicted, a two-way ANOVA (Test Format × Incentive) revealed substantially higher accuracy for the high-incentive conditions than for the moderate-incentive conditions, $F(1, 97) = 50.57$, $p < .0001$. Neither test format nor the interaction yielded significant effects.

Of further interest is whether recall accuracy was superior to recognition accuracy under the strong incentive of Experiment 3. One can see that recall accuracy was, if any-

Table 3
Means and Standard Errors of Quantity and Accuracy Memory Scores (Percentage Correct) for Free-Report Performance With Moderate and High Accuracy Incentives in Experiments 1 and 3

	Quantity		Accuracy	
Test format	Moderate[a]	High[b]	Moderate[a]	High[b]
Recall				
M	47.8	36.3	76.6	90.1
SD	3.2	2.4	2.3	2.1
Recognition				
M	61.5	46.4	76.9	92.7
SD	3.0	2.3	2.4	1.4

[a] Experiment 1. [b] Experiment 3.

thing, slightly inferior to recognition accuracy, though the difference was not significant, $t(54) = 1.06$. Also, the absence of a significant interaction in the preceding ANOVA indicates that motivation to be accurate did not exert a differential effect on recall and recognition accuracy.

Turning next to the quantity scores, a similar two-way ANOVA revealed a significant effect for test format, $F(1, 97) = 19.06$, $p < .0001$, with recognition evidencing higher quantity scores than recall. More important, the effect of accuracy incentive was also highly significant, $F(1, 97) = 23.89$, $p < .0001$. Both recognition and recall subjects in Experiment 3 yielded lower quantity scores than their counterparts in Experiment 1 (there was no interaction). Thus, subjects in Experiment 3 achieved higher accuracy, but with a sacrifice in quantity performance.

Once again, examination of the Phase 2 data may shed some light on the underlying mechanisms. When the results from Phases 1 and 2 are considered together, the effects of increased accuracy incentive appear to stem from a more stringent screening of the answers provided in Phase 1, not from differences in the amount of correct information accessible to the subjects. An estimation of the amount of accessible information in terms of the total number of correct answers in Phases 1 and 2 combined yields an overall quantity score that averaged 61.9 for the free-report conditions of Experiment 1 and 61.0 for Experiment 3. Subjects in the two experiments did differ, however, in their screening behavior: More answers were withheld in Experiment 3 (36.0 for recall, 29.8 for recognition) than in Experiment 1 (23.1 for recall, 11.8 for recognition), $F(1, 97) = 80.3$, $p < .0001$.

Of course, to achieve better accuracy performance, subjects in Experiment 3 had to withhold specifically more incorrect answers (27.3 for recall, 15.2 for recognition), than did subjects in Experiment 1 (19.1 for recall, 7.5 for recognition), $F(1, 96) = 32.37$, $p < .0001$. However, they also withheld more correct answers (8.6 for recall, 14.6 for recognition) than did subjects in Experiment 1 (3.1 for recall, 4.3 for recognition), $F(1, 96) = 95.34$, $p < .0001$, thereby reducing their quantity performance.

Discussion

The results of Experiment 3 indicate that, when given the option of free report, subjects can substantially boost their memory accuracy in response to increased accuracy motivation. This finding contrasts with the general failure to enhance memory quantity performance through monetary rewards (e.g., Nilsson, 1987). Thus, unlike memory quantity, memory accuracy appears to be under the strategic control of the subject. Importantly, the increased accuracy motivation did not produce an improvement in the overall correctness of the subjects' answers but rather appeared to encourage the use of a more stringent screening process. Because this process was not perfect, withholding more answers yielded a quantity–accuracy trade-off.

In addition, like the previous experiments, Experiment 3 also failed to support the claimed superior accuracy of recall over recognition. Thus, we find no evidence for such a superiority, even when a high priority is given to accurate reporting.

General Discussion

The results of the present experiments suggest that some of the confusion surrounding disparate claims in naturalistic and laboratory memory research may stem from critical differences in the way that memory is being evaluated. Of foremost importance is the distinction between quantity-based and accuracy-based memory assessment. On the one hand, traditional memory research, guided by the storehouse metaphor, has evaluated memory primarily in terms of the quantity of items that can be recovered (Schacter, 1989) under tightly controlled laboratory conditions (Banaji & Crowder, 1989). The more recent wave of naturalistic research, on the other hand, has brought with it a greater concern for the accuracy or faithfulness of memory reporting, coupled with an increased willingness to allow subjects control over their memory reporting (see Hilgard & Loftus, 1979; Neisser, 1988b). Our experimental findings indicate that one important aspect of such control, report option, may play different roles in accuracy-based and quantity-based memory assessment. Yet the effects of report option have often been either ignored or confused with those of test format. Thus, an important contribution of the present study lies in demonstrating the utility of the proposed classification of assessment methods in which these variables are orthogonally represented. We therefore begin our discussion by considering in more detail the mapping between the proposed classification and currently used assessment methods, particularly with respect to the option of free report.

Refining the Three-Variable Classification of Assessment Methods

The proposed classification of item-based assessment methods distinguishes three basic assessment variables—memory property, report option, and test format—yielding six operationally distinct types of assessment procedures

(see Figure 1). We will now attempt to further clarify the characteristics of each type and then compare our framework with that of the signal detection approach to memory.

Quantity-Based Methods

Consider first the four input-bound quantity-based methods. Here the two most widely used testing procedures are free production and forced selection. *Free production* includes any test in which the subjects produce (retrieve, generate, etc.) their own answers and also have the option to refrain from responding. Such tests encompass the recall of items from lists, the answering of open-ended questions, and also paired-associate or cued-recall tasks in which subjects are allowed to skip over items. *Forced selection* denotes any test in which the subject must make a response based on one or more presented options and is not allowed to skip any item. Included, for example, are multiple-choice tests of episodic or semantic memory, old–new and yes–no recognition tests, and so forth.

The two remaining quantity-based methods are much less prevalent, yet their use in conjunction with the former methods allows the unconfounding of test format and report option. *Forced production* is similar to free production except for the added requirement to answer all items. This procedure can be readily applied in connection with cued or paired-associate tests (e.g., Hart, 1967; Loftus & Wickens, 1970; Murdock, 1966), as well as with open-ended and fill-in-the-blank questionnaires (which are commonly used to assess implicit memory; see Richardson-Klavehn & Bjork, 1988; Schacter, 1987). Some researchers have applied this method to uncued list recall as well, by forcing subjects to provide the same number of items as were in the original study list, or some other large, preset quota (e.g., Cofer, 1967; Erdelyi & Becker, 1974; Erdelyi et al. 1989, Roediger & Payne, 1985).

Conversely, *free selection* differs from forced selection by removing the demand to answer all items, that is, by allowing a response of "don't know," which is scored as neither right nor wrong. These tests should be distinguished from tests that merely include response alternatives such as "none of the above" or "word not in list." Because such alternatives may still be evaluated as being correct or incorrect, tests incorporating them should properly be classified as *forced*-selection tests. In eyewitness-lineup studies, for example, the option to give a "target-missing" response does not in itself make the testing procedure free, unless subjects also are allowed to simply say "I don't know." Similarly, in "old–new" recognition tests, although subjects may mark any number of items as "old" or "new," they are required to make a decision regarding each and every item. Each of these decisions is necessarily right or wrong. This, then, distinguishes the old–new paradigm, which is commonly used in signal-detection studies of memory (see Murdock, 1982), from free-selection methods, in which subjects are allowed to abstain (see further discussion below). In fact, as an alternative to forced, signal-detection methods, free-selection quantity measures are often used in achievement and aptitude tests that discourage guessing, to derive an unbiased estimate of memory quantity (see Budescu & Bar-Hillel, 1993; Cronbach, 1984).

Accuracy-Based Methods

Let us now turn to the output-bound, accuracy-based methods. As we discussed earlier, quantity-based and accuracy-based memory measures differ operationally only under *free report* conditions, in which subjects can decide which answers to volunteer and which to withhold. Here, the output-bound accuracy measure uniquely reflects the dependability of the memory report, that is, the conditional probability that each volunteered item of information is correct. By contrast, under forced-report conditions the same measure can be taken to express either quantity or accuracy, because the input-bound proportion of questions correctly answered is equivalent to the output-bound proportion of provided or selected answers that are correct. Thus, the difference between the forced-report accuracy-based and quantity-based methods is essentially conceptual—whether subjects' responses are conceived in terms of correspondence with some to-be-remembered event or simply in terms of the amount of information recovered. In some cases an accuracy orientation is explicitly stated by the experimenter, as is typically the case in eyewitness research. In other cases, however, it can only be inferred from cues that disclose the experimenter's implicit treatment of the subject's responses (e.g., a qualitative analysis of memory errors).

We should emphasize that although memory accuracy may be evaluated using either free- or forced-reporting methods, the type of accuracy being assessed in each case is quite different. Under free-report conditions, subjects generally provide only information that they actually believe to be correct, so that their performance is mediated by a decision process used to screen out incorrect answers. Therefore, memory accuracy performance on free-report tests will depend critically on the effectiveness of *metamemory* processes (see Koriat & Goldsmith, 1993). Such is not the case with forced-report methods.

Comparison With Signal-Detection Methodology

This brings us to an important point that we are now ready to address. Clearly, there is an overall resemblance between our proposed framework and the signal-detection methodology for memory measurement (see, e.g., Banks, 1970; Bernbach, 1967; Kintsch, 1967; Klatzky & Erdelyi, 1985; Lockhart & Murdock, 1970; Murdock, 1966; Norman & Wickelgren, 1969). This methodology raises many of the same issues brought out in this article, yet there are crucial differences. It is important, then, to consider how signal-detection methods fit into the above classification.

The signal-detection methodology allows a separation between two parameters of memory performance, d' and β: The first parameter, d', is considered to measure "true" memory, whereas β reflects "response bias" (see, e.g.,

Klatzky & Erdelyi, 1985). The logic behind this methodology is best illustrated in the old–new recognition paradigm, in which it is typically applied. Here, subjects can arbitrarily increase their hit rate by lowering their criterion for responding "old." Thus, the false-alarm rate is used to adjust the hit rate for response bias. Both the hit rate and the false-alarm rate, then, are considered in tandem to compute a single measure of true memory, d' (or another similar measure).

Because signal-detection methods have been applied in conjunction with forced-selection procedures, they do not permit the derivation of independent memory quantity and accuracy measures, as do free-report assessment methods.[3] In fact, like other forced-report measures, whether d' is taken to express quantity or accuracy is simply a matter of the experimenter's interpretation. As a quantity measure, d' corrects the hit rate for false alarms and in this sense resembles other measures that correct memory quantity performance for guessing. It also, however, reflects the overall accuracy of the "old–new" responses. Thus, d' may be alternatively interpreted as reflecting either memory strength or memory sensitivity.

In contrast, free-report assessment methods permit an entirely different response decision—whether to volunteer or withhold a candidate answer. Indeed, it is precisely because omissions are allowed that under free-report conditions the signal-detection methodology cannot be applied (see Lockhart & Murdock, 1970). Moreover, although the logic of signal detection can be extended to free-report tasks (see Klatzky & Erdelyi, 1985; Koriat & Goldsmith, 1993), the motivation for doing so has generally been to control for criterion effects (differences in accuracy) when comparing quantity measures (e.g., Erdelyi & Becker, 1974), rather than to measure memory *accuracy* as a property of interest in its own right. Thus, although the signal-detection approach has contributed greatly to a consideration of the role of decision processes in forced recognition memory, it actually has little to say regarding the accuracy of a person's freely reported remembrances.

The proposed classification, then, may be seen as offering a general framework that supplements the signal-detection approach by incorporating both free and forced memory reporting (see also Koriat & Goldsmith, 1993). In the present article, we found this framework useful for analyzing some troubling issues in the literature. In the following section we examine the implications of our experimental work with regard to these issues. In the final section, we consider how the present framework might possibly be extended.

Implications of the Experimental Findings

Taken together, the three experiments we report in this article should underscore the distinction between the quantity-oriented and accuracy-oriented approaches to memory assessment. These experiments effectively dissociated the two memory properties, quantity and accuracy, demonstrating that quantity-based and accuracy-based measures are affected both by different variables and by the same variables in divergent ways: On the one hand, test format affected memory quantity performance but not memory accuracy. On the other hand, the option of free report increased memory accuracy but in itself had no effect on memory quantity performance. Furthermore, under free-report conditions, a strong accuracy incentive increased accuracy but decreased quantity (i.e., a quantity–accuracy trade-off). We now consider how these findings can help clarify some of the ambiguities evident in the memory literature. Our discussion will focus on three interrelated topics—the recall–recognition paradox, subject control, and the "context of inquiry" issue stemming from the everyday–laboratory controversy.

Recall–Recognition Paradox

In view of the differential effects of report option and test format on quantity and accuracy measures, the tendency in memory research to confound these two variables is regrettable. As a consequence, the recall–recognition paradox, which compares free recall with forced recognition, could be interpreted as reflecting a Property × Test Format interaction, with recognition testing eliciting more complete but less accurate memory reports than recall. Our results, however, indicate that the paradox actually stems from a three way, Property × Test Format × Report Option, interaction: The superior accuracy of free recall over forced recognition is due to report option, whereas the superior quantity of forced recognition over free recall is due to test format.

The common practice of using selection testing procedures in conjunction with forced reporting undoubtedly has contributed to the idea that recognition memory is inherently less accurate than recall. However, in the present study, test format had no effect on memory accuracy when report option was held constant. This finding questions the belief that recognition testing is necessarily contaminating and suggests a possible refinement of the established wisdom in eyewitness research that directed questioning is less reliable than is free-narrative reporting (e.g., Hilgard & Loftus, 1979). Although this generalization may be true, it could derive primarily—perhaps entirely—from report option rather than from test format per se. In fact, the present results suggest that in item-based testing, a free-selection procedure may actually be more effective than free-produc-

[3] As mentioned earlier, subjects in the old–new paradigm can control only the distribution of old and new responses—they are not free to abstain. Thus, although the hit- and false-alarm rates are sometimes considered to represent quantity and accuracy measures, respectively, note that unlike free-report measures, these proportions have no meaning independent of one another. In contrast, the free-report quantity measure cannot be arbitrarily raised to any desired level, and the free-report accuracy measure has its own meaningful interpretation in terms of the dependability of volunteered information. Of course, in free-report assessment it may still be important to consider quantity and accuracy performance together (see Klatzky & Erdelyi, 1985; Koriat & Goldsmith, 1993).

tion, given that substantially better quantity performance was achieved with no sacrifice in accuracy. Note, however, that in some real-life situations, such as eyewitness testimony, incorrect selection alternatives may be both highly plausible and forcefully presented.[4] Because such parameters might be expected to affect the quality of memory monitoring (see Koriat & Goldsmith, 1993; Weingardt, Leonesio, & Loftus, in press), the generalizability of the present results should be investigated further. We also note that directed questioning and selection procedures may in any case still have contaminating effects on subsequent testing occasions (e.g., Boon & Davies, 1988; Loftus, 1975).

Subject Control Over Memory Reporting

In helping resolve the recall–recognition paradox, our results clearly demonstrate the criticality of report option for the accuracy-based assessment of memory: Across the three experiments, the accuracy advantage of free report over forced report ranged from 61% to 89% for recall and from 15% to 38% for recognition. Moreover, the results indicate that memory accuracy performance is under strategic control: Given the option of free report, subjects were able to improve their accuracy performance in accordance with increased accuracy incentive (Experiment 3). The increased accuracy, however, was achieved at the expense of quantity performance (about a 25% decrease for both recall and recognition).

These results contrast sharply with findings indicating that subjects *cannot* improve their memory *quantity* performance when given incentives to do so (e.g., Nilsson, 1987; Weiner, 1966a, 1966b; but see Loftus & Wickens, 1970). For example, Nilsson (1987) gave recall and recognition subjects monetary incentives for producing as many correct answers as possible, with the incentives announced either before study or before test. In no case did incentive subjects provide more correct answers than control subjects who were not given any special incentive. Also, studies that have investigated the effects of recall criterion (e.g., Bousfield & Rosner, 1970; Britton, Meyer, Hodge, & Glynn, 1980; Cofer, 1967; Erdelyi, 1970; Erdelyi et al., 1989; Keppel & Mallory, 1969; Roediger & Payne, 1985; Roediger, Srinivas, & Waddil, 1989) have found that encouraging or forcing subjects to produce more items generally does not improve their memory quantity performance much or at all beyond that obtained under standard free-recall instructions.

The differential effects of report option and performance incentives on quantity and accuracy measures suggests that subject control over memory reporting may need to be treated quite differently in quantity-oriented and accuracy-oriented research. Indeed, although the general failure to take subject control into account in quantity-oriented assessment methods may perhaps be justified on empirical grounds (see Roediger et al., 1989), subject control must be taken into account in accuracy-oriented research, because here the effects on memory accuracy can be quite substantial. Furthermore, although quantity motivation is generally ineffective in enhancing quantity performance, accuracy motivation may actually be detrimental to such performance, given the potential quantity–accuracy tradeoff. Thus, for both theoretical and practical reasons, it will be important to investigate further the functions relating accuracy and quantity performance to report option and accuracy incentive and to clarify the mechanisms underlying the strategic regulation of memory performance (see Koriat & Goldsmith, 1993).

In this regard, our results disclosed two trends that point to the contribution of the metamemory processes of *monitoring and control:* First, memory performance computed on the basis of volunteered and withheld answers combined (Phases 1 and 2) indicated that the option of free report did not enhance the overall correctness of the subjects' answers. Second, the answers that were volunteered by free-report subjects in Phase 1 of the experiments had a much higher likelihood of being correct than did those that were withheld (across the three experiments, the accuracy of the volunteered items was about 4–5 times higher for recall and about 2–2.5 times higher for recognition). Nevertheless, some correct answers were also withheld—a significantly greater number when the motivation for accuracy was increased. Taken together, these results suggest that the improved memory accuracy was achieved by using an effective but imperfect screening policy, eliminating answers that were likely to be incorrect.

Thus, beyond the fundamental distinction between accuracy-based and quantity-based memory measures, the experimental results demonstrate the criticality of subject control over memory reporting, particularly for accuracy-oriented memory assessment, and they implicate an important mediating role for metamemory processes (cf. Klatzky & Erdelyi, 1985; Metcalfe, 1993; Nelson & Narens, 1990). We have reported elsewhere (Koriat & Goldsmith, 1993) research designed to clarify the operation of these processes and the manner in which they mediate the effects of report option and accuracy motivation on both memory accuracy and memory quantity performance.

Context of Inquiry and the Everyday–Laboratory Controversy

Let us now consider how the present results bear on the context of inquiry issue, naturalistic versus laboratory. Although the context of inquiry was not manipulated in the present study, this variable is important to consider, because it generally is confounded with both memory property and subject control over memory reporting. We argue that this confounding can often complicate the evaluation of claimed differences between memory performance in naturalistic contexts versus controlled laboratory conditions (cf. Conway, 1991; Neisser, 1988b).

As far as memory property is concerned, the focus of everyday memory researchers on accuracy rather than quan-

[4] We thank Ulric Neisser for pointing this out to us.

tity could sometimes give the impression that memory performance is considerably better in naturalistic settings than in laboratories, unless one notices that the same property is not being assessed in each context. As our results clearly demonstrate, free-report accuracy performance can vastly exceed quantity performance: Across the three experiments, this superiority ranged from 60% to 200% for recall and from 25% to 100% for recognition.

Now consider report option: Real-life situations generally offer much greater control over reporting than is allowed in traditional laboratory research. Moreover, in such situations, functional incentives may induce a stronger motivation for accuracy than do typical laboratory conditions. (In traditional free-recall tasks, for instance, the standard instruction is to list all the words that one can remember, with no explicit instruction to avoid commission errors.) Failure to take into account such systematic differences between the two research contexts could again lead to unwarranted conclusions.

To illustrate, it would be tempting to attribute the remarkable recall accuracy that has been demonstrated in naturalistic settings to the unique functional and motivational variables that affect memory retrieval under real-life conditions (cf. Neisser, 1988b). However, our results indicate that the option of free report, combined with a strong accuracy incentive, will yield impressive recall accuracy (and recognition accuracy) in the laboratory as well. Similarly, context-specific factors appear to be superfluous in resolving the seeming incongruence between naturalistic and laboratory findings embodied in the recall–recognition paradox. Indeed, we demonstrated the superiority of free-report recall accuracy performance over forced-report recognition quantity performance under typical laboratory conditions, for such banal a task as memorizing a list of unrelated words.

Of course, these remarks should not be taken to imply that there are no actual differences in the dynamics of memory between real-life and laboratory settings. On the contrary, our results not only lead us to expect marked differences between the two contexts, they also help identify some of the variables contributing to such differences. For instance, as just noted, subject control and accuracy motivation may vary greatly across different memory contexts, thereby accounting for some of the variance in observed performance levels. The point is that only by isolating the variables that differ between memory contexts can the effects of memory context be demystified and investigated.

As a further example, consider control over the grain size of memory reports. In traditional laboratory research (as in the present study) the grain of responses is typically controlled by the experimenter (or by the nature of the stimuli themselves), whereas in real-life settings the level of detail is more often controlled by the person who is making the report. Consequently, the free-narrative mode of reporting commonly used in naturalistic research, for instance, allows the rememberer to choose that level of generality at which accuracy is likely to be high (Neisser, 1988b). Because such methods are rarely used in traditional laboratory research, subjects in that context are generally deprived of a powerful means of enhancing the accuracy of their reports. This difference need not be treated as simply a methodological problem: Although subject control over grain size may underlie differences in memory performance between the two contexts, once identified, such control can be operationalized and investigated as well (see Yaniv & Foster, 1993).

In sum, the present article essentially delivers a double message regarding the everyday–laboratory controversy: First, at the empirical level, many of the seeming discrepancies between everyday and laboratory findings can perhaps be clarified by considering the different assessment methods characteristic of each context. Second, however, the methodological biases prevalent in each context appear to reflect a more fundamental difference in the underlying conception of memory (see also Koriat & Goldsmith, 1994).

Concluding Remarks

In concluding this article, we return briefly to the metatheoretical foundation of the present work, namely, the contrast between the storehouse and correspondence conceptions of memory. As we argued in the introduction, we believe that the quantity-oriented and accuracy-oriented methodologies actually reflect two different underlying conceptions of memory. Behind the traditional, quantity-oriented approach lies a conception of memory as a storage place where items are deposited and later retrieved. In contrast, the accuracy-oriented approach derives from a treatment of memory as a representation or perception of the past and hence leads to a greater concern with the veridicality and dependability of memory reports. To experimentally compare the two approaches in the present study, we focused on a particular subset of their distinguishing features—input-bound versus output-bound measures, and subject control. The compromises required by our experimental paradigm, however, severely restricted the extent to which the unique aspects of the correspondence metaphor could be expressed. Indeed, the proposed metatheoretical framework clearly extends beyond the item-based domain investigated here. Thus, we wish to point to some of the broader aspects of this framework, which will need to be addressed in future research (for a fuller treatment, see Koriat & Goldsmith, 1994).

Most prominent are those aspects pertaining to the wholistic nature of correspondence and miscorrespondence. These aspects are better revealed when memory concerns complex scenes, events, and stories that have an internal structure. Such materials afford greater opportunities for demonstrating that the changes in memory that occur over time are not confined to the mere loss of individual elements (i.e., forgetting) but also include a variety of qualitative changes, such as distortion, reorganization, confabulation, simplification, and the like (see, e.g., Alba & Hasher, 1983; Bahrick, 1984; Bartlett, 1932; Brewer & Nakamura, 1984; Dawes, 1966; Loftus, 1982). Furthermore, because wholis-

tic correspondence can be achieved at various levels and in many different ways (see McCauley, 1988; Neisser, 1981, 1986, 1988c), subject control and functional variables (e.g., Baddeley, 1988; Bruce, 1989; Neisser, 1981, 1988a, 1988b; Ross, 1989; Winograd, in press) may play an even greater role.

Because the correspondence metaphor embodies a different way of thinking about memory than the storehouse metaphor, correspondence-oriented research should differ from traditional, quantity-oriented research in many significant ways—in the preference for complex stimulus materials having an internal structure, in the focus on the many qualitative ways in which memory can change over time and on the processes underlying these changes, in allowing for the contribution of subject variables and subject control to memory performance, and in the study of motivational and functional factors that affect such contributions.

Indeed, there is a growing body of research that exhibits many of these features, particularly in connection with the study of everyday memory phenomena. The development of memory assessment methods that can adequately deal with this type of complexity, however, remains a major hurdle. Unlike quantity-oriented memory research, which has benefited from a great deal of systematic methodological analysis, there has been relatively little effort invested in appraising the unique requirements of accuracy-oriented research and the logic underlying the assessment of memory accuracy (see Koriat & Goldsmith, 1994). This, then, should provide an important challenge for memory research in the years to come.

References

Alba, J. W., & Hasher, L. (1983). Is memory schematic? *Psychological Bulletin, 93,* 203–231.

Algom, D. (1992). Memory psychophysics: An examination of its perceptual and cognitive prospects. In D. Algom (Ed.), *Psychophysical approaches to cognition* (pp. 441–513). Amsterdam: Elsevier.

Algom, D., Wolf, Y., & Bergman, B. (1985). Integration of stimulus dimensions in perception and memory: Composition rules and psychophysical relations. *Journal of Experimental Psychology: General, 114,* 451–471.

Allen, G. L., Siegel, A. W., & Rosinski, R. R. (1978). The role of perceptual context in structuring spatial knowledge. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 617–630.

American Psychologist. (1991). *46* (1).

Baddeley, A. D. (1988). But what the hell is it for? In M. Gruneberg, P. Morris, & R. Sykes (Eds.), *Practical aspects of memory: Current research and issues* (Vol. 1, pp. 3–18). Chichester: Wiley.

Baddeley, A. D., Lewis, V., & Ninno-Smith, I. (1978). When did you last . . .? In M. Gruneberg, P. Morris, & R. Sykes (Eds.), *Practical aspects of memory* (pp. 77–83). London: Academic Press.

Bahrick, H. P. (1984). Replicative, constructive, and reconstructive aspects of memory: Implications for human and animal research. *Physiological Psychology, 12,* 53–58.

Bahrick, H. P., Hall, L. K., & Dunlosky, J. (1993). Reconstructive processing of memory content for high versus low test scores and grades. *Applied Cognitive Psychology, 7,* 1–10.

Banaji, M. R., & Crowder, R. G. (1989). The bankruptcy of everyday memory. *American Psychologist, 44,* 1185–1193.

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin, 74,* 81–99.

Barclay, C. R. (1986). Schematization in autobiographical memory. In D. C. Rubin (Ed.), *Autobiographical memory* (pp. 82–89). Cambridge, England: Cambridge University Press.

Barclay, C. R. (1988). Truth and accuracy in autobiographical memory. In M. Gruneberg, P. Morris, & R. Sykes (Eds.), *Practical aspects of memory: Current research and issues* (Vol. 2, pp. 289–294). Chichester: Wiley.

Barclay, C. R., & Wellman, H. M. (1986). Accuracies and inaccuracies in autobiographical memories. *Journal of Memory and Language, 25,* 93–103.

Bartlett, F. C. (1932). *Remembering.* Cambridge, England: Cambridge University Press.

Bernbach, H. A. (1967). Decision processes in memory. *Psychological Review, 74,* 462–480.

Boon, J. C. W., & Davies, G. (1988). Attitudinal influences on witness memory: Fact and fiction. In M. Gruneberg, P. Morris, & R. Sykes (Eds.), *Practical aspects of memory: Current research and issues* (Vol. 1, pp. 53–58). Chichester: Wiley.

Bousfield, W. A., & Rosner, S. R. (1970). Free vs. uninhibited recall. *Psychonomic Science, 20,* 75–76.

Brewer, W. F. (1988). Memory for randomly sampled autobiographical events. In U. Neisser & E. Winograd (Eds.), *Remembering reconsidered: Ecological and traditional approaches to the study of memory* (pp. 21–90). Cambridge, England: Cambridge University Press.

Brewer, W. F., & Nakamura, G. U. (1984). The nature and function of schemas. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition.* Hillsdale, NJ: Erlbaum.

Britton, B. K., Meyer, B. J., Hodge, M. H., & Glynn, S. M. (1980). Effects of organization of text on memory: Tests of retrieval and response criterion hypotheses. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 620–629.

Brown, E. L., Deffenbacher, K. A., & Sturgill, W. (1977). Memory for faces and the circumstances of encounter. *Journal of Applied Psychology, 62,* 311–318.

Brown, J. (Ed.). (1976). *Recall and recognition.* London: Wiley.

Brown, R., & McNeill, D. (1966). The "tip of the tongue" phenomenon. *Journal of Verbal Learning & Verbal Behavior, 5,* 325–337.

Bruce, D. (1989). Functional explanations of memory. In L. W. Poon, D. C. Rubin, & B. E. Wilson (Eds.), *Everyday cognition in adulthood and late life* (pp. 44–58). Cambridge, England: Cambridge University Press.

Budescu, D., & Bar-Hillel, M. (1993). To guess or not to guess: A decision-theoretic view of formula scoring. *Journal of Educational Measurement, 38,* 277–291.

Byrne, R. W. (1979). Memory for urban geography. *Quarterly Journal of Experimental Psychology, 31,* 147–154.

Cofer, C. N. (1967). Does conceptual organization influence the amount retained in immediate free recall? In B. Kleinmuntz (Ed.), *Concepts and the structure of memory* (pp. 181–214). New York: Wiley.

Cohen, G. (1989). *Memory in the real world.* Hillsdale, NJ: Erlbaum.

Conway, M. A. (1991). In defense of everyday memory. *American Psychologist, 46,* 19–27.

Cronbach, L. J. (1984). *Essentials of psychological testing.* New York: Harper & Row.

Dawes, R. M. (1966). Memory and distortion of meaningful written material. *British Journal of Psychology, 57,* 77–86.

Deffenbacher, K. A. (1988). Eyewitness research: The next ten years. In M. Gruneberg, P. Morris, & R. Sykes (Eds.), *Practical aspects of memory: Current research and issues* (Vol. 1, pp. 20–26). Chichester: Wiley.

Deffenbacher, K. A. (1991). A maturing of research on the behavior of eyewitnesses. *Applied Cognitive Psychology, 5,* 377–402.

Erdelyi, M. H. (1970). Recovery of unavailable perceptual input. *Cognitive Psychology, 1,* 99–113.

Erdelyi, M. H., & Becker, J. (1974). Hypermnesia for pictures: Incremental memory for pictures but not words in multiple recall trials. *Cognitive Psychology, 6,* 159–171.

Erdelyi, M. H., Finks, J., & Feigin-Pfau, M. B. (1989). The effect of response bias on recall performance, with some observations on processing bias. *Journal of Experimental Psychology: General, 118,* 245–254.

Fisher, R. P., Geiselman, R. E., & Amador, M. (1989). Field test of the cognitive interview: Enhancing the recollection of actual victims and witnesses of crime. *Journal of Applied Psychology, 74,* 722–727.

Fisher, R. P., Geiselman, R. E., & Raymond, D. S. (1987). Critical analysis of police interview techniques. *Journal of Police Science and Administration, 15,* 177–185.

Flanagan, E. J. (1981). Interviewing and interrogation techniques. In J. J. Grau (Ed.), *Criminal and civil investigation handbook* (Pt. 4, pp. 3–24). New York: McGraw-Hill.

Gentner, D., & Grudin, J. (1985). The evolution of mental metaphors in psychology: A 90-year retrospective. *American Psychologist, 40,* 181–192.

Goldmeier, E. (1982). *The memory trace: Its formation and its fate.* Hillsdale, NJ: Erlbaum.

Gorenstein, G. W., & Ellsworth, P. C. (1980). Effect of choosing an incorrect photograph on a later identification by an eyewitness. *Journal of Applied Psychology, 65,* 616–622.

Gruneberg, M. M., Monks, J., & Sykes, R. N. (1977). Some methodological problems with feeling of knowing studies. *Acta Psychologica, 41,* 365–371.

Gruneberg, M. M., Morris, P. E., & Sykes, R. N. (1991). The obituary on everyday memory and its practical applications is premature. *American Psychologist, 46,* 74–76.

Hall, D. F., Loftus, E. F., & Tousignant, J. P. (1984). Postevent information and changes in recollection for a natural event. In G. L. Wells & E. F. Loftus (Eds.), *Eyewitness testimony: Psychological perspectives* (pp. 124–141). Cambridge, England: Cambridge University Press.

Hart, J. T. (1967). Memory and the memory-monitoring process. *Journal of Verbal Learning and Verbal Behavior, 6,* 685–691.

Hart, R. A. (1979). *Children's experience of place.* New York: Irvington.

Hart, R. A. (1981). Children's spatial representation of the landscape: Lessons and questions from a field study. In L. S. Liben, A. H. Patterson, & N. Newcombe (Eds.), *Spatial representation and behavior across the life span* (pp. 195–233). New York: Academic Press.

Hilgard, E. R., & Loftus, E. F. (1979). Effective interrogation of the eyewitness. *International Journal of Clinical and Experimental Hypnosis, 27,* 342–357.

Huttenlocher, J., Hedges, L. V., & Bradburn, N. M. (1990). Reports of elapsed time: Bounding and rounding processes in estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 196–213.

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review, 98,* 352–376.

Huttenlocher, J., Hedges, L. V., & Prohaska, V. (1988). Hierarchical organization in ordered domains: Estimating the dates of events. *Psychological Review, 95,* 471–484.

Keppel, G., & Mallory, W. (1969). Presentation rate and instructions to guess in free recall. *Journal of Experimental Psychology, 79,* 269–275.

Kintsch, W. (1967). Memory and decision aspects of recognition learning. *Psychological Review, 74,* 496–504.

Klatzky, R. L., & Erdelyi, M. H. (1985). The response criterion problem in tests of hypnosis and memory. *International Journal of Clinical and Experimental Hypnosis, 33,* 246–257.

Kolers, P. A., & Roediger, H. L. (1984). Procedures of mind. *Journal of Verbal Learning and Verbal Behavior, 23,* 425–449.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100,* 609–639.

Koriat, A., & Goldsmith, M. (1993). *Monitoring and control processes in the strategic regulation of memory accuracy and memory quantity performance.* Manuscript submitted for publication.

Koriat, A., & Goldsmith, M. (1994). *Memory metaphors and the everyday–laboratory controversy: Comparing the storehouse and correspondence conceptions of memory.* Manuscript submitted for publication.

Linton, M. (1975). Memory for real-world events. In D. A. Norman & D. Rumelhart (Eds.), *Explorations in cognition* (pp. 376–404). San Francisco: Freeman.

Linton, M. (1978). Real world memory after six years: An in vivo study of very long term memory. In M. Gruneberg, P. Morris, & R. Sykes (Eds.), *Practical aspects of memory* (pp. 69–76). London: Academic Press.

Lipton, J. P. (1977). On the psychology of eyewitness testimony. *Journal of Applied Psychology, 62,* 90–95.

Lockhart, R. S., & Murdock, B. B. (1970). Memory and the theory of signal detection. *Psychological Bulletin, 74,* 100–109.

Loftus, E. F. (1975). Leading questions and eyewitness report. *Cognitive Psychology, 7,* 560–572.

Loftus, E. F. (1979a). *Eyewitness testimony.* Cambridge, MA: Harvard University Press.

Loftus, E. F. (1979b). The malleability of memory. *American Scientist, 67,* 312–370.

Loftus, E. F. (1982). Memory and its distortions. In A. G. Kraut (Ed.), *G. Stanley Hall lectures* (pp. 119–154). Washington, DC: American Psychological Association.

Loftus, E. F., & Hoffman, H. G. (1989). Misinformation and memory: The creation of new memories. *Journal of Experimental Psychology: General, 118,* 100–104.

Loftus, E. F., & Marburger, W. (1983). Since the eruption of Mt. St. Helens, has anyone beaten you up? Improving the accuracy of retrospective reports with landmark events. *Memory & Cognition, 11,* 114–120.

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 19–31.

Loftus, G. R., & Wickens, T. D. (1970). Effect of incentive on storage and retrieval processes. *Journal of Experimental Psychology, 85,* 141–147.

Malcolm, N. (1977). *Memory and mind.* Ithaca, NY: Cornell University Press.

Marshall, J. C., & Fryer, D. M. (1978). Speak, memory! An introduction to some historic studies of remembering and forgetting.

In M. Gruneberg & P. Morris (Eds.), *Aspects of memory* (pp. 1–25). London: Methuen.

McCauley, R. N. (1988). Walking in our own footsteps: Autobiographical memory and reconstruction. In U. Neisser & E. Winograd (Eds.), *Remembering reconsidered: Ecological and traditional approaches to the study of memory* (pp. 126–144). New York: Cambridge University Press.

McNamara, T. P. (1986). Mental representations of spatial relations. *Cognitive Psychology, 18,* 87–121.

Metcalfe, J. (1993). Novelty monitoring, metacognition, and control in a composite holographic associative recall model: Implications for Korsakoff amnesia. *Psychological Review, 100,* 3–22.

Murdock, B. B. (1966). The criterion problem in short term memory. *Journal of Experimental Psychology, 72,* 317–324.

Murdock, B. B. (1982). Recognition memory. In C. R. Puff (Ed.), *Handbook of research methods in human memory and cognition* (pp. 1–26). New York: Academic Press.

Neisser, U. (1981). John Dean's memory: A case study. *Cognition, 9,* 1–22.

Neisser, U. (Ed.). (1982). *Memory observed: Remembering in natural contexts.* San Francisco: Freeman.

Neisser, U. (1986). Nested structure in autobiographical memory. In D. C. Rubin (Ed.), *Autobiographical memory* (pp. 71–81). Cambridge, England: Cambridge University Press.

Neisser, U. (1988a). The ecological approach to perception and memory. *New Trends in Experimental and Clinical Psychiatry, 4,* 153–166.

Neisser, U. (1988b). Time present and time past. In M. Gruneberg, P. Morris, & R. Sykes (Eds.), *Practical aspects of memory: Current research and issues* (Vol. 2, pp. 545–560). Chichester: Wiley.

Neisser, U. (1988c). What is ordinary memory the memory of? In U. Neisser & E. Winograd (Eds.), *Remembering reconsidered: Ecological and traditional approaches to the study of memory* (pp. 356–373). New York: Cambridge University Press.

Neisser, U., & Harsch, N. (1992). Phantom flashbulbs: False recollections of hearing the news about Challenger. In E. Winograd & U. Neisser (Eds.), *Affect and accuracy in recall: Studies of "flashbulb memories"* (pp. 9–31). New York: Cambridge University Press.

Neisser, U., & Winograd, E. (Eds.). (1988). *Remembering reconsidered: Ecological and traditional approaches to the study of memory.* New York: Cambridge University Press.

Nelson, T. O., & Chaiklin, S. (1980). Immediate memory for spatial location. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 529–545.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.), *The psychology of learning and motivation* (pp. 125–173). New York: Academic Press.

Nelson, T. O., & Narens, L. (in press). Why investigate metacognition? In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing.* Cambridge, MA: MIT Press.

Nilsson, L.-G. (1987). Motivated memory: Dissociation between performance data and subjective reports. *Psychological Research, 49,* 183–188.

Norman, D. A., & Wickelgren, W. A. (1969). Strength theory of decision rules and latency in retrieval from short-term memory. *Journal of Mathematical Psychology, 6,* 192–208.

Pick, H. L., & Lockman, J. L. (1981). From frames of reference to spatial representations. In L. S. Liben, A. H. Patterson, & N. Newcombe (Eds.), *Spatial representation and behavior across the life span* (pp. 39–61). New York: Academic Press.

Puff, C. R. (Ed.). (1982). *Handbook of research methods in human memory and cognition.* New York: Academic Press.

Richardson-Klavehn, A., & Bjork, R. A. (1988). Measures of memory. *Annual Review of Psychology, 39,* 475–543.

Riley, D. (1962). Memory for form. In L. Postman (Ed.), *Psychology in the making: Histories of selected research problems* (pp. 402–464). New York: Knopf.

Roediger, H. L. (1980). Memory metaphors in cognitive psychology. *Memory & Cognition, 8,* 231–246.

Roediger, H. L., & Payne, D. G. (1985). Recall criterion does not affect recall level or hypermnesia: A puzzle for generate/recognize theories. *Memory & Cognition, 13,* 1–7.

Roediger, H. L., Srinivas, K., & Waddil, P. (1989). How much does guessing influence recall? Comment on Erdelyi, Finks, and Feigin-Pfau. *Journal of Experimental Psychology: General, 118,* 253–257.

Ross, M. (1989). Relation of implicit theories to the construction of personal histories. *Psychological Review, 96,* 341–357.

Schacter, D. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 501–518.

Schacter, D. (1989). Memory. In M. I. Posner (Ed.), *Foundations of cognitive science* (pp. 687–725). Cambridge, MA: MIT Press.

Shepard, R. N. (1967). Recognition memory for words, sentences and pictures. *Journal of Verbal Learning and Verbal Behavior, 6,* 156–163.

Siegel, A. W. (1981). The externalization of cognitive maps by children and adults: In search of ways to ask better questions. In L. S. Liben, A. H. Patterson, & N. Newcombe (Eds.), *Spatial representation and behavior across the life span* (pp. 167–194). New York: Academic Press.

Siegel, A. W., & Schadler, M. (1977). Young people's cognitive maps of their classroom. *Child Development, 48,* 388–394.

Spence, D. P. (1982). *Narrative truth and historical truth.* New York: W. W. Norton.

Timm, H. W. (1983). The factors theoretically affecting the impact of forensic hypnosis techniques on eyewitness recall. *Journal of Police Science and Administration, 11,* 442–450.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381–403). New York: Academic Press.

Tulving, E. (1983). *Elements of episodic memory.* Oxford, England: Clarendon Press.

Tversky, B. (1981). Distortions in memory for maps. *Cognitive Psychology, 13,* 407–433.

Tversky, B., & Schiano, D. J. (1989). Perceptual and conceptual factors in distortions in memory for graphs and maps. *Journal of Experimental Psychology: General, 118,* 387–398.

Wagenaar, W. A. (1986). My memory: A study of autobiographical memory over six years. *Cognitive Psychology, 18,* 225–252.

Wagenaar, W. A., & Boer, H. P. A. (1987). Misleading postevent information: Testing parameterized models of integration in memory. *Acta Psychologica, 66,* 291–306.

Wallace, W. P. (1980). On the use of distractors for testing recognition memory. *Psychological Bulletin, 88,* 696–704.

Waterman, S., & Gordon, D. (1984). A quantitative–comparative approach to analysis of distortion in mental maps. *Professional Geographer, 36,* 326–337.

Weiner, B. (1966a). Effects of motivation on the availability and retrieval of memory traces. *Psychological Bulletin, 65,* 24–37.

Weiner, B. (1966b). Motivation and memory. *Psychological Monographs: General and Applied, 80* (18, Whole No. 626).

Weingardt, K. R., Leonesio, R. J., & Loftus, E. F. (in press). Viewing eyewitness research from a metacognitive perspective. In J.

Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing.* Cambridge, MA: MIT Press.
Wells, G. L., & Loftus, E. F. (Eds.). (1984). *Eyewitness testimony: Psychological perspectives.* Cambridge, England: Cambridge University Press.
White, R. T. (1982). Memory for personal events. *Human Learning, 1,* 171–183.
Winograd, E. (in press). Comments on the authenticity and utility of memories. In U. Neisser & R. Fivush (Eds.), *The remembered self.* Cambridge, England: Cambridge University Press.
Winograd, E., & Neisser, U. (Eds.). (1992). *Affect and accuracy in recall: Studies of "flashbulb memories".* New York: Cambridge University Press.
Yaniv, I., & Foster, D. P. (1990). Judgment, graininess, and categories. *Proceedings of the Annual Conference of the Cognitive Science Society, 12,* 130–140.
Yaniv, I., & Foster, D. P. (1993). *On graininess of judgement under uncertainty: An accuracy–informativeness tradeoff.* Manuscript submitted for publication.

Received December 8, 1993
Revision received April 6, 1994
Accepted April 11, 1994 ■