Vital and Health Statistics

Cognitive Processes in Long-Term Dietary Recall

Series 6:
Cognition and Survey Measurement
No. 4

This report describes a program of research designed to explore the cognitive processes and representations that subserve dietary reporting in surveys. Specific experiments investigated recall of which foods were eaten, estimation of the frequency with which foods are eaten, and judgments of portion sizes. Implications for cognition, surveys, and nutritional epidemiology are discussed.

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service
Centers for Disease Control
National Center for Health Statistics

Hyattsville, Maryland
October 1991
DHHS Publication No. (PHS) 92-1079

Suggested citation

Smith AF. Cognitive processes in long-term dietary recall. National Center for Health Statistics. Vital Health Stat 6(4). 1991.

Library of Congress Cataloging-in-Publication Data

Smith, Albert F.
Cognitive processes in long-term dietary recall.
p. cm. -- (Vital and health statistics. Series 6, Cognition and survey measurement ; no. 4) (DHHS publication ; no. (PHS) 92-1079)
By Albert F. Smith.
Includes bibliographical references.
ISBN 0-8406-0446-7
1. Reducing diets—Evaluation. 2. Food habits— Evaluation. 3. Recollection (Psychology) 4. Health surveys—United States. I. National Center for Health Statistics (U.S.) II. Title. III. Series: Vital & health statistics. Series 6. Cognition and survey measurement; no. 4. IV. Series: DHHS publication ; no. (PHS) 92-1079
[DNLM: 1. Cognition. 2. Diet Surveys. 3. Eating. 4. Food Habits. 5. Memory. W2 A N148vf no. 4]
RM222.2.S614 1991
613.2'5'0723—dc20
DNLM/DLC
for Library of Congress

91-13676
CIP

This research was part of a larger project, entitled "National Laboratory for Collaborative Research in Cognition and Survey Methodology," being conducted by the National Center for Health Statistics under grant SES-86123320 from the National Science Foundation (NSF). Monroe G. Sirken, Ph.D., was the principal investigator for this grant, and Murray Aborn, Ph.D., was the NSF program director.

This research was performed under Department of Health and Human Services contract 282-87-0039 from the National Center for Health Statistics to the Research Foundation of the State University of New York. Albert F. Smith, Ph.D., was the principal investigator. Jared B. Jobe, Ph.D., was the NCHS project officer and is the technical editor and reviewer for series 6. David J. Mingay was the alternate project officer.

Preface

Although my name is the only name on the byline of this report, many people have contributed to the work that is described. It is a pleasure to acknowledge them.

In my laboratory, I was assisted in the collection and analysis of data by a group of talented people that included Kathryn Murphy, Joseph Belluck, David Manzer, Adrian Clark, and Elsa Issa. Important contributions to specific experiments were made by Jody Layer and Deborah Prentice. Several colleagues in the Department of Psychology, particularly Patricia DiLorenzo and Ralph Miller, made constructive suggestions at many points during this project.

Without the extraordinary support of people in the Offices of Sponsored Program Development and Sponsored Funds Administration at the State University of New York at Binghamton—Heather Tomashek, Paul Parker, Dennis Saunders, and Joe Walker—it would have been impossible to carry out this work.

Much of this report was drafted at the Division of Biostatistics, Columbia University School of Public Health, where Patrick Shrout, Michael Parides, Alan Weinberg, and Carl Pieper provided an encouraging environment.

The project was vastly enriched by the enthusiasm of the National Center for Health Statistics personnel who have been involved in this research since its inception. Jared Jobe, David Mingay, Gordon Willis, and Monroe Sirken all made contributions that have vastly improved the research and the report.

Thanks to Deborah Prentice for the companionship and help that she has provided since the inception of this project and to Snack Smith for the companionship that he has provided for as long as he has been around.

Albert F. Smith
Binghamton, New York

Contents

Appendixes

List of text figures

List of text tables

Symbols

- - -	Data not available
. . .	Category not applicable
–	Quantity zero
0.0	Quantity more than zero but less than 0.05
Z	Quantity more than zero but less than 500 where numbers are rounded to thousands
*	Figure does not meet standard of reliability or precision
#	Figure suppressed to comply with confidentiality requirements

Cognitive Processes in Long-Term Dietary Recall

by Albert F. Smith, Ph.D., State University of New York at Binghamton

Summary

To quantify dietary intake, three types of information are needed— information about what individuals eat, how often they eat those items, and the sizes of the portions they eat. This report describes eight experiments that were conducted to investigate dietary reporting performance.

Dietary recall was studied under various combinations of acquisition, retention, and retrieval conditions. When asked to report their dietary intake for a specified reference period, subjects appear to rely largely on generic knowledge of their diets—they tend to report items that they are likely to have eaten or items that they routinely eat rather than items that they specifically remember having eaten during the designated period. This tendency increases as the time of the recall test becomes more remote from the reference period.

The accuracy of estimates of the frequency with which foods are eaten during a specified period also deteriorates as the amount of time between the end of the reference period and the frequency test increases. Individuals appear to remember, with reasonable accuracy, how often they ate various foods. However, there is sufficient disparity among individuals in their frequency calibration that the results of analyses of the frequency with which different individuals eat particular items—the sorts of analyses that are of most interest to epidemiologists—are less satisfactory.

Two experiments suggest that current methodologies are not satisfactory for collecting high-quality data on portion sizes. Individuals appear to have fragile memories about the sizes of food portions they view, and they tend to be insensitive to the definitions of portion size provided on food frequency questionnaires.

Introduction

Because diet is believed to play a significant role in determining health status, health scientists from various disciplines are interested in evaluating relationships between nutritional inputs and indexes of human health. Nutritionists, for example, are concerned especially with the influence of dietary intake on growth and maturation; epidemiologists are concerned with relationships between the intake of various nutrients and the incidence of specific diseases (1,2). Efficient evaluation of research hypotheses about the health consequences of diet requires accurate information about the dietary intake of individuals.

Collecting information about the dietary intake of individuals has long been recognized as a challenging problem in nutrition and epidemiology (3). Because people tend to eat frequently and to eat a variety of foods at any given meal and over time, tracking the dietary intake of any individual is difficult. To observe the diets of noncaptive individuals is intrusive; for individuals to keep records of their intake is burdensome. Hence, most information about dietary intake is obtained by conducting surveys, that is, by asking individuals to report what they have eaten during some period of time.

Surveys rely on respondents' memories, and so collectors and users of retrospective self-report data must be concerned about the accuracy of memory-based reports (4). Epidemiologists almost routinely express concern about the likely fallibility of memory and the possible consequences of memory error, not only for dietary survey data but for all survey data concerning health-related events and behaviors (5–9). Thus, researchers who collect dietary information invest substantial effort in assessing the accuracy of their data-collection techniques.

Despite much speculation about how memory and other cognitive processes influence dietary reporting, researchers have not explicitly investigated these processes. This report describes a program of research designed to explore the cognitive processes and representations that subserve dietary reporting. An improved understanding of these processes could enhance understanding of the nature of dietary survey responses, suggest the sorts of dietary data that might reasonably be collected, and guide the development of procedures to improve reporting accuracy. The general strategy of the research program was to approach divergently the processes hypothesized to be involved in dietary reporting, that is, to identify the components of the various tasks in which respondents to dietary surveys engage and to explore these in parallel.

In the remainder of this section, some of the techniques used to collect reports about dietary intake are described, the psychological questions that are motivated by considering these methods are discussed, and hypotheses about some of the cognitive processes involved in dietary reporting are presented. In subsequent sections, experiments that were conducted to address various aspects of the dietary reporting problem are described. In the final section, the implications of the experiments for the collection of dietary data are discussed.

Methods of dietary assessment

Three principal classes of survey methods are used to collect dietary information—the diet history, dietary recall, and food frequency procedures (2,3,10,11). All of these ask respondents to retrieve and report memories of their dietary experiences.

To obtain a diet history, a trained nutritionist conducts a detailed interview with a respondent, who is asked to describe the routine features of his or her diet during some period of time. The goal is not to obtain specific information about what the respondent has eaten, but rather to develop a general characterization of the respondent's "typical" diet.

Recall procedures ask respondents to report everything they ate or drank during a designated period; customary reference periods in epidemiology are 24 hours, 3 days, and 1 week. In some recall procedures, the respondent is asked simply to list the foods and drinks that come to mind; in others, he or she is "guided" mentally through the reference period by the interviewer. When participating in recall procedures, respondents may be asked to supplement their statements with descriptions of the portions of the items they ate. In any case, in contrast to the diet history, the recall procedure is intended to collect specific information about dietary intake. In principle, respondents' answers could be scored as correct or incorrect according to whether they matched some record of what the respondent actually ate and drank during the reference period.

Food frequency procedures require respondents to provide either count or rate information about their intake of each of a set of food and beverage items during some reference period. For example, a respondent might be asked, "How many artichokes did you eat during the last year?" or "How often did you eat artichokes during the last year?" The first of these questions asks for a count; the second asks for a rate. Appropriate (and equivalent) answers to these two questions would be 52 times and once per week, respectively. The customary reference periods of food frequency questionnaires range from 1 month to 1 year.

The usefulness of the data collected with any of these procedures in identifying associations between dietary intake and health outcomes depends on two related aspects of the data's validity. The first is the extent to which the respondents' reports are accurate; the second is the extent to which the respondents' reports are representative of their dietary intake. Intuitions about memory decay and dietary variability suggest that there is probably a tradeoff between these properties of dietary reports. The most accurate data may be the least representative, whereas the potentially least accurate data may be the most representative. An individual will presumably remember and report best the bite of food that he or she has just swallowed, but unless that individual eats only one foodstuff, the report of that just-swallowed item will not be representative of his or her diet nor reflect its variability. As the period of time about which an individual is asked to report is extended—bite by bite and meal by meal—into the past, dietary variability should be represented in reports. However, those reports about past intake would be increasingly likely to contain errors.

Validation of dietary assessment methods

Nutritionists and epidemiologists have invested substantial effort in evaluating the validity of reports about dietary intake. True validation of a survey method requires that dietary reports be compared to some record of the respondent's intake.[1] Typically, the record is presumed to be accurate, although it too may contain errors (3).

Consider two examples of the strategies used by epidemiologists to evaluate the validity of dietary reports: Hankin, Rhoads, and Glober (12) asked 50 individuals to keep a checklist diary for 7 days, indicating each occasion on which they ate any of 33 target food items and recording the quantity that they ate. On the day after the end of the recording period, subjects were interviewed about their intake of the 33 foods on the preceding day and during the preceding 7 days. For each food item and for each time period, Hankin et al. reported the proportion of subjects who reported having eaten the item, conditional on their having recorded it. Hankin et al. also reported the correlation, over subjects, between estimated quantity of intake from the record and estimated quantity from the report. Willett et al. (13) asked 173 nurses to keep diet records for 4 weeks and, subsequently, to complete a food frequency questionnaire. The validity of the questionnaire was assessed by estimating each individual's mean daily intake of 16 nutrients from both the records and the frequency questionnaire and by calculating the correlation, over subjects, of these estimates.

Two aspects of these studies are of special concern. First, studies of the validity of dietary reports are useful only insofar as their conclusions can be generalized to the reports of individuals for whom criterion information is not available. In general, respondents to dietary surveys have not kept records of their dietary intake, but it is the reporting accuracy of such individuals that is of general interest. If keeping dietary records modifies subsequent reports, the generalizability of validity estimates to individuals who have not kept records would be severely limited. Recordkeeping might improve memory directly, resulting in more accurate reports of dietary intake. The close attention to dietary intake required to keep a record might result in the encoding of information that is better than normal. Alternatively, recordkeeping might affect reports indirectly: Compared to individuals who have not kept records, respondents who have kept a record might exert greater retrieval effort when tested; alternatively, they might respond more conservatively, given that they know that the investigator possesses the true answers (in the records). Because typical survey respondents have not kept diaries, the performance of participants in validity studies may differ from the performance of subsequently surveyed respondents. To use diaries to assess the accuracy of subsequent reports requires first establishing that diarykeeping does not affect subsequent reports (14). Neither Hankin et al. (12) nor Willett et al. (13) checked the effect of recordkeeping on reports, although both groups of investigators were aware that recordkeeping might affect reports.

Second, there appears to be no general consensus among nutritionists and epidemiologists on standards of validity for dietary reports or even on how validity should be measured (2,3). Some investigators directly compare the foods reported by a respondent to the foods that he or she recorded, sometimes taking quantity into account (12). This procedure is most tractable for a closed set of foods—for example, the 33 foods studied by Hankin et al. (12). Recently, however, indirect scoring methods of the sort used by Willett et al. (13) have been more commonly used. Food composition tables are used to translate the recorded items and the reported items into average daily values of nutrients of interest. Then the association between records and reports is calculated for each nutrient, over respondents. Reports are declared valid if some criterion level of association is met. For example, Willett et al. (13) obtained 16 correlation coefficients, one for each nutrient, and made separate decisions about the validity of reports for each nutrient.

[1]Often, in nutritional epidemiology, comparing data obtained by one reporting technique to that obtained by another, more favored technique is called validation (3).

Such indirect approaches may fail to detect certain types of reporting errors that may be informative about what dietary information is available in respondents' memories and about the processes they use while reporting. Researchers who use indirect methods tend to treat target nutrients individually rather than as an interdependent set. Therefore, if respondents erroneously report having eaten items that are similar to consumed items in some nutrients but not in others, such analyses may falsely suggest that reports were accurate. This point is discussed in more detail in appendix I.

In sum, the decision about how validity is measured is significant. One might reasonably argue that the measure used to assess validity need not be the measure that is ultimately the focus of epidemiological analyses. For example, it may be appropriate to assess response validity using direct scoring methods, and then to transform responses to nutrient values for subsequent analyses of the relation of nutrient intake to health status. This report uses direct scoring methods to study the reporting performance of individual respondents.

Psychological processes in survey responding

Respondents to dietary surveys are asked what they have eaten or how often they have eaten various foods and in what quantities, and, if the survey instrument is considered valid, the responses are used to make scientific inferences about diet-disease relationships and to formulate health policy. During the last 10 years, survey methodologists have become increasingly sympathetic to the notion that analysis of the cognitive processes that subserve survey question responding can contribute to improved survey design and to the reduction of response error (15–18). An enhanced understanding of people's knowledge of their dietary intake might improve dietary survey data by guiding the development of answerable survey questions. Initial steps toward such an understanding of people's knowledge might be based on an examination of the types and frequency of errors exhibited by respondents to dietary surveys.

In this section, the cognitive demands of dietary survey procedures are considered, and relevant theory and data from cognitive psychology are summarized. The intent of this review is to illustrate what might be reasonably expected of a dietary survey respondent, to speculate about what sorts of errors might occur, and to suggest some hypotheses about and explanations of performance in dietary survey procedures.

Survey responding can be decomposed into question comprehension, information retrieval, and response formulation components; each of these aspects of answering survey questions may be addressed empirically (15–17,19,20). In dietary surveys, question comprehension includes understanding what sort of information is sought by the interviewer, understanding to what foods the interviewer is referring, and understanding references to and making judgments about groups of foods (for example, collard, turnip, and mustard greens). Information retrieval refers to searching memory for information relevant to a question. Response formulation refers to all of the mental processes involved in producing a response (for example, using specific information retrieved from memory, using norms believed to be relevant to the question, or using some combination of these).

Completely accurate dietary recall would require that the respondent have a cognitive representation of diet that included information about what he or she had eaten, some kind of temporal information, and quantity information. In addition, an exhaustive search process would be required to operate on this memory representation to ensure the retrieval of all items eaten during the designated period. The target information would have to be sufficiently discriminable for retrieval to be limited to the items eaten during the designated period.

To fulfill all of these requirements would be implausible. In general, long-term memory for any set of items is far from perfect (21). In contrast, memory for certain attributes of experience has proved to be quite accurate. For example, studies of memory for the frequency of occurrence of events have shown that people can make reasonably accurate judgments of relative frequency (22,23). To correctly report portion sizes would require either that the respondent have precise information about portion sizes in memory or that he or she be able to transform the contents of some representation into standard units of size. Little is known about long-term memory for size.

What sort of internal representation of dietary information might serve as the basis of responding in both the recall and food frequency procedures? Consider as a working hypothesis the following model, based loosely on Anderson and Bower's (24) two-process model of free recall: Suppose that an individual has general knowledge about his or her diet and that this knowledge can be represented as an associative network among the various foods that the individual eats. (Foods eaten occasionally may, of course, be added to the network; no specific position is taken on the relation of the individual's knowledge about his or her diet and that individual's general knowledge about food.) Suppose further that, for each occasion on which an individual eats a food item, a marker is associated to the mental representation of that food item and that these markers decay over time. In essence, then, there are two types of information in this representation. The network of representations of food items constitutes an individual's generic dietary knowledge about his or her own diet. The markers, which may be associated not only to the mental representation of food items but also to other contextual information, instantiate specific dietary memories. When asked to recall dietary intake, a respondent might use a memory search process that accesses an appropriately marked representation of a food item and that follows associative paths to other items.

To report frequency information, the respondent might carry out some kind of counting operation on the markers associated to the representation of the named food item (23).

A survey respondent asked to report dietary intake for a designated period is asked for specific dietary information and presumably attempts to report appropriate items—items whose cognitive representations are tagged with markers that reflect the appropriate time period. However, memory for dietary intake may be imperfect, and imperfect memory might result in two types of reporting errors: The respondent may fail to report items that he or she ate during the reference period (omissions) or may report items that he or she did not eat during the reference period (intrusions). Intrusions might be observed if the respondent's report combined reports of specific dietary memories with plausible inferences from generic dietary knowledge. Because it is implicit in health survey situations that respondents should be able to answer the questions, they will be inclined to answer, regardless of whether their knowledge is sufficient. Thus, respondents will likely report having eaten items during the specified period that they did not eat.

These expectations are consistent with the results of laboratory experiments on memory. Generally, in such experiments, subjects do not exhibit perfect performance when asked to recall a previously presented set of target items (for example, a list of words). Typically, when the to-be-recalled items are unrelated, errors are primarily omissions and intrusions are rare (25). As the retention interval (the time between the presentation of the target items and the test) is lengthened, the reported proportion of target items decreases, but there is only a modest increase in the intrusion rate (26). A quite different pattern is observed when the target items are related (for example, when they are exemplars of a single semantic category): Intrusions are observed, and the intrusion rate increases as the retention interval is lengthened (27). The items that an individual eats during any given period are related both by being members of a single semantic category and by being members of the more specialized category, "things that this individual eats." Thus, it is reasonable to expect that the recall of dietary intake would follow the recall pattern of related items.

For judgments of frequency, laboratory studies show that subjects are highly sensitive to variation in the frequency with which items are presented and estimate frequency quite accurately (22,23). Most systematic knowledge about the acquisition of frequency information comes from experiments in which subjects are exposed to a relatively small set of events, each of which occurs not more than a few times, and are tested immediately. Respondents to food frequency questionnaires are asked to make judgments about a stimulus list that did not have a well-marked beginning, that occurred over an extended period, and that contained events that occurred with a wide range of frequencies. Controlled experiments that evaluate performance under these conditions have not been reported in the psychological literature. Nevertheless, data concerning memory for frequency suggest that people may reasonably be expected to estimate frequencies accurately.

The next sections of this report describe experiments that addressed these issues. Recall of foods eaten, judgments of frequency, and judgments of portion size are addressed in separate experiments. The goal of these experiments was to examine some of the cognitive processes suspected to underlie dietary survey responding in tasks that approximate those used in nutritional surveys. The reader is advised from the outset that these are psychological experiments, designed to address psychological questions that appear relevant to some dietary survey measurement problems. Therefore, the specific information collected in the various experiments and the focuses of the analyses are not necessarily identical to the information that would be collected by nutritionists or epidemiologists or to the analyses that would be carried out by investigators in those disciplines.

Dietary recall

In the dietary recall procedure, the respondent is asked to think back over some period of time and to report everything that he or she ate or drank during that period. Although there is general consensus that respondents err in the dietary recall procedure, neither a complete description nor a full explanation of these errors has been advanced. To develop such an account requires identifying the nature of the information that is available in the memories of respondents. An understanding of what information is available could help clarify what sorts of questions might reasonably be asked of respondents to dietary surveys.

Two hypotheses about performance in the dietary recall procedure stem directly from the model of the representation and retrieval of dietary information proposed in the introduction. First, because of the decay of the markers that represent the particular occasions on which food items are consumed, the contribution to reports of specific memories should decrease as the length of the retention interval increases. Second, as the length of the retention interval increases, respondents' reports will consist increasingly of reports of generic dietary information. The experiments reported in this section were conducted to collect data that would help evaluate these predictions.

The general strategy of these experiments was to ask people, after the end of a period during which they had recorded their dietary intake, to report what they had eaten during that period. Reporting accuracy was assessed by comparing the reports to the diary records. To test specific issues of theoretical concern, various aspects of the conditions under which subjects recorded and reported their dietary intake were manipulated. These variables included the duration of the diarykeeping period, the amount of time between the end of the diarykeeping period and the memory test, and the reporting instructions given to subjects at the time of the memory test. These experiments also addressed several methodological issues, the resolution of which was necessary for meaningful interpretation of the accuracy data.

In this section, three experiments that involved two groups of subjects are described. Experiment 1, conducted with a group of university undergraduates, evaluated whether diarykeeping is a suitable method for validating subsequent reports, whether participating in a recall procedure prior to keeping the diary would influence subsequent performance, and whether reporting performance depends on retention interval. Experiment 2 exploited experiment 1's findings concerning the suitability of diarykeeping to explore further, with a heterogeneous community sample, the effect of the duration of the retention interval on reporting performance. In addition, experiment 2 examined reporting performance for reference periods of different lengths and, indirectly, the organization of dietary memory. In each of these experiments, analyses of the data were concerned only with recall of items that were eaten during the diarykeeping period, not with the frequency with which those items were eaten nor with the quantities in which they were eaten. Experiment 3, conducted with subsets of the participants in experiments 1 and 2, assessed the ability of individuals to recall a set of food items that they had listed prior to diarykeeping.

Experiment 1: Effects of recordkeeping and retention interval

To develop a model of dietary recall performance requires an understanding of the factors that influence reporting accuracy. However, to evaluate accuracy of dietary reports requires an acceptable technique for validating responses. The most obvious possibility is a written record, but to use this technique requires establishing that recordkeeping does not affect subsequent reports about the recorded events (14).

Similar considerations apply to the administration of a dietary recall procedure prior to any recordkeeping. Data collected in an initial recall trial have a variety of potential uses in research of this type. However, if participation in such a procedure affects a subject's subsequent performance, it would be inappropriate to conduct the procedure.

This experiment was conducted to evaluate diarykeeping as a method for validating responses, to determine whether a preliminary recall task influenced subsequent reports, and to examine the effects of variation in the length of the retention interval on performance.

Method

Subjects—Ninety-six university undergraduates were recruited for a two-session study of health-related behavior. Subjects were given credit toward the research participation requirement of their introductory psychology course and were paid $12.50 to keep a diary for 1 week.

Design—To determine whether diarykeeping influenced reporting performance, half of the subjects were assigned to record their food and drink intake for 1 week (food diary condition); the remaining subjects recorded either their mass media consumption (books, newspapers, magazines, radio, television, etc.), their social interactions, or their telephone activity (nonfood diary condition). To evaluate whether engaging in the dietary recall procedure prior to recording would influence subsequent performance, half of the subjects were asked at the beginning of the first laboratory session to report their dietary intake for the preceding week (initial recall); the remainder did not complete this task (no initial recall). Finally, to study the nature of reporting performance as a function of time since the diarykeeping period, half of the subjects were assigned to return on the day after the last day of the diarykeeping week (immediate test); the remainder were scheduled to return 3 weeks later (delayed test).

Procedure—On the subjects' first visits to the laboratory, those assigned to the initial recall condition attempted to recall their food intake for the preceding week. All subjects were instructed on how to keep their diaries. Food diary subjects were asked to record every item that they ate or drank during the next 7 days; nonfood diary subjects were asked to record one of the alternate sets of events described above. Each subject was given a set of diary forms and envelopes, was asked to start a new form on each day of the diarykeeping period, and was instructed to return the preceding day's form each day to one of three secure campus locations. Subjects engaged in several additional tasks that are irrelevant to the present discussion, were scheduled to return for their second sessions, and were dismissed.

In their second sessions, all subjects were asked to recall their food intake for the week during which they had kept the diary. Subjects were asked to report whatever they could remember eating and drinking during that week. Subjects reported orally, and reports were tape-recorded. Then, those subjects who had engaged in the initial recall task were asked also to attempt to recall the set of items they had listed during the first session. (Analyses of these reports are described as experiment 3.)

Data analysis

Two types of data are reported. For some of the research questions, the appropriate measures are counts of items recorded in the diaries or reported during the recall test. When simple counts are reported, the number of distinct items (that is, the number of items after the list of recorded or reported items has been edited to remove duplicates) is of usual interest. For other questions, the appropriate measures are indexes of the correspondence between the items reported during the recall test and those recorded in the diary.

To score the dietary reports, each set of items (the reported items and the recorded items) was edited to eliminate duplicate entries and to consolidate different labels for single items (for example, "green salad" and "tossed salad" would have been combined). After editing, each distinguishable item that a subject recorded or reported was treated as having occurred only once in each set. The edited sets of items were then classified into three mutually exclusive groups: Reported items that had been recorded in the diary were classified as matches; recorded items that were not reported were labeled omissions; reported items that were not recorded were called intrusions.[2] These counts were used to calculate the match and intrusion rates of each subject. The match rate is the proportion of recorded items that were reported (the ratio of matches to the sum of matches and omissions), and the intrusion rate is the proportion of reported items that were intrusions (the ratio of intrusions to the sum of matches and intrusions). (These measures and the rationale for their use are discussed more fully in appendix II.)

To assess the reliability of these scoring procedures, two individuals scored the reports of 24 subjects. The correlations between the resulting scores were 0.95 for match rates and 0.83 for intrusion rates. Thus, the judgments of individual scorers were deemed acceptable for the protocols of the remaining subjects.

Results and discussion

Two subjects were dropped from the study because they were unable to return for the second session at their scheduled times. Each of the other subjects submitted the seven required daily diaries and returned for the second session according to the prescribed schedule. This experiment was run as two replications of 48 subjects; one analysis is reported only for data collected in the first replication.

Effect of diarykeeping—Logically, the first issue of concern is whether recordkeeping influenced performance. If the subjects' reports depended on whether they had kept a food diary, then the usefulness of diaries as a validation method would be suspect. To answer this question requires comparing the reports of subjects who kept food diaries to those of subjects who did not. Thus, some characteristic of the subjects' reports other than accuracy (for example, number of items reported) must be

[2]For purposes of this report, intrusions are defined strictly as reported items that had not been recorded. It is always possible that a reported item was eaten, but not recorded, in which case it would be scored as an intrusion despite its having been eaten during the reference period. To conduct research of the type decribed in this report, some standard has to be accepted as representative of the truth; in these experiments, the diaries were taken as this standard. Smith et al. (28), in experiment 2, showed that subjects were highly consistent, over recording periods, in how many distinct items they recorded. This suggests that the subjects in these experiments took seriously the diarykeeping task and that the quality of subjects' diaries did not deteriorate over time.

examined, because there is no way to evaluate the accuracy of the reports of subjects who did not keep diaries.

Table A shows the numbers of items reported by subjects in the two diary conditions for each retention interval. The type of diary kept by the subjects did not influence systematically how many items they reported ($F(1,90) < 1$), and the interaction of the length of the retention interval and diary type was not significant ($F(1,90) = 2.49$, $p > 0.10$). Not surprisingly, the number of reported items depended on the length of the retention interval. Subjects who were tested immediately after the reference period reported significantly more items than did those who reported 3 weeks later ($F(1,90) = 5.81$, $p < 0.02$). Taken together, these results indicate that diarykeeping is a suitable validation procedure for these studies. (For a fuller discussion, see Smith et al. (14).)

Effect of initial recall—Data from subjects' reports about their dietary intake for a period that precedes the diarykeeping period have several potential uses in research of this type. However, to collect such data would be appropriate only if the procedure did not affect later performance. If, when recalling for the diarykeeping period, subjects who had engaged previously in such a procedure reported more items or reported more accurately than subjects who did not, the procedure should not be carried out. Again, this is because the typical respondent to a dietary survey has not engaged in such a previous recall attempt, and one of the research objectives was to collect data that are at least somewhat generalizable to such respondents.

The initial recall procedure affected neither the count nor the accuracy measure of performance. On average, subjects who had engaged in the initial recall procedure reported 38 different items, whereas those who had not reported 36.2. This difference was not significant ($F(1,90) = 1.33$), and the initial recall variable did not interact with any other experimental variable. Among subjects who had kept food diaries, subjects who had participated in the initial recall procedure did not differ on the accuracy indexes from those who had not ($F(2,42) < 1$). The match and intrusion rates of subjects who participated in the initial recall procedure were 0.48 and 0.30, respectively; for subjects with no initial recall, these rates were 0.46 and 0.26, respectively.

Retention interval—The major substantive question of this experiment was whether reporting performance depends on the length of the retention interval and, if so,

Table A. Number of items reported by subjects in experiment 1, by retention interval and diary type

Diary type	Retention interval: Immediate test	Retention interval: Delayed test	Mean
	Number		
Food	39.4	36.2	37.9
Nonfood	43.7	28.3	36.2
Mean	41.5	32.3	37.0

NOTE: 24 subjects participated in each diary group that received an immediate test; 23 subjects participated in each diary group that received a delayed test.

how. Two analyses are reported—an overall analysis of accuracy for the entire diarykeeping period (as described above), in which a reported item was scored as a match if it had been recorded at any time during the diarykeeping period, and an analysis by days, in which a match rate was calculated separately for each day of the diarykeeping period.

Consistent with extensive data from other memory experiments, reporting performance deteriorated as the retention interval was lengthened—the average match rate decreased, and the average intrusion rate increased. For subjects who were tested immediately after the end of the diarykeeping period, match and intrusion rates were 0.55 and 0.22, respectively; for subjects who were tested after a 3-week delay, the match and intrusion rates were 0.38 and 0.34, respectively. The subjects in the two retention interval groups differed significantly on each measure: For match rate, $F(1,43) = 17.36$, $p < 0.0001$; for intrusion rate, $F(1,43) = 7.82, p < 0.01$.

If the probability of reporting a consumed item decreases with the amount of time that has elapsed since eating that item, one would expect, at least for subjects tested immediately after the end of the diarykeeping period, that memory for intake on the later days of the reference period would be superior to memory for intake on the earlier days. This hypothesis was evaluated by calculating, for each subject in the first replication of the experiment, a match rate for each day of the reference period: The number of matches for each day was divided by the number of items recorded on that day. Then, for each subject, these daily match rates were regressed on day of the reference period; the slope of the estimated regression line (change in match rate per day) was taken as an index of memory stability. For example, a slope of 0 would indicate the absence of any systematic change in match rate over the days of the reference period, whereas a positive slope would indicate higher match rates for the later days of the reference period than for the earlier ones.

The daily match rates of subjects tested immediately after the diarykeeping period increased from the early days of the reference period to the later ones, whereas those of subjects tested after a 3-week delay showed no systematic relationship to day of the reference period. The mean slopes for subjects in the immediate- and delayed-test conditions were 0.028 and –0.001, respectively; these differed significantly, $t(45) = 3.28, p < 0.005$. In terms of the model of memory proposed in the introduction, these results suggest that specific memories deteriorate day by day, so that the reports of subjects tested after a 3-week delay are likely to consist less of specific dietary memories than of generic dietary information. This hypothesis is pursued further in experiment 2.

Experiment 2: Effects of retention interval and instructions

Experiment 1 showed that diaries were a suitable standard against which reports of dietary intake could be scored, that subsequent performance was not affected by engaging in a recall trial prior to recording, and that dietary reporting performance decreased as the retention interval increased. The dependence of reporting performance on the retention interval was shown in both a general and a more refined way: Overall performance was better for subjects tested immediately after the retention interval than for those tested after a 3-week delay; more specifically, the match rates of the former subjects were, on average, best for the last day of the reference period and deteriorated back toward the first day at a rate of about 3 percent per day.

Subjects in the delayed-test condition of experiment 1 were tested 29 days after the beginning of the reference period. Their match rates were considerably above zero, but their daily match rates did not depend on the day of the reference period. Obviously, then, the reports of these subjects were based, at least in part, on memories other than those whose deterioration was revealed by the systematic decline in daily match rate for subjects in the immediate-test condition. Experiment 2 was designed to pursue the hypothesis that individuals' reports of dietary intake are based, in large part, on generic knowledge about their diets.

To study the dietary reporting performance of individuals who were more similar to national health survey respondents than are university undergraduates, a heterogeneous sample of community volunteers was recruited. Three variables were manipulated to evaluate specific hypotheses about how dietary intake is represented in memory and, in particular, to examine the proposed distinction between generic knowledge and specific dietary memories. The length of the retention interval was varied to replicate and extend, over a greater span of time, the manipulation of experiment 1. The observed effect of the length of the retention interval in that experiment was interpreted as evidence that specific dietary memories become unavailable over time. The length of the reference period was manipulated to evaluate the notion that there is a tradeoff, reflected in performance measures, between dietary representativeness and reporting accuracy: Because people have the opportunity to eat a greater variety of items during a long period than during a short one, reporting performance for longer reference periods should be less sensitive to the length of the retention interval than that for shorter reference periods. Reporting instructions were manipulated to test the hypotheses that different sets of instructions—in the form of cues intended to help respondents retrieve memories—might provide different entry points into memory and that different patterns of responses in different instructional conditions might be informative concerning the organization of dietary memory.

Method

Subjects—Subjects were recruited by advertising in the Binghamton, New York, area for participants in a study of health-related everyday behavior. Subjects were assigned randomly to conditions, but under the constraint that the ages and genders of the subjects be roughly balanced over the eight experimental conditions defined by the lengths of the reference period and of the retention interval. Data from 170 subjects are reported; table B classifies these subjects by demographic categories and experimental conditions and shows also how many subjects failed to complete the experiment. Data from seven subjects who completed the procedure were lost before they were analyzed. Subjects were paid $16 for the laboratory sessions and $12.50 per week for diarykeeping.

Design—Eight experimental conditions were defined by the orthogonal combination of two reference period lengths and four retention intervals. Within each of these conditions, subjects were assigned at random to three instruction conditions. Approximately half of the subjects were assigned to keep a diary for 2 weeks; the remainder were asked to keep a diary for 4 weeks. Approximately one-quarter of the subjects were assigned to each of the four retention interval conditions—immediately after the reference period (0 weeks) or 2, 4, or 6 weeks after the end of the diarykeeping period. When asked to recall their dietary intake for the reference period, subjects were assigned randomly to receive one of three different sets of recall instructions—to report in reverse chronological order, to report by food groups, or to report by meal. These instructions involved presentation of prompts, or cues (e.g., "meats" and "fruits" for the food group instructions; "breakfast" and "lunch" for meals; a calendar marked with dates for reverse chronological).

Procedure—Subjects attended two laboratory sessions and kept a food diary. In the first session, approximately half of the subjects attempted to recall what they had eaten and drunk during the preceding 2 or 4 weeks, depending on the length of time for which they would be asked to keep a diary. Then all subjects were instructed on how to keep the food diary. They were asked to record each item that they ate or drank during their assigned

Table B. Number of subjects in experiment 2, by demographic category and experimental condition

Condition	Total	Demographic category								Drop
		F–1	F–2	F–3	F–4	M–1	M–2	M–3	M–4	
	Number									
Target	24	4	5	5	2	2	3	2	1	0
Total	170	27	37	36	15	12	18	20	5	14
2–0	24	2	7	6	1	1	2	4	1	0
2–2	23	4	3	5	3	2	2	4	0	0
2–4	19	3	6	3	2	1	2	2	0	4
2–6	22	4	5	4	1	2	3	2	1	1
4–0	21	3	5	4	2	2	2	2	1	1
4–2	22	4	4	5	2	2	2	2	1	2
4–4	19	4	3	4	2	1	3	2	0	2
4–6	20	3	4	5	2	1	2	2	1	4
Drop	14	3	2	3	0	3	2	1	0	...

NOTE: Conditions are designated by duration of the reference period, in weeks, and the retention interval, in weeks (for example, 2–0 refers to a 2-week reference period and a 0-week retention interval). Demographic groups are designated by gender (M = male; F = female) and age category (1 = 21–29; 2 = 30–44; 3 = 45–64; 4 = 65 and over). Drop refers to individuals who failed to complete the experiment. Data from 7 participants who completed the experiment were lost prior to data analysis and are therefore not included in this table.

reference periods. Subjects were given a supply of forms and envelopes, told to start a new form each day, and asked to submit their records by mail twice per week. Finally, subjects were told that a followup session was required, and second sessions were scheduled according to the requirements of the retention interval manipulation. There was no indication that the followup session would involve memory tests.

In the second session, subjects were asked to report all items that they had eaten or drunk during the diary-keeping period. Each subject was given one of three sets of instructions for reporting: Reverse chronological order, food groups, and meals. Subjects were told that these cues might help them organize their reports. They were told that they should not feel constrained by the cues, but rather that they should report whatever came to mind. Subjects engaged in several additional tasks, some of which will be described in subsequent sections of this report, and were debriefed.

Results

Although the experimental participants were generally very cooperative in submitting their diaries, some were not received. For the 2-week and 4-week recording periods, diaries were missing for 6 subject-days and 33 subject-days, respectively. These constituted 0.5 percent and 1.4 percent, respectively, of all the subject-days for the two conditions. It was assumed that these losses would have negligible effects on the results.

The data were edited and scored using the procedures described for experiment 1. Ninety-six report protocols were scored by two judges: The correlations between the match and intrusion rates calculated by the two judges were 0.91 and 0.90, respectively.

To enhance the clarity of exposition, the principal results are described here, but presentation of detailed results, including values of test statistics, is deferred to appendix III. Unless otherwise noted, statements concerning the effects of experimental manipulations are made with a p of 0.01 or less, and statements concerning the lack of effects are made with a p of 0.25 or greater. To further simplify the presentation of the results, the effects of the length of the retention interval and the reference period length are discussed first, and then the effect of the reporting instructions is examined.

Effect of retention interval and reference period length—Table C shows the mean number of items recorded by subjects in each of the conditions defined by the lengths of the reference period and of the retention interval. The length of the reference period affected the number of items recorded in an orderly and sensible way: Subjects who kept diaries for 4 weeks recorded more different items (but only about 1½ times as many) than did subjects who kept diaries for 2 weeks. As some items are eaten on multiple occasions during the reference period, the rate of growth of the to-be-reported set of items is negatively accelerated. Because the retention interval followed the recording period, it should have had no effect on the number of items recorded. Table C shows that, indeed, the retention interval did not affect the mean number of items recorded by subjects during either reference period. There was no interaction of the lengths of the reference period and of the retention interval on the number of items recorded.

Table C. Number of items recorded by subjects in experiment 2, by reference period length and retention interval

Retention interval	Reference period length: 2 weeks	4 weeks	Mean
	Number		
0 week	62.9	107.1	83.5
2 weeks	63.4	105.5	84.0
4 weeks	69.7	108.7	89.2
6 weeks	71.8	104.3	87.3
Mean	66.7	106.4	85.9

NOTE: For number of subjects per condition, see table B.

Table D shows the mean match and intrusion rates for subjects in the eight experimental conditions defined by reference period length and retention interval. Two features of these results are particularly salient. First, the average match rate over experimental conditions was only 0.38; among the conditions, the highest mean match rate was only 0.49. Second, on average, approximately one-third of the reported items were intrusions; in no condition was the mean intrusion rate less than 0.27. In general, then, the subjects' free reports of their dietary intake did not effectively describe their intake during the specific period about which they were asked.

The pattern of match and intrusion rates shown in table D indicates, consistent with the results of experiment 1, that dietary reporting performance deteriorated as the length of the retention interval increased. Consistent with other data on memory decay, the decrease in these rates was negatively accelerated. (All tests of the effect of retention interval length reported in this section used a contrast with a negatively accelerated "shape": The weights applied to the 4- and 6-week levels of retention interval were equal and, in sum, opposite to the weight applied to the 0-week level.)

Although the average match rates of subjects assigned to the two reference periods did not differ, subjects who reported for a 2-week period exhibited higher intrusion

Table D. Match and intrusion rates for subjects in experiment 2, by reference period length and retention interval

	Reference period length					
	2 weeks		4 weeks		Mean	
Retention interval	p(mtc)	p(int)	p(mtc)	p(int)	p(mtc)	p(int)
			Rate			
0 week	0.49	0.27	0.46	0.29	0.48	0.28
2 weeks	0.42	0.39	0.37	0.28	0.39	0.34
4 weeks	0.28	0.40	0.34	0.34	0.31	0.37
6 weeks	0.30	0.42	0.32	0.33	0.31	0.38
Mean	0.38	0.37	0.37	0.31	0.38	0.34

NOTE: *P*(mtc) is the match rate; *p*(int) is the intrusion rate. For number of subjects per condition, see table B.

rates than did subjects who reported for 4 weeks. Further, as shown in table D, the mean intrusion rate increased quite steeply over retention intervals for subjects in the 2-week reference period condition but increased only very slightly for those in the 4-week reference period condition. (The interaction of reference period length and retention interval was significant at $p < 0.05$.)

Effect of reporting instructions—At the time of the memory test, the subjects were given one of three sets of retrieval cues. They were told that these were intended to aid rather than to constrain their reports and that they should feel free at all times to say whatever came to mind. Because subjects could use any preferred reporting strategy, the instruction manipulation was not expected to affect performance markedly. This prediction was incorrect; the observed results are potentially informative concerning the organization of memory for foods and the way in which personal dietary experiences are retrieved from this memory.

Table E shows how many items, on average, were recorded and reported by subjects who received each of the three sets of instructions for each reference period length. Appropriately, the reporting instructions did not influence the number of items recorded during the diarykeeping period. However, subjects who received food group cues reported significantly more items than did those who received reverse chronological or meal cues; the mean numbers of items reported by subjects in the latter two conditions did not differ.

Table F shows the mean match and intrusion rates for subjects in the three instruction conditions. Match rates did not differ significantly over the three conditions, and none of the interactions of instructions with other experimental conditions was significant. However, intrusion rates did differ among the instruction conditions. Subjects who received food group cues exhibited significantly higher intrusion rates than did subjects who received the reverse chronological or meal cues; the intrusion rates of subjects in these latter two conditions did not differ.

Table E. Numbers of items recorded and reported by subjects in experiment 2, by instruction condition and reference period length

Reference period	Instructions			
	Backward	*Food group*	*Meal*	*Mean*
	Number of items recorded			
2 weeks	67.4	67.0	65.9	66.7
4 weeks	100.2	110.1	106.2	106.4
Mean	82.3	88.9	85.1	85.9
	Number of items reported			
2 weeks	35.5	46.9	36.6	42.6
4 weeks	51.2	66.0	56.2	59.0
Mean	42.6	56.6	45.9	49.2

NOTE: Number of subjects per condition was as follows (from left to right and top to bottom): 23, 32, 33, 19, 33, and 30.

Table F. Match and intrusion rates for subjects in experiment 2, by instruction condition and retention interval

Retention interval	*Instruction*							
	Backward		*Food group*		*Meal*		*Mean*	
	p(mtc)	p(int)	p(mtc)	p(int)	p(mtc)	p(int)	p(mtc)	p(int)
	Rate							
0 week	0.51	0.17	0.48	0.34	0.46	0.26	0.48	0.28
2 weeks	0.46	0.28	0.39	0.38	0.35	0.33	0.39	0.34
4 weeks	0.26	0.34	0.33	0.42	0.34	0.37	0.31	0.37
6 weeks	0.31	0.40	0.33	0.40	0.28	0.33	0.31	0.38
Mean	0.37	0.31	0.39	0.38	0.36	0.32	0.38	0.34

NOTE: *P*(mtc) is the match rate; *p*(int) is the intrusion rate. For number of subjects per cell, see table E.

Discussion

These results build on those of experiment 1 in suggesting the important contribution of generic dietary knowledge to dietary reporting. Three features of the results particularly encourage this conclusion.

The first is the retention interval effect: As the retention interval was lengthened, reporting performance declined to the point that only 30 percent of recorded items were reported and 40 percent of reported items were intrusions. Even for the shortest retention intervals (which, in this experiment, admittedly followed long acquisition periods), subjects reported only half of the items they had recorded and exhibited intrusion rates of about 30 percent. People, when reporting their dietary intake for extended periods, clearly do not have access to memory representations that include accurate temporal information. Specific memories surely contribute to dietary reports: After all, match rates are highest for subjects tested immediately after the end of the diarykeeping period, and, as was seen in experiment 1, subjects tested immediately after the end of a 1-week diarykeeping period report more items from the last days than from the earlier days of the reference period. Nevertheless, dietary reports appear to consist, in large part, of individuals' guesses about what they probably ate. Intrusions must have some cognitive origin, and generic knowledge about diet is a likely source. If subjects have general knowledge about their diets but imperfect representation of when they ate various foods, then, when they are asked to recall what they have eaten for a period of 2 or 4 weeks, they may list foods that they routinely eat without regard to when they ate them.

The second result that encourages the generic knowledge hypothesis is the interaction of the retention interval and the reference period length on the intrusion rate. The absolute intrusion rates of subjects who reported for a 4-week period did not depend on the length of the retention interval, whereas those of subjects who reported for a 2-week period increased as a function of the length of the retention interval. Because subjects who reported for a 4-week period had more opportunity, during their diarykeeping periods, to eat the routine elements of their diets, they had considerably more latitude than did the

2-week subjects to misremember and misdate their dietary experiences. When accuracy is scored on the basis of whether an item was ever eaten during the reference period, the longer the reference period is, the more accurate subjects can be. To the extent that they can think of items to report, the greater will be the convergence between what they are supposed to report and what they do report. If, as the retention interval is increased, individuals' reports degenerate to descriptions of their generic diets, reports and records would become increasingly discrepant for subjects reporting for the 2-week interval, and this is precisely what was observed. These results are consistent with those of laboratory experiments involving the recall of lists of categorically related words, which show that intrusion rates increase as retention intervals increase and, necessarily, decrease as the number of to-be-remembered words from a category increases (25,29,30).

The third result that supports the conjecture concerning the contribution of generic dietary knowledge to reports is the observed effect of reporting instructions on intrusion rates. The subjects' general knowledge about food may be organized by food groups—food group instructions increased the number of items that subjects reported. It is of particular interest that these instructions elevated only the intrusion rate, not the match rate. In other words, food group cues appear to have given subjects mental access to more responses but not to the specific responses that would have improved performance.

Experiment 3: Free recall of an unambiguous target list

Individuals' dietary reports appear to be based on generic knowledge of their diets. One might be concerned that the patterns of results suggesting this conclusion are an artifact of the unusual conditions that prevailed during the acquisition of the to-be-reported items. Although the results of experiments 1 and 2 accord reasonably well with relevant published findings on free recall of related items, a distinctive feature of experiments 1 and 2 was that acquisition occurred over very long periods of time. Because irrelevant events were interspersed, during the acquisition phase, with the to-be-reported events, subjects may have been confused about exactly what items were to be reported. Experiment 3 was designed to examine recall of an unambiguously defined list of dietary items. Each subject was asked to report a set of food items that he or she had listed during a single brief period. Specifically, each subject who had been assigned to the initial recall condition of experiment 1 or 2 was asked to recall the set of items that he or she had recorded during the first laboratory session. Because that list was created during a discrete period, there should have been no ambiguity at the time of recall about what constituted the target set.

Method

Subjects—The two groups of participants in this experiment were those subjects who completed the initial recall task in experiments 1 and 2. For this experiment, these groups will be labeled groups 1 and 2, respectively.

Design and procedure—Subjects assigned to the initial recall conditions of experiments 1 and 2 were asked during their first laboratory sessions to write down everything that they had eaten and drunk during the preceding week (experiment 1) or during the preceding 2 or 4 weeks (experiment 2), depending on the amount of time for which they would subsequently be asked to keep a diary. In the second session, following the dietary recall procedure, subjects were asked to recall the items that they had recorded during the first laboratory session. For group 1, two retention intervals (1 week and 4 weeks) were crossed with two types of diary that had been kept during the reference period (food and nonfood). For group 2, four retention intervals were crossed with two reference period lengths.

Results and discussion

The recall protocols were scored in the fashion described for experiment 1, with the initially recorded items serving as the standard set. Table G shows mean match and intrusion rates for subjects in group 1 by type of diary and length of retention interval. Although the mean match rate was higher and the mean intrusion rate was lower for subjects who recalled their list after 1 week than for those who recalled their list after 4 weeks, the differences were quite small—0.06 for match rate and 0.07 for intrusion rate. Numerically, the effect of the retention interval on these data was substantially smaller than that observed for reports of intake during the reference period. A multivariate analysis of variance (MANOVA) on these data showed a significant effect of retention interval ($F(2,42) = 3.28, p < 0.05$) but neither a significant effect of diary type ($F(2,42) = 1.94, p > 0.10$) nor a significant interaction ($F(2,42) < 1$).

Table H shows the match and intrusion rates for subjects in group 2 by retention interval and reference period length. Neither of these variables affected systematically either of the measures of correspondence between

Table G. Match and intrusion rates for subjects in experiment 3, group 1, by retention interval and diary type

	Retention interval					
	Immediate test		*Delayed test*		*Mean*	
Diary type	*p*(mtc)	*p*(int)	*p*(mtc)	*p*(int)	*p*(mtc)	*p*(int)
			Rate			
Food	0.49	0.26	0.40	0.29	0.44	0.28
Nonfood	0.52	0.26	0.48	0.37	0.50	0.32
Mean	0.50	0.26	0.44	0.33	0.47	0.30

NOTE: *P*(mtc) is the match rate; *p*(int) is the intrusion rate.

Table H. Match and intrusion rates for subjects in experiment 3, group 2, by reference period length and retention interval

Retention interval	Reference period length: 2 weeks p(mtc)	2 weeks p(int)	4 weeks p(mtc)	4 weeks p(int)	Mean p(mtc)	Mean p(int)
	Rate					
0 week.	0.41	0.45	0.43	0.48	0.42	0.47
2 weeks	0.46	0.40	0.48	0.44	0.47	0.42
4 weeks	0.38	0.43	0.43	0.48	0.41	0.46
6 weeks	0.41	0.48	0.44	0.45	0.42	0.47
Mean	0.42	0.44	0.45	0.46	0.44	0.45

NOTE: *P*(mtc) is the match rate; *p*(int) is the intrusion rate.

the two sets of responses: For reference period, $F(2,73) < 1$; for retention interval, $F(6,146) < 1$; for the interaction of reference period and retention interval, $F(6,146) < 1$. This is not to say that the length of the reference period had no effect on performance. On average, both during the preliminary recall and during the recall of that list, subjects whose reference period was 4 weeks recorded and recalled more items than did subjects whose reference period was 2 weeks. The average numbers of items recorded during the preliminary recall task were 45.2 and 55.5 by subjects whose reference periods were 2 and 4 weeks, respectively; the average numbers recalled by these two groups of subjects were 32.6 and 47.8. Subjects working with a 4-week reference period record and report more items than do subjects whose reference period is 2 weeks. However, for neither these counts nor for the match and intrusion rates was there any dependence, for subjects in group 2, on retention interval.

Subjects appear to have been relatively indifferent to what they recorded at the initial laboratory session. The high intrusion rates suggest that at the time of recall, and quite possibly at the time of initial recording as well, subjects simply produce a list of items that they are likely to have eaten.

General discussion

The results of experiments 1–3 converge to suggest that individuals who are asked to recall their dietary intake for a specific extended period rely increasingly on generic knowledge about their diets as the time delay from the reference period to the recall test increases. Although specific memories clearly contribute to dietary reports, those memories are rapidly lost.

Several types of evidence lead to this conclusion. The deterioration in reporting performance over time from the end of the reference period, reflected both in decreasing match rates and in increasing intrusion rates, suggests that the representation of temporal information associated with routine dietary intake decays. However, match rates do not drop to zero. Even 6 weeks after the end of a 2-week reference period, the mean match rate was 0.30. If subjects use generic dietary knowledge, they would be expected to report from that knowledge even if they had no specific memories of the reference period (31). The observed interaction in experiment 2 of reference period length and retention interval on intrusion rates and the observed effect of reporting instructions are consistent with this view.

An additional analysis of the data from experiment 1 provides further support for the hypothesis that as retention intervals are lengthened, subjects rely increasingly on generic knowledge to report their intake for a designated period. If, as retention intervals are lengthened, reports degenerate to descriptions of generic dietary knowledge, then, if reports were scored against a list of items that constituted the salient components of generic dietary knowledge, the reports of subjects tested immediately would contain more unique items—items that are not in that list of elements of generic knowledge—than would the reports of subjects who are tested 3 weeks after the end of the diarykeeping period.

To test this prediction requires some estimate of the contents of generic dietary knowledge. Because the participants in experiment 1 were university undergraduates, many of whom lived in university residence halls and participated in a campus meal plan, it was assumed that the generic dietary knowledge of the subjects was sufficiently similar that a common estimate could be used for all of them. That estimate was constructed as follows: Each of 23 undergraduate participants in an unrelated psychological experiment was asked to list the items that a typical person of his or her age and gender would eat in a typical week. From these lists, a list was created that contained items listed by at least 10 of the 23 subjects; these items are considered stereotypic of the diets of these individuals. The 1-week dietary report of each experiment 1 subject was scored against this list to find the proportion of reported items not on the stereotype list—this was called the proportion of unique items (that is, the proportion of the reported items unique to the subject). For a subject who reported only items from the list of generic items, the proportion of unique items would have been 0; for a subject who reported no items from the list of generic items, the proportion of unique items would have been 1.

Table J shows the mean proportion of unique items for subjects in the four conditions defined by diary type and retention interval. Consistent with the prediction, the

Table J. Mean proportion of unique items reported by subjects in experiment 1, by retention interval and diary type

Diary type	Retention interval: Immediate test	Delayed test	Mean
Food	0.38	0.28	0.33
Nonfood	0.41	0.32	0.37
Mean	0.39	0.30	0.35

NOTE: Higher values indicate a greater proportion of listed items not on the stereotype list.

proportion of unique items in the subjects' reports was significantly higher for subjects who reported immediately than for subjects who received a delayed test, $F(1,89) = 14.43, p < 0.001$. The dietary reports of subjects who were tested immediately following the reference period corresponded more poorly to the generic list than did the reports of subjects who were tested after a 3-week delay (14).

Although people have access to and report some specific memories about their dietary experiences, these are incomplete, and dietary reports degenerate rapidly to reports of general knowledge about dietary intake.

Frequency judgments

To characterize completely the nutritional intake of individuals, it is important to know not only what foods they eat but also how often they eat those foods: Experiments 4–6 investigated several properties of frequency reports.

The most popular type of instrument for collecting dietary information in large-scale epidemiological studies is the food frequency questionnaire. Respondents to such questionnaires are asked to indicate either the rate at which or how many times they ate each of several items during some period—for example, 1 month or 1 year. Is it reasonable to ask people questions of this sort? How accurate are the responses? What is the nature of the representation in memory on which those responses are based?

Several distinct lines of research in experimental psychology converge to show that people are quite sensitive to variations among items in frequency of occurrence (26,32,33). Numerous experiments show that subjects can report experienced frequency quite accurately when they are tested shortly after experiencing the target set of events. Thus, insofar as dietary experiences during some period can be construed as a list of stimulus items, subjects might be expected to report, with reasonable accuracy, how many times they ate various foods during that period. However, the conditions under which subjects experience the "list" of food items differ in several significant ways from those of typical laboratory experiments on memory for frequency: "List items" are experienced over periods that range from 1 month to 1 year, the number of list items exceeds by far the number found in typical laboratory experiments, and the range of frequencies with which list items occur is much larger. The experiments described in this section examined acquisition, retention, and reporting of frequency information under conditions that approximate more closely those under which food frequency information is acquired and remembered than do the conditions of most laboratory experiments on memory for frequency.

Two measurement issues require some prefatory remarks. The first concerns the measurement of memory for frequency. Several measures of memory for frequency are used in this section. One of these is the correlation between the actual frequency of occurrence of food items (as recorded in diaries) and subjects' estimates of those frequencies. The correlation coefficient measures the linear relationship between the two counts. Its advantage over other measures is that it standardizes the variabilities of the actual and estimated frequencies, but its disadvantage is that a correlation between actual and estimated frequencies could be quite high even if the estimates failed to reflect the absolute variation in actual frequencies. For example, if reported frequencies were consistently one-tenth of the actual frequencies, the correlation between these sets of counts would be 1. The absolute variation in frequency is reflected by the slope of the function estimated by regressing reported frequencies on actual counts. When feasible, this measure is reported. Various measures of memory for frequency are discussed in detail by Naveh-Benjamin and Jonides (34).

A second measurement issue that requires clarification concerns the distinction between analysis of memory for frequency within subjects and analysis of memory for frequency for items, across subjects. Most psychological experiments on memory for frequency assess performance within subjects. Such experiments ask whether, for a given subject, there is a functional relationship between the actual and the reported frequencies, and, if so, about the nature of that relationship. These questions are addressed by analyzing the relationship between the frequencies of the items presented to the subject and his or her frequency estimates. Typically, epidemiologists are not interested in the association between estimated and actual frequency for an individual over items. Instead, they are interested in the association between estimates and actual frequency for a given item over respondents. In other words, epidemiologists want to know not whether an individual who eats chicken more often than hamburger reports this, but whether an individual who eats more chicken than does a second individual reports a higher estimate for frequency of eating chicken than does the second. Such a between-subjects analysis is reported, where appropriate.

The three experiments described in this section investigated various aspects of the ability of people to make food frequency judgments. Experiments 4 and 5 investigated the accuracy of frequency reports for previously recorded dietary events and the conditions that influence the accuracy of those judgments. Experiment 4, carried out with a subset of the participants in experiment 2, examined the effects of reference period length and retention interval on frequency judgments. Subjects' frequency estimates were regressed on counts of the number of occasions on which they had recorded target items in their diaries. To prepare materials for these subjects, recorded items were collapsed into more general classes. Because subjects and investigators might have disagreed about this categorization, experiment 5 asked subjects to report how often they had eaten each of a closed set of items which were described identically both on a checklist diary form and on the frequency questionnaire. Experiment 5 also evaluated whether engaging in the dietary recall procedure prior to making frequency judgments affected those frequency judgments.

The approach of experiment 6 to the study of frequency judgments was somewhat different. Experiment 6 investigated the sensitivity of frequency judgments to retrieval cues. Prior to making a judgment about how often they had eaten some food during a specified time period, subjects in experimental conditions were instructed to engage in thoughts that were designed either to promote or inhibit the retrieval from memory of instances of target foods.

Experiment 4: Effects of retention interval and reference period length

Numerous experiments have found a systematic relationship between the remembered frequency of occurrence of events and the actual frequency with which those events occurred. Underwood et al. (26) showed that as the time between the presentation of the target set of events and a frequency test increases, the ability of subjects to discriminate between different frequencies deteriorates, so that the slopes estimated by regressing estimated frequencies on actual frequencies decrease over retention intervals. Experiment 4 evaluated whether such a systematic effect of the length of the retention interval on the association between actual and estimated counts would be observed for judgments of food frequency.

A subset of the participants in experiment 2, after reporting their dietary intake for their diary period, were asked to indicate how often, during the reference period, they had eaten each of a set of food items. A special test set of foods was constructed for each participant, based on the contents of his or her diary.

Judgments of frequency might be based on counts of the markers that represent specific dietary memories, or they might be reports of normative frequencies that the individual supposes are typical of his or her diet (23,35). When subjects report frequencies for some reference period after a delay, estimation may involve counting markers that are decayed sufficiently that they may plausibly be records of events that occurred during the reference period. Such markers would likely be less discriminable from markers established during adjacent periods than would the stronger markers established during an immediately past reference period. Given any such process or mechanism, one would expect, for reasons similar to those discussed in experiment 2, that the performance of subjects who report about 2-week reference periods would deteriorate as the lengths of their retention intervals increase. One would further expect that among subjects who report about 4-week reference periods, the performance of those tested immediately would be best, and the performance of those tested after longer retention intervals would be worse but roughly constant over those retention intervals.

Method

Subjects—Of the participants in experiment 2, 128 served in this experiment.

Materials—For each subject, the frequency of occurrence of each item in the diary was counted. If several diary entries could be construed as referring to the same food item or to closely related food items, they were combined into a single category and their counted frequencies were summed (for example, two occurrences of "blueberry pie" and one occurrence of "apple pie" would have been counted as three occurrences of "pie"). The frequency questionnaire for each subject was constructed to include items with a broad range of frequencies of recorded occurences. The recorded items were sorted by frequency and up to 25 items were selected, subject to the constraints that the 5 most frequent items be tested and that approximately equal numbers of items be sampled from each of four quartiles of the remainder of the distribution. The selected items were randomized, and a questionnaire with those items was given to the subject.

Design and procedure—The design paralleled that of experiment 2. Subjects had been assigned to keep a food diary for 2 weeks or 4 weeks, and returned for memory tests at the end of the diarykeeping period or 2, 4, or 6 weeks later. After they had completed the dietary reporting task described in experiment 2, subjects were asked to indicate how many times during the diarykeeping period they had eaten each of the items that appeared on their questionnaires.

Results and discussion

Because the set of food items differed for each subject, only within-subjects analyses were conducted on these data. The question addressed by this experiment was whether, as retention interval was lengthened, there was deterioration in the ability of subjects to discriminate, in their frequency estimates, among items that occurred with different frequencies.

Index of performance—Subjects in the 4-week reference period ate their most frequently eaten items substantially more often than did those in the 2-week reference period condition. In an attempt to equate the variances of recorded and reported frequencies over the experimental conditions, both sets of counts were transformed logarithmically, and the use of the slopes of regressions of transformed estimates on transformed recorded frequencies was considered as a measure of memory. Even after applying the logarithmic transformation, significant differences remained among the experimental conditions in the variances of both estimates and counts of recorded occurrences. However, the ratio of the standard deviations of these counts depended neither on the reference period length nor on the retention interval: The slope found by regressing log(estimates) on log(recorded occurrences) normalized by the ratio of the standard deviations of these variables is the Pearson correlation. The comparison of performance in different conditions was carried out on z-transformed correlations.

Effects of retention interval and reference period length—Table K shows the mean correlation, over subjects, for each of the conditions defined by reference period length and retention interval. The means are all positive and reasonably high, indicating that the relationships between the logarithms of the estimates and the counts of recorded occurrences are reasonably linear. As predicted, for subjects who reported about frequencies of intake during a 2-week period, the correlation decreased as the length of the retention interval increased. For subjects who reported about 4-week periods, however, the correlation was highest for those who reported immediately after the end of the reference period and was lower, but roughly constant, for subjects who reported after longer retention intervals. A contrast on the eight means that tested this pattern was significant ($F(1,120) = 28.16$, $p < 0.0001$); the residual was not significant ($F(1,120) = 1.04$, $p > 0.25$).

This pattern is consistent with the results of experiment 2. Consider first the performance of subjects whose reference period was 4 weeks. When the test was administered immediately after the end of the reference period, available memories about the specific dietary experiences of the reference period evidently enhanced the accuracy of frequency estimates, but when the test was delayed, performance did not depend on the length of the retention interval. Dietary intake may be sufficiently similar from one 4-week period to another that estimates for any 4-week period, based on general knowledge of one's own diet, are reasonably accurate. Performance of subjects whose reference period was 2 weeks decreased systematically as the retention interval was lengthened. Such periods are sufficiently short that specific experiences of the target 2-week period must be retrieved for responses to be reasonably accurate.

Among subjects who were tested immediately after the ends of their reference periods, the mean correlation of subjects who reported about 2-week periods was higher than that of subjects who reported about 4-week periods. Available specific information certainly contributes to accuracy. However, after that information is lost, generic memory appears to be more effective for estimating frequencies for relatively longer, hence more representative, periods than arbitrary shorter ones.

Table K. Correlations of log(frequency estimates) with log(number of recorded occurrences) for subjects in experiment 4, by reference period length and retention interval

Retention interval	*Reference period length*		
	2 weeks	*4 weeks*	*Mean*
0 week	0.86	0.84	0.85
2 weeks	0.79	0.76	0.77
4 weeks	0.70	0.77	0.74
6 weeks	0.68	0.77	0.73
Mean	0.75	0.78	0.77

NOTE: Number of subjects per condition was as follows (from left to right and top to bottom): 11, 9, 17, 17, 19, 17, 18, and 20.

Experiment 5: Effects of recall using closed-ended diaries

Experiment 4 showed that people's food frequency estimates depend, in interpretable ways, on the lengths of the reference period and the retention interval. Consistent with previous experimental research on memory for frequency, the accuracy of food frequency estimates deteriorated as the length of the retention interval increased, and the decline was more pronounced for judgments about a 2-week period than for judgments about a 4-week period (13,26).

Experiment 4 might be criticized on several methodological grounds. First, the standard against which each subject's frequency estimates were scored was compiled by an experimenter with whom the subject might have disagreed about the classification of recorded items. Discrepancies between the subject's estimates and the counts from the diaries may have been due, at least in part, to such disagreements. Second, all of the subjects in experiment 4 completed the frequency questionnaires after attempting to recall their dietary intake for their reference period (see experiment 2). Engaging in the recall procedure may have influenced the frequency judgments systematically. Third, correlation coefficients are not the most desirable index of memory for frequency. A preferable measure of relative frequency is the slope of the line estimated by regressing estimates on counts of recorded occurrences. The unequal variances over conditions in recorded occurrences and estimates precluded the analysis of slopes in experiment 4. Fourth, because each subject in experiment 4 estimated the frequency of a unique set of items, the data could not be subjected to a between-subjects analysis for individual items.

Experiment 5 addressed all of these problems. All subjects kept a checklist diary about a closed set of dietary items for 1 month and then returned for a frequency test. Because the items on the diary form and the items on the questionnaire were the same, the compilation of the standard against which estimates were scored did not depend on the intervening judgment of another person. To assess whether a prior recall experience influences frequency estimates, half of the subjects engaged in the recall procedure prior to making frequency judgments. Because all subjects kept diaries for the same amount of time, the variances of the record counts and the estimates were approximately equal for the subjects in different conditions, so slopes could be analyzed. In addition, because judgments were about a closed set of items, a between-subjects analysis of the estimates could be conducted.

Method

Subjects—Thirty-one subjects participated. They were recruited through newspaper and television advertisements without regard to their demographic characteristics.

Design and procedure—Each subject was given a supply of postcards, on each of which a list of 58 food items or groups of items had been printed. Subjects were asked to use a new postcard each day to record each occasion on which they ate any of the food items on the list. Each day, the previous day's postcard was to be mailed back to the investigators.

Each subject was scheduled to return to the laboratory for a followup interview 29 days after the original visit. Prior to the second laboratory session, half of the subjects were assigned randomly to the recall condition. When these subjects arrived for their followup sessions, they were asked to report each item that they had eaten or drunk during the preceding 4 weeks. All subjects completed a questionnaire that asked them to indicate how often during the diarykeeping period they had eaten each of the items listed on the postcards. The 58 items or groups of items appeared in a different random order on each questionnaire.

Results and discussion

Effect of recall—For each subject, estimated food frequencies were regressed on the sum of recorded occurrences for each item, and the slope was obtained. Because all subjects had kept diaries for the same period of time, no transformation was applied to the counts. Whether subjects attempted to recall their intake during the diarykeeping period prior to making the frequency judgments did not affect the frequency estimates. The mean slopes of the regression lines for the recall and no-recall conditions were 1.01 and 0.90, respectively; these did not differ significantly ($t(29) = 1.22, p > 0.25$).

Within-subjects analysis—On average, the relationship between estimated and recorded frequencies was fit well by a linear function, and estimates increased as a function of recorded occurrences by about one occasion per recorded occasion. Over all subjects, the mean slope of the regression functions was 0.96 and the mean correlation was 0.86, which is quite similar to the value observed in experiment 4. These data suggest that relative frequency estimation is excellent.

Reasonably accurate estimates of frequency at extremes of recorded frequency might have inflated these estimates. Subjects tended to indicate correctly the items that they had not eaten at all during the reference period; this, together with high estimates for a few frequently consumed items, could improve measured performance. Therefore, additional regression analyses were conducted in which the range of recorded frequencies that was analyzed was iteratively narrowed. When attention was restricted to items with recorded frequencies of 1 through 10, the mean correlation was 0.62 and the mean slope was 1.01. Although the average fit of these functions was worse than was the fit over the entire set of data for each subject, the slopes remained stable at around 1. This suggests that the frequency estimates of individual subjects are reasonably accurate. Individuals give larger estimates for the items that they have eaten more often than they do for items that they have eaten less often, and their estimates reflect a sensitivity to the range of frequencies with which they ate those items.

Between-subjects analysis—Epidemiologists tend to be more concerned with whether the estimates of people who ate a particular item with different frequencies reflect, at least ordinally, those different frequencies. For example, if one individual eats potatoes more frequently than does a second, will the former's frequency estimate be greater than that of the latter? If frequency judgments were perfectly accurate, then, of course, both within-subjects analyses (over items), reported in the last section, and between-subjects analyses (for each item), reported in this section, would yield the same results.[3]

For each item, the estimates of the 31 subjects were regressed on their counts. Over items, the average correlation was 0.43 and the average slope of the regression functions was 0.56. These between-subjects analyses present a bleaker view of performance than do the within-subjects analyses. The discrepancy between the within-subjects analysis and the between-subjects analysis indicates that although the judgments of individuals tend to be internally consistent, people are not sufficiently calibrated with each other for their frequency estimates to adequately serve the purposes for which epidemiologists collect them. The development of techniques to standardize respondents with each other could improve this state of affairs.

Experiment 6: Question-induced biasing of frequency judgments

Experiments 4 and 5 indicate that respondents can provide orderly numerical estimates of food frequency. These experiments also demonstrate that such judgments are sufficiently inaccurate that they may not be adequate for their intended purposes. Experiment 4 showed that the quality of estimates for a specified period deteriorates over time. The results were generally consistent with the informal model proposed in the introduction, but they revealed little else about the processes that subserve frequency judgments.

---

[3]If a subject were perfectly accurate in this task, regression of estimated frequencies on actual frequencies would yield a correlation of 1, a slope of 1, and an estimated-frequency intercept of 0.

If dietary experiences are represented in a memory network such that the different episodes of eating any food item are interconnected, then a search of that network for representations of instances of eating an item should lead to the retrieval of relevant instances. A careful search of the memory representation for instances of eating a food should result in higher frequency estimates than judgments made without having engaged in such a search. If subjects fail to engage in a thorough search, perhaps by engaging in some mental activity that competes with a search, frequency estimates should be lower than when such a search is completed.

In experiment 6, each subject was asked to estimate how many times he or she had eaten just one food item during a specified reference period. Prior to answering the frequency question, some subjects were instructed to engage in mental activity that was hypothesized to promote or prevent a search of memory for instances relevant to the frequency question.

Table L. Geometric means of frequency estimates by subjects in experiment 6, by cognitive context, food item, and reference period length

Food item and reference period	Cognitive context: Control	Recent occasion	All occasions
Apples		Number	
1 month	5.5	4.8	6.0
1 year	33.9	27.5	28.8
Chicken			
1 month	14.8	8.1	14.5
1 year	70.8	57.5	61.7
Pizza			
1 month	9.8	7.1	11.7
1 year	61.7	52.5	67.6

NOTE: Number of observations per condition was as follows (reading from left to right and top to bottom): 24, 19, 15; 21, 20, 21; 22, 22, 23; 21, 19, 19; 23, 23, 20; and 22, 23, 24.

Method

Subjects—Responses were collected from 417 introductory psychology students who received course credit for participating in a session in which they completed a variety of research instruments for several investigators.

Design and procedure—Six frequency questions were generated by crossing three different food items with two reference periods (see table L). Each subject answered one such question. Certain subjects were instructed to engage in a specific mental activity prior to answering the frequency question. One-third of the subjects were asked to think and make notes about all the occasions on which they ate their target food item; another third of the subjects were asked to think and make notes about the most recent occasion on which they ate the target food. The all-occasions instruction was intended to promote retrieval from memory of instances relevant to the frequency question; the recent-occasion instruction was expected to prevent such retrieval by exaggerating the cognitive salience of that occasion. The remaining third of the subjects were given no special instructions.

Results and discussion

Because the three food items were eaten with vastly different frequencies and because people reported having eaten the foods with different frequencies during the two reference periods, subjects' responses were logarithmically transformed. The geometric mean responses for three food items in the two reference periods are shown in table L.

For each food item for both reference periods, the mean frequency estimate by subjects who had been given all-occasions instructions exceeded that of subjects who had been given recent-occasion instructions. The means for these instructional conditions were 22.1 and 15.4, respectively; these differed significantly, $F(1,401) = 3.80$, $p < 0.052$.

These results show that frequency estimates are sensitive to the cognitive activity in which individuals engage prior to making the estimate. Moreover, they support the notion that memories of the various episodes of eating a particular food item are linked sufficiently that, with retrieval effort, episodes that are not immediately accessible can be retrieved. In this study, responses were not validated against any sort of external record, so the accuracy of the responses is indeterminate. If additional research were to show that subjects in the all-occasions condition responded more accurately than did those in the recent-occasion condition, it would suggest that such a manipulation of cognitive activity could be used to improve the accuracy of survey responses. (For a fuller discussion of this experiment, including more detailed analyses of subsets of the data, see Smith, Jobe, and Mingay (36).)

Portion size estimates

To quantify the dietary intake of individuals, the third type of information about dietary intake that is required—after the identity of the foods that are eaten and the frequency with which they are eaten—is portion size.

Both food frequency questionnaires and dietary recall procedures sometimes involve asking respondents for information about their typical portion sizes. Obviously, an individual's amount of intake depends on both frequency and portion size: A person who eats ½ cup of spinach daily eats the same amount of spinach as a person who eats 1 cup every other day. If these individuals reported accurately the frequency with which they eat spinach, an analysis that took only frequency information into account would characterize the former individual as eating twice as much spinach as the latter. If people can report portion size information accurately, estimates of intake will clearly be improved.

Two methods are used to collect most data about portion sizes. Commonly, when dietary data are collected during an interview by a nutritionist, respondents are asked to indicate which of a set of food models most closely resembles their portion of each item. On food frequency questionnaires, or when data collection time is scarce, respondents are asked to describe their typical portion of each food item as "large," "medium," or "small," given some quantitative definition of "medium." For example, a respondent might be asked whether his or her typical serving of coffee is "large," "medium," or "small," given that "medium" is a 6-ounce cup. Respondents are clearly expected to know what quantities are signified by these definitions. It is not unusual for food frequency questionnaires to provide different standards for seemingly similar foods. For example, on the Cancer Supplement Questionnaire to the 1987 National Health Interview Survey (37), a medium serving of beans was defined as ½ cup, whereas a medium serving of carrots was defined as ¾ cup. The experiments in this section were designed to investigate what people know about portion size and some of the determinants of the statements that they make about portion sizes.

The two experiments described in this section investigated different aspects of portion size reports. The first (experiment 7) was concerned with the types of memories that are presumably required to accurately report portion sizes using food models: Subjects were asked to provide a numerical description of displayed food portions either relative to portions that had just been presented or relative to portions that had been presented sometime earlier. The dependent measure of interest was the variability of the ratios of the responses to the displayed and reference portions. The second (experiment 8) investigated the meaningfulness to respondents of the definitions of portion sizes that are presented on food frequency questionnaires. A medium portion was defined differently for different subjects, and distributions of portion size responses were analyzed to assess whether subjects were sensitive to this manipulation.

Experiment 7: Estimation of sizes of food portions

Subjects in this experiment were shown photographed portions of 11 different food items and were to assign numbers to describe the sizes of the portions. Each displayed serving was to be judged relative to the previously displayed portion of that food. Subjects were instructed to respond so that the ratio of the responses to successive exemplars of each food item would equal the ratio of the two portion sizes. For example, if a displayed serving of scrambled eggs appeared twice as large as the previously seen serving of scrambled eggs, the numerical response was to be twice as large.

Two experimental conditions differed in the location of the reference stimulus (the previously displayed serving of the food item displayed on a particular trial) in the stimulus sequence. In the blocked condition, sequences of slides of a single type of food item were presented, so that each judgment of subjects in the blocked condition was to be relative to the stimulus and response of the preceding trial. In the mixed condition, the same slides were presented, but the slides of the different types of food items were intermixed. Therefore, each judgment of subjects in the mixed condition was to be relative to a stimulus and response on a trial that had occurred, on average, 11 trials earlier.

Because subjects were instructed to respond according to the ratio of the size of a displayed stimulus to the size of a specific reference stimulus, the ratios of responses to these two stimuli can be calculated, and their consistency can be evaluated. If intervening trials interfere with memory for size, the response ratios should be more variable for subjects in the mixed condition than for subjects in the blocked condition, because only subjects in the mixed condition experienced intervening trials between the displayed and reference stimuli.

Method

Subjects—Twenty-two university undergraduates, participating to partially fulfill a requirement of their introductory psychology course, served in the experiment. Eight subjects were assigned to the blocked condition, and the remaining 14 were assigned to the mixed condition. Subjects participated in small groups; all subjects in any group served in the same experimental condition.

Stimuli—The stimuli were color slides of servings of food. Eleven different-sized samples of each of 11 different food items were used, for a total of 121 different stimuli. For each food item, portion sizes increased in equal arithmetic steps from the smallest to the largest. For example, for orange juice, the smallest portion was 2 ounces, and the serving sizes increased in 1-ounce steps to the largest portion, which was 12 ounces. For each food item, the median portion was designated as the standard (S_0).

Design—Each of the 121 stimuli was presented four times, so each subject responded to 484 stimulus presentations. In the stimulus sequence for the blocked condition, the 44 exemplars of each food item were arranged consecutively, so that each stimulus followed another stimulus of its own type. The appropriate standard was shown before the first trial of each food-item trial block. The stimulus sequence for the mixed condition was constructed by interleaving the 11 item sequences for the blocked condition, subject to the constraint that one exemplar of each food item be included in each group of 11 slides. All 11 standards were shown prior to the first block of trials.

In each condition, the 11 blocks of 44 slides were treated as three sets of three blocks and one set of two blocks, and these four units were presented in different orders to different groups of subjects.

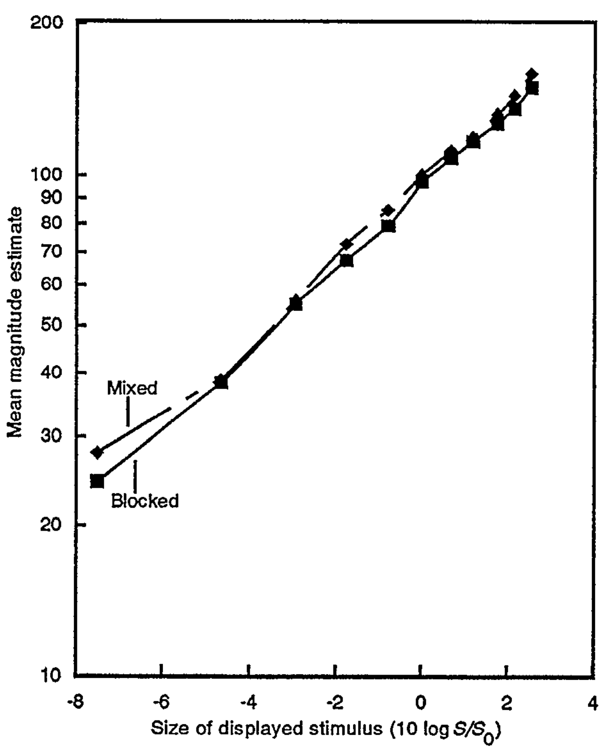
Procedure—The subjects in each experimental condition were told that they would view slides that showed portions of various food items and that the portions would vary in size. The subjects were instructed to respond, on each trial, with a number that reflected the ratio of the size of the stimulus to the size of the preceding stimulus of the same item. The role of the standards was explained. Subjects were told that their first judgments for each food item should be relative to the standard portion for that item and that the value 100 was to be assigned to the standards.

Results and discussion

Data editing and aggregation—Each subject's data were edited to exclude highly deviant responses. First, the standard deviations of the logarithms of responses were calculated for each size level of each food item, and these standard deviations were averaged. Then, at each size level of each food item, if the most deviant response differed from the mean of the remaining responses by more than four average standard deviations, it was removed from the data. Approximately 2 percent of the responses were removed in this way. The mean numbers of responses removed for subjects in the blocked and mixed conditions (11.63 and 10.64, respectively) did not differ significantly, $t(20) < 1$.

Judgments of portion size—Figure 1 shows, for each stimulus-sequence condition, geometric mean responses as a function of the size of the displayed portions (expressed as $\log(S/S_0)$). The response functions from the two experimental conditions are essentially superimposed. In each condition, numerical responses increased linearly and monotonically with the sizes of the displayed portions; the linear component of the stimulus size effect was the only significant effect in these data, $F(1,20) = 1683.04$, $p < 0.0001$.

Response ratios—To calculate the coefficient of variation of response ratios, the trials were first sorted by the ratio of the size of the displayed stimulus to the size of the reference stimulus for that trial. (In the blocked condition, the reference stimulus was always the stimulus that had been shown on the preceding trial; in the mixed condition, the reference stimulus was the previously presented exemplar of the displayed food item.) For each trial, the



NOTE: Geometric mean responses are shown as a function of the logarithm of the ratio of the size of the displayed stimulus to the size of the standard stimulus. Data are averages across food items and across subjects in the blocked and mixed conditions.

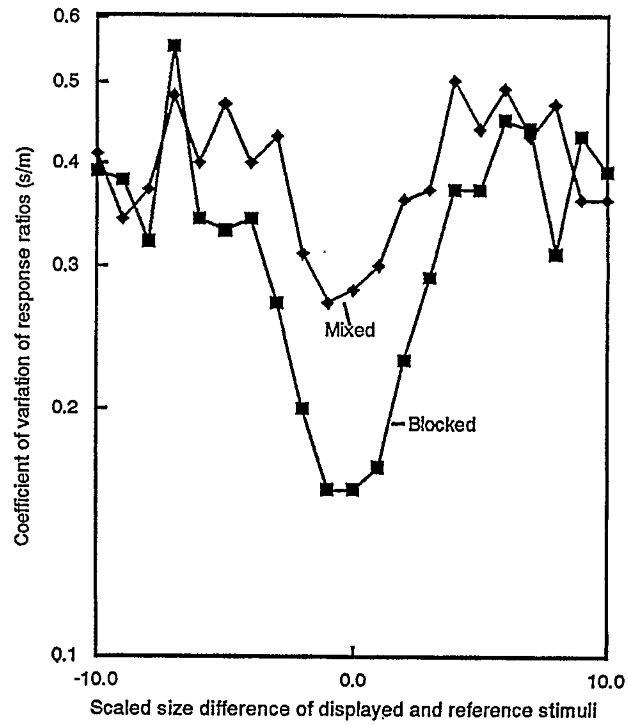
Figure 1. Mean magnitude estimate, by size of displayed stimulus

ratio of responses to the displayed and reference stimuli was calculated. Then, for each class of trials defined by a ratio of stimulus sizes, the mean and standard deviation of the response ratios were calculated; the ratio of the standard deviation to the mean is the coefficient of variation of response ratios.[4] For example, for every trial on which the stimulus was twice as large as that trial's reference stimulus, the ratio of the response on that trial to the response to the reference stimulus was found. The standard deviation of these ratios, divided by their mean, is the coefficient of variation for this value of stimulus size ratio.

Figure 2 shows the coefficient of variation of the response ratios of the subjects in each stimulus-sequence condition as a function of the difference in sizes between the displayed and reference stimuli (expressed as $\log(S/S_{ref})$, where the subscript denotes the reference stimulus). For example, zero on the abscissa indicates that the sizes of the target and reference stimuli were the same, positive values indicate that the displayed stimulus was larger than the reference stimulus, and negative values indicate that the displayed stimulus was smaller than the reference stimulus. For each subject, trials for different food items were pooled to calculate the coefficient of variation function, and these were averaged across subjects in each condition.

For subjects in the blocked condition, the shape of the coefficient of variation function is a pronounced V, with a minimum value for stimuli that were of the same size as the reference stimulus. For subjects in the mixed condition, the minimum coefficient of variation of response ratios was also at stimulus-size ratios of around 1, but the dip is far less pronounced than for the blocked condition. The difference between the conditions in the depth of the V was confirmed by an analysis of variance that showed that the interaction of sequential condition (blocked versus mixed) and a quadratic contrast on log stimulus ratios was significant, $F(1,20) = 5.10, p < 0.05$.

When the size difference between the displayed stimulus and the reference stimulus was small, the variability of the response ratios of subjects who had to remember



NOTE: The coefficient of variation of the ratio of responses to the displayed and reference stimuli is shown as a function of the logarithm of the ratio of the sizes of the displayed and reference stimuli. Data are averages across food items and across subjects in the blocked and mixed conditions.

Figure 2. Coefficient of variation of response ratios, by scaled size difference of displayed and reference stimuli

back only one trial was considerably less than was the variability of the response ratios of subjects who had to remember back further. When the size difference of successive exemplars of a particular food was large, the normalized variability of response ratios did not depend on whether the reference stimulus for the trial had just been presented or had been presented some number of trials earlier. For extreme size ratios of target to reference stimuli, subjects may be unable to use the ratio response rule and may judge the size of the target stimulus absolutely (38,39).

Although estimates of size averaged over many observations did not depend on the sequential arrangement of the judged stimuli (see figure 1), memory-dependent relative judgments of size are least variable when the reference stimulus has just been presented and is of approximately the correct size. This suggests that the best estimates of size, using food models, would be obtained immediately after the judged food portion has been viewed, and this does not occur in practice. Most interviews take place hours or days after the portions about which the respondents report have been viewed. The results of this experiment suggest that food-model-based portion size estimates for individuals are likely to be inaccurate, although averages of estimates for groups of individuals may be quite reasonable.

[4]The coefficient of variation is a normalized measure of variability, the standard deviation divided by the mean, that is studied when the variability of a measure is proportional to its mean. Consider, for example, the weights, in kilograms, of elephants and cats: The standard deviation of the weights of a sample of elephants will be larger than that for a sample of cats simply because the weight of an average elephant is thousands of kilograms, whereas the weight of the average cat is around 5 kilograms. The weights of elephants can deviate from their mean much more than the weights of cats can deviate from theirs. For the response ratios described in the main text, when a large stimulus is judged relative to a small one, the ratio will be large, and the variability of several replications of this judgment will be large because, in absolute values, there is greater potential for variation than when values are small. When a small stimulus is judged relative to a large one, the ratios are small and their variability is small. The coefficient of variation makes the variabilities of these ratios comparable.

Experiment 8: Comprehension and utilization of defined portion sizes

Many dietary questionnaires ask respondents to quantify their typical portion sizes for various foods by indicating whether their portions are "small," "medium," or "large," with "medium" given a particular quantitative definition. For example, respondents might be asked about their typical portions of coffee, with "medium" defined as 6 ounces, or about their typical portions of spinach, with "medium" defined as ½ cup. This method of portion size assessment tends to be used on self-completed questionnaires and on fast-paced general health questionnaires that might be administered as part of a household survey.

To respond accurately to such questions, a respondent must have precise internal representations of the sizes of his or her typical portions (which, of course, might be quite variable) and of the quantity that is defined as "medium." A respondent who has specific quantitative knowledge of his or her portion sizes may respond accurately to such questions. Lacking such knowledge, or lacking knowledge of the sizes of standard portions, this task might be quite difficult. Nevertheless, respondents will likely respond to such questions anyway.

This experiment explored the ability of people to provide meaningful reports about portion sizes with this assessment method. The specific question addressed was whether people are sensitive to the defined portion sizes. Subjects were asked to indicate whether their typical portion sizes for selected foods were "small," "medium," or "large." Over subjects, the quantitative definition of the response "medium" was varied. For some subjects, "medium" was defined as a relatively small portion; for others, as a relatively large portion; and for the remainder, as an intermediate-sized portion. If subjects have stable knowledge of their portion sizes and stable representations of the quantities in terms of which "medium" is defined, then the distribution of responses over the three alternatives should depend systematically on the definitions. Subjects for whom "medium" was defined as relatively small should tend to respond "large" more often than subjects who received larger definitions of medium; subjects for whom "medium" was defined as relatively large should tend to respond "small" more often than subjects who received smaller definitions of medium.

Method

Each of 414 introductory psychology students completed a food frequency and portion size questionnaire; each subject responded to four items sampled from among eight used in the experiment. For each food item, medium was defined, over questionnaires, by three different quantities; for every item, the largest definition of medium was at least twice that of the smallest. Table M lists the food items and shows the ratio of the largest to the smallest definition of medium for each item.

For each of four questions, the subject was asked to make a frequency judgment and to indicate whether his or her typical portion was "small," "medium," or "large." The quantitative definition of "medium" appeared in parentheses adjacent to that alternative. Among the four items presented to each subject was at least one with a low definition of "medium," one with an intermediate definition of "medium," and one with a high definition of "medium." Over 200 responses were collected for each food item, with approximately 70 subjects receiving each definition of "medium."

Results and discussion

Table M shows the distribution of responses over the three response alternatives for each food item. To evaluate whether there was an orderly shift over the definitions of "medium" in the distributions of responses, ridit analysis was used (40). This technique permits comparison among different samples of distributions of an ordered set of responses. In this application, each sample consisted of the subjects who read a particular definition of "medium" for a given item. For each condition of each food item, a mean ridit was calculated to summarize the distribution of responses in a single condition relative to the distribution of responses pooled over conditions. The condition mean ridit is an estimate of the probability that a random response from that condition is larger than a response sampled at random from the distribution of responses

Table M. Response distributions of reported portion size in experiment 8, by food item and definition of medium

Food (large: small)	*Definition of medium*	*Number of responses* Small	Medium	Large	*Mean ridit*	χ^2
Sliced cheese . . . (2:1)	Low	24	35	13	0.5333	2.01
	Middle	23	36	2	0.4694	
	High	35	28	16	0.4933	
Cream cheese . . . (4:1)	Low	20	35	20	0.5238	3.32
	Middle	24	26	23	0.5171	
	High	18	28	6	0.4416	
French fries. . . . . (2:1)	Low	13	37	15	0.5046	3.46
	Middle	8	47	16	0.5369	
	High	18	38	12	0.4571	
Fruit juice . . . . . . (2:1)	Low	5	28	33	0.5690	[1]10.56
	Middle	3	39	23	0.5131	
	High	14	42	20	0.4289	
Green salad . . . . (2:1)	Low	10	39	20	0.4824	0.69
	Middle	7	41	24	0.5180	
	High	8	40	21	0.4988	
Ice cream . . . . . . (2:1)	Low	18	31	17	0.4995	[1]6.47
	Middle	10	42	19	0.5551	
	High	19	41	8	0.4429	
Pasta. . . . . (2:1)	Low	9	33	27	0.5209	1.40
	Middle	7	40	24	0.5079	
	High	7	45	18	0.4714	
Salty snacks . . . . (4:1)	Low	6	39	27	0.5413	[2]4.69
	Middle	9	37	23	0.5077	
	High	15	31	17	0.4444	

[1] $p < 0.05$.
[2] $0.05 < p < 0.10$.

pooled across conditions. Table M shows the condition mean ridits for each food item.

If people are sensitive to the presented definitions of portion sizes, then, assuming the average portion sizes of the subjects assigned to each condition are equal, as the definition of "medium" is increased, more "small" responses should be observed and condition mean ridits should decrease. Table M shows that for only four of the items (cream cheese, fruit juice, pasta, and salty snacks) did condition mean ridits decrease in this fashion, and for only juice were the differences among the condition mean ridits significant. Although the condition mean ridits for ice cream differed overall, pairwise comparisons showed that the mean ridits of the low- and high-definition conditions did not differ significantly.

These results suggest that, in general, either people lack stable representations of the quantities of food that they eat, or the defined portion sizes are not meaningful to them, or both. For half of the items used in this experiment, response distributions over the conditions did not conform to the normatively expected pattern—that as the defined size of a medium portion is increased, the distribution of responses shifts toward "small." For the other four items, condition mean ridits decreased as the defined size of medium increased, but for three of these, the shift in the response distributions was insufficient to yield a statistically significant result. Inspection of table M shows that this is due, at least in part, to a strong tendency for these subjects to respond "medium," regardless of the quantitative definition ascribed to "medium."

The one food item for which condition mean ridits decreased systematically and significantly was fruit juice. Ounces, particularly in the numbers used—4, 6, and 8—may correspond to natural cognitive units of juice. People often drink containers of juice whose size, in ounces, they know. In contrast, people do not typically eat green salad from containers whose size, in cups, they know (although this may change if the 1990's become the age of the fast-food salad).

Response alternatives on surveys in general and on dietary surveys in particular must give respondents an appropriate means to communicate their knowledge about whatever the data collector wants to know. This will likely be best achieved if the format of response alternatives corresponds to that of the respondents' knowledge. Dietary surveys ask respondents to describe food frequency and portion sizes rather than to describe average daily intake of specific micronutrients because it is assumed, probably correctly, that most individuals have no idea what their average daily intakes of those micronutrients are. Respondents may not know what a cup of green salad is, either. The meaningfulness of response options must be evaluated before they are used to collect data from respondents.

Conclusions

Overview

Dietary data are collected from individuals in order to identify and investigate relationships between nutritional intake and health status. Nutritional epidemiologists are aware that dietary reports contain some degree of response error, but, nevertheless, they use data collected with a variety of methods, including recall and frequency procedures (2,11). Any response error in survey data reduces the power of analyses of those data. Such loss of power may result in the failure to detect relationships that are present in the studied population or necessitate incurring additional expenses to collect data on larger sample sizes (41).

The research program described in this report was motivated by the supposition that an understanding of the mental representations and processes that underlie dietary reporting can improve the design of dietary survey methods. Such knowledge could inform the development of methods to obtain better data than can be collected with current methods and could improve the understanding of the data collected with current methods. Toward this end, several aspects of dietary reporting performance were studied under conditions that were manipulated to reveal the functioning of the mental representations and processes assumed to be responsible for those reports.

Three types of information are essential to quantifying dietary intake— what people eat, how often they eat what they eat, and the quantities in which they eat what they eat. Each of these aspects of dietary intake may be studied somewhat independently. The experiments described in this report examined dietary recall performance, food frequency estimation, and judgments of portion size.

Experiments 1–3 examined dietary recall under various acquisition, retention, and retrieval conditions. Differences in reporting performance among experimental conditions were used to make inferences about the memorial representation of dietary information. Unconstrained reports of dietary intake for reference periods of 1 week to 1 month deviated markedly from records of intake that were maintained during those periods: Experimental subjects both failed to report items that they recorded, and reported substantial numbers of items that they did not record. As the amount of time between the end of the reference period and the recall test was increased, reporting accuracy decreased, with performance appearing to level off after a retention interval of 4 weeks.

Taken together, the results of experiments 1–3 encourage the notion that reports of dietary intake are, to a considerable extent, reports of generic knowledge of diet rather than reports of memories of specific experiences (28). Generic knowledge of diet is hypothesized to be organized information about the routine elements of an individual's diet. Recall of dietary intake for extended periods appears to be a task that people carry out quite inefficiently. It is possible that, with appropriate cues, the accuracy of dietary recall might be improved, although this would be at a considerable cost in time. On the other hand, the information provided by subjects, although inaccurate relative to the intake recorded for a specific period, may characterize what respondents typically eat.

Experiments 4–6 addressed the determinants of accuracy of people's reports about the frequency with which they eat various foods. In general, subjects provided reasonable estimates of the relative frequency with which they ate various foods. The judgments of any subject were such that estimates for more frequently eaten items were higher than estimates for less frequently eaten items. In fact, the data of experiment 5 showed that, on average, estimates increased by about one occurrence per recorded occurrence, indicating that subjects were highly sensitive to variations in experienced frequency. However, of primary interest in epidemiological research are between-subjects analyses of frequency estimates—that is, analyses of whether the responses of different individuals who eat some food with different frequencies are at least ordinally consistent with the true frequencies. Despite the excellent within-subjects frequency estimates observed in experiment 5, there is sufficient variability over people in the accuracy of responses that analyses of frequency judgments for items over subjects were less indicative of accurate judgments. Nevertheless, the within-subjects consistency of the judgments encourages the idea that, if respondents could be calibrated to a common standard, between-subjects consistency might be improved.

Several other qualifications concerning frequency estimates must be made. Experiment 4 showed that the accuracy of frequency judgments deteriorates over time and that relative judgments for longer, presumably more representative, periods of time are more stable than are those for shorter periods of time. Experiment 6 showed that frequency judgments were sensitive to such external manipulations as instructions to engage in mental activity that would either promote or preclude the retrieval of information relevant to a frequency judgment.

The results of the experiments on portion size judgments suggest that people do not have high-resolution long-term memories for size information, nor do they code in memory the sizes in standard units of their typical portions of various foods. In experiment 7, in which subjects were asked to respond numerically to the sizes of displayed portions, the relative variability of response ratios was low only when the reference stimulus for the judgment had just been presented and was close in size to the target stimulus. In the absence of these two conditions, as would typically be the case in nutritional surveys, the relative variability of response ratios was high. In experiment 8, for only one of eight studied food items were response distributions sensitive to changes in the quantitative definition of a standard portion.

Implications

What do these results imply for the collection of dietary survey data for epidemiological purposes? This penultimate section is organized around several of the techniques used in nutritional epidemiology for the collection of such data.

Dietary recall

The results of experiments 1 and 2 suggest that the reports collected in dietary recall procedures for extended reference periods do not accurately characterize dietary intake during those periods. Many items are omitted, and many of the reported items are intrusions. The results of experiment 1 suggest that the data collected using 24-hour recall procedures may be reasonably accurate reports about intake on the preceding day, but, of course, one day's intake is unlikely to be representative of the diet of an individual. (Because the procedure used in experiment 1 did not permit the calculation of a daily intrusion rate, only this weak statement about the accuracy of data collected using 24-hour recall can be made.)

The results of experiments 1 and 2 suggest that taking diet histories—reports about the routine elements of an individual's diet—may be a reasonable way to collect information. If what is needed for research purposes is a set of items that are typical of the individual's diet rather than a specific list of items eaten during a specific period of time, the diet history may be adequate to collect such a list. In such a case, the respondent would be instructed to report the contents of his or her generic dietary knowledge, rather than to attempt to search memory for specific dietary memories. People may be quite able to report what they typically eat, even if they cannot report exactly what they ate during some specific period of time.

Frequency questionnaires

Food frequency questionnaires are used to collect information of two types—frequency information and categorical portion size information. The results of experiments 4 and 5 show that people report relative frequency information with reasonable accuracy, although the accuracy of their estimates—especially for short reference periods—declines as the amount of time since the end of the reference period increases. Unfortunately, the within-subjects consistency in frequency estimates observed in experiment 5 does not carry over to the analyses of items (over subjects) that are used commonly in epidemiology. This discrepancy is probably due to intersubject discrepancies in calibration for frequency judgments. It would be desirable to develop ways to align the frequency estimates of different respondents to food frequency questionnaires.

The results of experiment 6 showed that people's frequency estimates may be influenced by the thoughts in which they engage prior to making those estimates. Although it would be unusual for a respondent to a food frequency questionnaire to be asked to think about the most recent time that he or she had eaten an item prior to judging the frequency with which he or she eats that item, the results of experiment 6 raise the general concern that responses to food frequency questionnaires might be influenced by the sequential arrangement of questions. For example, a question about one food might lead subjects, on a subsequent question about some other food, to consider only a restricted range of occasions on which that food is eaten. This problem, if it is a problem, could be circumvented by randomizing questions over forms of the questionnaire.

The results of experiment 8 raise concerns about the categorical portion size information collected on questionnaires. Subjects in experiment 8 displayed a general insensitivity to variations in the defined sizes of food portions. This indicates that portion size questions of the type used in the 1987 National Health Interview Survey (37) may be meaningless to respondents.

Food models

In dietary interviews, food models are used to collect portion size information. After naming a food item that he or she has eaten, the respondent is asked to indicate which of a set of models matches the size of the consumed portion. The results of experiment 7 suggest that precise memories for portion sizes do not survive exposure to intervening items. Food models can likely be used to collect only very crude portion size information.

Final comment

Perhaps the most surprising result of the experiments on recall and frequency estimation was that, from a psychological perspective, the results were not surprising.

Despite the differences between the conditions under which individuals acquire information about their dietary experiences and those under which subjects acquire information in laboratory experiments, and despite differences in the nature of the to-be-remembered events, the results of these experiments on memory for dietary information are generally consistent with the results of relevant laboratory research on memory. By virtue of this consistency, these results strengthen those principles.

References

1. Surgeon General. The Surgeon General's report on nutrition and health. DHHS No. 88–50210. Washington: U.S. Government Printing Office. 1988.
2. Willett WC. Nutritional epidemiology. New York: Wiley. 1990.
3. Block G. A review of validations of dietary assessment methods. Am J Epidemiol 115:492–505. 1982.
4. Krall EA, Dwyer J, Coleman KA. Factors influencing accuracy of dietary recall. Nutri Res 8:829–41. 1988.
5. Coughlin SS. Recall bias in epidemiologic studies. J Clin Epidemiol 43:87–91. 1990.
6. Harlow SD, Linet MS. Agreement between questionnaire data and medical records. Am J Epidemiol 129:233–48. 1989.
7. Lippman A, Mackenzie SG. What is "recall bias" and does it exist? In: Marois M, ed. Prevention of physical and mental congenital defects: Part C: Basic and medical science, education, and future strategies. New York: Alan R. Liss, 205–9. 1985.
8. Mackenzie SG, Lippman A. An investigation of report bias in a case-control study of pregnancy outcome. Am J Epidemiol 129:65–75. 1989.
9. Raphael K. Recall bias: A proposal for assessment and control. Int J Epidemiol 16:167–70. 1987.
10. Marr JW. Individual dietary surveys: Purposes and methods. World Rev Nutr Diet 13:105–64. 1971.
11. Young CM. Dietary methodology. In: Committee on Food Consumption Patterns, National Research Council, Assessing changing food consumption patterns. Washington: National Academy Press, 89–118. 1981.
12. Hankin JH, Rhoads GG, Glober GA. A dietary method for an epidemiologic study of gastrointestinal cancer. Am J Clin Nutr 28:1055–61. 1975.
13. Willett WC, Sampson L, Stampfer MJ, et al. Reproducibility and validity of a semiquantitative food frequency questionnaire. Am J Epidemiol 122:51–65. 1985.
14. Smith AF, Jobe JB, Mingay DJ. Concerning the suitability of diarykeeping to validate recall of dietary information. Manuscript submitted for publication. 1990.
15. Bradburn NM, Rips LJ, Shevell SK. Answering autobiographical questions: The impact of memory and inference on surveys. Science 236:157–61. 1987.
16. Fienberg SE, Loftus EF, Tanur JM. Cognitive aspects of health survey methodology: An overview. Milbank Mem Fund Q 63:547–64. 1985.
17. Fienberg SE, Loftus EF, Tanur JM. Cognitive aspects of health surveys for public information and policy. Milbank Mem Fund Q 63:598–614. 1985.
18. Jabine TB, Straf ML, Tanur JM, Tourangeau R, eds. Cognitive aspects of survey methodology: Building a bridge between the disciplines. Washington: National Academy Press. 1984.
19. Jobe JB, Mingay DJ. Cognition and survey measurement: History and overview. Appl Cognit Psychol 5:175–92. 1991.
20. Tourangeau R. Cognitive science and survey methods. In: Jabine TB, Straf ML, Tanur JM, Tourangeau R, eds. Cognitive aspects of survey methodology: Building a bridge between disciplines. Washington: National Academy Press, 73–100. 1984.
21. Crowder RG. Principles of learning and memory. Hillsdale, New Jersey: Lawrence Erlbaum. 1976.
22. Hasher L, Zacks RT. Automatic processing of fundamental information: The case of frequency of occurrence. Am Psychol 39:1372–88. 1984.
23. Hintzman DL. Judgments of frequency and recognition memory in a multiple-trace memory model. Psychol Rev 95:528–51. 1988.
24. Anderson JR, Bower GH. Recognition and retrieval processes in free recall. Psychol Rev 79:97–123. 1972.
25. Cofer CN. Does conceptual organization influence the amount retained in immediate free recall? In: Kleinmuntz B, ed. Concepts and the structure of memory. New York: Wiley, 181–214. 1967.
26. Underwood BJ, Zimmerman J, Freund JS. Retention of frequency information with observations on recognition and recall. J Exp Psychol 87:149–62. 1971.
27. Rabinowitz JC, Mandler G, Patterson KE. Determinants of recognition and recall: Accessibility and generation. J Exp Psychol Gen 106:302–29. 1977.
28. Smith AF, Jobe JB, Mingay DJ. Retrieval from memory of dietary information. Appl Cognit Psychol 5:269–96. 1991.
29. Cohen BH. Recall of categorized word lists. J Exp Psychol 66:227–34. 1963.
30. Deese J. Influence of inter-item associative strength upon immediate free recall. Psychol Rep 5:305–12. 1959.
31. Graesser AC. Prose comprehension beyond the word. New York: Springer-Verlag. 1981.
32. Jonides J, Naveh-Benjamin M. Estimating frequency of occurrence. J Exp Psychol (Learn Mem Cogn) 13:230–40. 1987.
33. Lichtenstein S, Slovic P, Fischoff B, et al. Judged frequency of lethal events. J Exp Psychol (Hum Learn Mem) 4:551–78. 1978.
34. Naveh-Benjamin M, Jonides J. On the automaticity of frequency coding: Effects of competing task load, encoding strategy, intention. J Exp Psychol (Learn Mem Cogn) 12:378–86. 1986.

35. Blair E, Burton S. Cognitive processes used by survey respondents to answer behavioral frequency questions. J Consumer Res 14:280–8. 1987.
36. Smith AF, Jobe JB, Mingay DJ. Question-induced cognitive biases in reports of dietary intake by college men and women. Health Psychol 10:244–51. 1991.
37. Schoenborn CA, Marano M. Current estimates from the National Health Interview Survey. National Center for Health Statistics. Vital Health Stat 10(166). 1988.
38. Green DM, Luce RD, Smith AF. Individual magnitude estimates for various distributions of signal intensity. Percept Psychophys 27:483–8. 1980.
39. Luce RD, Baird JC, Green DM, Smith AF. Two classes of models for magnitude estimation. J Math Psychol 22:121–48. 1980.
40. Fleiss JL. The design and analysis of clinical experiments. New York: Wiley. 1986.
41. Freudenheim JL, Johnson NE, Wardrop RL. Nutrient misclassification: Bias in the odds ratio and loss of power in the Mantel test for trend. Int J Epidemiol 18:232–8. 1989.
42. Green DM, Swets JA. Signal detection theory and psychophysics. Huntington, New York: Robert E. Krieger. 1974.
43. Hilgard ER. Methods and procedure in the study of learning. In: Stevens SS, ed. Handbook of experimental psychology. New York: Wiley, 517–67. 1951.

Appendixes

Contents

Appendix tables

Appendix I
Indirect scoring of a simulated memory experiment

This appendix illustrates the argument that indirect response validity assessment methods, which transform both recorded and reported food items to mean daily nutrient intakes, might lead to the false conclusion that reports are highly accurate.

An analog to indirect assessment of response accuracy, in a psychological experiment on memory for words, might involve decomposing both the to-be-remembered and the reported words into their component letters and, for each letter, calculating over subjects the correlation between the counts of to-be-remembered and reported letters. In this example, the 26 letters, the components of words, are considered to be analogous to the nutrients contained in foods. If memory performance were evaluated with such a set of correlations, a subject's erroneous report of *Mississippi* for *stresses* would not affect the correlation calculated for the *s*'s, but would affect the correlations for the other letters. The erroneous report of an anagram of a to-be-remembered word (for example, *clean* for *lance*) would not be reflected as an error. The notion of nutritional anagrams, or near-anagrams, could easily be developed.

To evaluate the conjecture that the correlations might suggest good performance even in cases in which other measures would suggest otherwise, a set of memory experiments was simulated and scored as follows: Each simulation assumed 40 subjects, to each of whom a unique memory set of 50 items, sampled from a 596-item pool, was presented. Simulated subjects reported a specified number of responses from this pool, with specified match and intrusion rates. The match rate is the reported proportion of to-be-reported items; the intrusion rate is the proportion of reported items that were not in the memory set. (These measures are discussed in detail in appendix II.) Memory and response sets were transformed to counts of letters, and, for each of the 26 letters, counts from the reports were regressed on counts from the memory items and the correlation between the counts over subjects was calculated.

Table I shows, for values of the three performance parameters (number of items reported, match rate, and intrusion rate), the mean of the slopes and the intercepts of the 26 regression functions and how many of the 26 correlations were significant at the 5-percent level. The top line shows the results for perfect performance; the bottom line shows the results for subjects who report 25 items, all of which are intrusions. The key results are on the intermediate lines, which show that many significant correlations between the counts of components of the memory and response sets may be observed even if memory performance is quite poor. For example, the third line of the table shows that when match rates are only 0.5 and intrusion rates are 0.5, 23 significant correlations were observed.

Table I. Results of indirect scoring of simulated memory experiments

N(report)	*P*(mtc)	*P*(int)	*Slope*	*Intercept*	*Correlations*
50	1.0	0.0	1.00	0.00	26
50	0.8	0.2	0.79	4.28	26
50	0.5	0.5	0.47	9.91	23
25	0.5	0.0	0.52	0.16	26
25	0.0	1.0	−0.09	11.32	0

NOTE: *P*(mtc) is match rate; *p*(int) is intrusion rate; "Correlations" is the number of correlations (out of 26) significant at the 5-percent level.

Appendix II
Indexes of reporting performance

This appendix clarifies the relation between the two indexes of performance that are described in this report and explains why they were chosen over certain possible alternatives. This discussion is adapted from the appendix of Smith et al. (28).

The indexes

For experiments 1 and 2, subjects recorded items in diaries and, subsequently, were asked to report as many of those items as they could. Subjects reported items that had not been recorded as well as items that had been recorded. After preliminary editing of both the diaries and the report protocols, the reported items were divided into two sets according to whether they matched items in the diary, and the diary items were divided into two sets according to whether they were matched by items in the report.

Let D be the number of items in the diary, R be the number of items in the report, and m be the number of items that are matches between the diary and the report. Also, for convenience, let the number of intrusions be $i = R - m$, and let the number of omissions be $o = D - m$. The two indexes of performance are defined as

$$p(\text{match}) = m/D \qquad \text{(A1)}$$

and

$$p(\text{intrusions}) = 1 - (m/R). \qquad \text{(A2)}$$

Note that the first of these measures is a fraction of the number of items recorded, and the second is a fraction of the number of items reported. By definition, the relationship between the two indexes is

$$\frac{p(\text{match})}{1 - p(\text{intrusions})} = \frac{R}{D} \qquad \text{(A3)}$$

Restrictions on values

In principle, the only restriction on the possible values of p(intrusions) is that if p(match) is greater than 0, p(intrusions) must be less than 1, because there will be at least one match. Otherwise, for any p(match) greater than 0, p(intrusions) can range from 0, if all of the reported items are matches, to arbitrarily close to 1, if i (the number of intrusions) is sufficiently close to R for i/R to be large. Note particularly that for any D, p(match) does not depend on i, the number of intrusions. (If p(match) is equal to 0 and R is greater than or equal to 1—that is, if at least one item is reported—then p(intrusions) must be equal to 1.)

By way of illustration, let the set of items reported by an individual be labeled $x(1), x(2), \ldots x(R)$. If these are arranged as if the individual reported first his or her matches and then his or her intrusions, so that $x(1), \ldots x(m)$ are the matches and $x(m + 1), \ldots x(R)$ are the intrusions, then, because D is fixed, $p(\text{match}) = m/D$, and this value would have been determined once the subject had reported the mth item. Then, as the subject reported his or her intrusions, items $x(m + 1)$ through $x(R)$, only p(intrusions) would have changed, growing from 0 before the subject reported the $(m + 1)$st item to $(R - m)/R$ when he or she reported the Rth item. Although p(match) and p(intrusions) are related, as described by equation A3, the relationship is through both D and R.

Empirical relationship

P(match) and p(intrusions) are generally unrelated. For example, over all of the subjects of experiment 2, the correlation between these indexes was –0.16, which, although statistically significant at $p < 0.05$, is quite small. The four extreme points in the p(match) x p(intrusions) scatterplot were (0.167,0.125), (0.186,0.797), (0.673,0.132), and (0.683,0.423), which shows that at the extreme levels of p(match), the range of p(intrusions) was reasonably large. In the data of experiment 2, the correlation of these measures in the eight experimental conditions defined by combinations of reference period length and retention interval ranged from –0.37 to 0.48, with a mean of 0.02. In only one condition was the correlation statistically significant.

Choice of the indexes

These indexes are not ideal. It would be preferable to have measures corresponding to the hit and false alarm rates of signal detection theory that would permit the calculation of analogs of the sensitivity and response bias parameters of signal detection theory (42).

The counts that contribute to the calculation of these indexes are summarized in table II. Examination of that table clarifies why measures analogous to those of signal detection theory cannot be computed. Although p(match) corresponds to signal detection theory's hit rate, to

Table II. Summary of counts required to calculate performance indexes

Items recorded	*Items reported* Yes	No	Sum
Yes	*m*	*o*	*D*
No	*i*	(1)	(1)
Sum	*R*	(1)	(1)

[1]Count is neither known nor estimable.
NOTE: *m* represents matches, the count of reported items that match recorded items; *i* represents intrusions, the count of reported items which were not recorded; *o* represents omissions, the count of recorded items which were not reported; *R* represents the count of reported items; *D* represents the count of recorded items.

calculate a false alarm rate requires that i be divided by the sum of entries in the second row of the table, the value of which is neither known nor estimable. False alarm rates can be computed in recognition memory experiments because test items either were or were not presented to the subject during the acquisition phase of the experiment, but in free recall situations, although the items presented during the acquisition phase are known, there is no reasonable way to define the number of items that were neither recorded nor reported.

The weakness of the measures used in this report is that a match between a reported item and a recorded item cannot, by itself, be taken as evidence that the subject deliberately retrieved and reported a target. In general, in any situation in which an individual makes a choice response, a correct response is not necessarily evidence that the subject knew which response to make, and so some sort of correction for guessing is frequently applied. In other words, in such situations, the observed set of correct responses is assumed to be a mixture of genuinely correct responses and guesses that turned out to be correct. Thus, the meaningfulness of p(match) as a measure of memory performance is not clear: p(match) is surely an overestimate of the remembered fraction of the target list, but the extent to which it is an overestimate is not known.

Each of the indexes used in this report describes a seemingly important aspect of performance. Gordon B. Willis, in personal communication during June 1988, suggested that p(match) could be considered a measure of sufficiency—the extent to which the reported items achieve the goal of completely matching the recorded items—and that p(intrusions) could be considered a measure of inefficiency—the extent to which the report protocol was cluttered with items that do not contribute toward the goal of matching the recorded items. Clearly, performance is considered superior and is presumed to be based to a greater extent on retrieval from memory of the target items when a report protocol is both highly sufficient and efficient than when it is highly inefficient, regardless of its sufficiency.

The question, then, is whether there is a reasonable way to consolidate these aspects of performance into a single measure. Graesser (31), following Hilgard (43), suggested calculating a memory score defined as

$$\text{Memory score} = \frac{p(\text{match}) - p(\text{intrusions})}{1 - p(\text{intrusions})} \qquad \text{(A4)}$$

Willis independently suggested that another measure,

$$\text{Memory score} = \frac{m - i}{m + o} = \frac{[D \cdot p\ (\text{match})] - [R \cdot p\ (\text{intrusions})]}{D} = \frac{(2 \cdot m) - R}{D} \qquad \text{(A5)}$$

be used to characterize the memory performance of the subjects in the experiments discussed in this report. Each of these measures is rooted in the same underlying philosophy—to offset the count of matches of reported items to recorded items by subtracting information about the number of items reported, which is assumed to reflect the probability of matching by chance.

The problem with these measures, and the reason that they are not used, is that they make very strong assumptions about how intrusions should offset nominally correct responses. For example, the measure described in equation A4 is one of a family of measures that can be written as

$$\text{Memory score} = \frac{p(\text{match}) - [k \cdot p(\text{intrusions})]}{1 - [k \cdot p(\text{intrusions})]} \qquad \text{(A6)}$$

with k having some value between 0 and 1 that describes how the intrusion rate is charged against the match rate in order to produce a compromise score. In equation A4, $k = 1$. In some situations (for example, choice response), the appropriate value of k is obvious, and in others, reasonable arguments can be made for a particular value of k. For example, Graesser (31) reported experiments in which two versions of stories were presented to different groups of subjects, and intrusion rates were calculated by counting as intrusions only items reported by one group that had been presented to the other; items that were reported but that had not been presented to the other group were ignored. Thus, Graesser based his calculations on a closed set of items; because he knew how many items had not been presented, he could actually have calculated signal detection measures rather than the memory score that he reported. In contrast, the sets of items analyzed in the free recall experiments described in this report were open sets of items. In the absence of any knowledge about how to charge intrusions against nominally correct reports, that is, about how to specify a value for k in equation A6, it seems preferable to present and interpret the two measures separately. The measure defined in equation A4 was calculated for the data of experiment 2, and both the pattern of that measure and inferences based upon it were consistent with the reported interpretation of the two indexes that were used.

Appendix III Detailed results of experiment 2

This appendix presents, in more detail and with more formality than the main text, the results of experiment 2.

Effect of retention interval and reference period length

Table C shows the mean number of items recorded by subjects in each of the experimental conditions defined by retention interval and reference period length. Subjects who kept diaries for 4 weeks recorded significantly more different items than did those who kept diaries for 2 weeks, $F(1,162) = 112.52, p < 0.0001$. The length of the retention interval did not affect the number of items recorded, $F(3,162) < 1$. Table C shows that the mean number of different items recorded during a period of fixed length was quite stable over the groups of subjects assigned to each reference period length condition. There was no interaction of reference period length and retention interval on the number of items reported, $F(3,162) < 1$.

Table D shows the mean match and intrusion rates for subjects in the eight experimental conditions defined by retention interval and reference period length. Consistent with the results of experiment 1, dietary reporting performance declined as the length of the retention interval increased. Consistent with other data on memory decay, the change over time in the match and intrusion rates was negatively accelerated, so all tests of the effect of retention interval length used a contrast with a negatively accelerated "shape" (for example, the weights applied to the 4- and 6-week levels of retention interval were equal and, in sum, opposite to that applied to the 0-week level).

A multivariate analysis of variance (MANOVA) of the match and intrusion rates showed that the contrast on the levels of retention interval was significant, $F(2,161) = 46.82$, $p < 0.0001$; the residual was not significant. The pattern of means of each of the performance indexes was well described by this contrast: For match rate, $F(1,162) = 78.40$, and for intrusion rate, $F(1,162) = 15.19, p < 0.0001$.

The MANOVA showed a significant main effect of reference period length ($F(2,161) = 4.28$, $p < 0.02$) and a significant interaction of the reference period length and the contrast on retention intervals ($F(2,161) = 4.00$, $p < 0.05$). Univariate analyses showed these effects to be attributable to variations in intrusion rates. Although the average match rates of subjects assigned to the two reference periods did not differ ($F(1,162) < 1$), the intrusion rates of subjects who reported for a 2-week period were significantly higher than were those of subjects who reported for 4 weeks, ($F(1,162) = 8.58, p < 0.01$), and the interaction of the reference period length and the contrast on retention intervals was significant ($F(1,162) = 4.50$, $p < 0.05$). Table D shows that the absolute mean intrusion rate of subjects who kept diaries for 2 weeks increased steeply over retention intervals, whereas that of subjects who kept diaries for 4 weeks was roughly constant.

Effect of reporting instructions

Table E shows the number of items recorded and reported by subjects who received each of the three sets of instructions for each reference period length. Instructions did not influence the number of items recorded, $F(2,146) < 1$. However, subjects who received food group cues reported significantly more items than did those who received reverse chronological or meal cues, $F(1,146) = 10.97, p < 0.0002$; the mean numbers of items reported by subjects in the latter two conditions were not statistically different, $F(1,146) < 1$.

Table F shows the mean match and intrusion rates for subjects in the three instructional conditions. A MANOVA conducted on these measures showed an overall performance difference between subjects who received food group cues and those who received the two other types of instructions ($F(2,145) = 7.61$, $p < 0.001$), but showed no difference in the performance between the meal and reverse chronological conditions ($F(2,145) = 1.40$, $p > 0.25$). Univariate analyses showed that the differences among instruction conditions were due to differences in intrusion rates. Subjects who received food group cues exhibited significantly higher intrusion rates than subjects who received the two other types of cues, $F(1,145) = 15.05$, $p < 0.001$. The intrusion rates of subjects in the two other conditions did not differ significantly, $F(1,145) < 1$. Match rates did not differ significantly over the three conditions: For food group versus the other two conditions, $F(1,145) < 1$; for meal versus reverse chronological, $F(1,145) = 2.30$, $p > 0.1$. None of the interactions of instructions with other experimental conditions was significant.

☆ U.S. GOVERNMENT PRINTING OFFICE: 1991 — 2 8 1 - 7 9 8/ 4 0 0 1 1

Vital and Health Statistics series descriptions

SERIES 1. **Programs and Collection Procedures**—Reports describing the general programs of the National Center for Health Statistics and its offices and divisions and the data collection methods used. They also include definitions and other material necessary for understanding the data.

SERIES 2. **Data Evaluation and Methods Research**—Studies of new statistical methodology including experimental tests of new survey methods, studies of vital statistics collection methods, new analytical techniques, objective evaluations of reliability of collected data, and contributions to statistical theory. Studies also include comparison of U.S. methodology with those of other countries.

SERIES 3. **Analytical and Epidemiological Studies**—Reports presenting analytical or interpretive studies based on vital and health statistics, carrying the analysis further than the expository types of reports in the other series.

SERIES 4. **Documents and Committee Reports**—Final reports of major committees concerned with vital and health statistics and documents such as recommended model vital registration laws and revised birth and death certificates.

SERIES 5. **Comparative International Vital and Health Statistics Reports**—Analytical and descriptive reports comparing U.S. vital and health statistics with those of other countries.

SERIES 6. **Cognition and Survey Measurement**—Reports from the National Laboratory for Collaborative Research in Cognition and Survey Measurement using methods of cognitive science to design, evaluate, and test survey instruments.

SERIES 10. **Data From the National Health Interview Survey**—Statistics on illness, accidental injuries, disability, use of hospital, medical, dental, and other services, and other health-related topics, all based on data collected in the continuing national household interview survey.

SERIES 11. **Data From the National Health Examination Survey and the National Health and Nutrition Examination Survey**—Data from direct examination, testing, and measurement of national samples of the civilian noninstitutionalized population provide the basis for (1) estimates of the medically defined prevalence of specific diseases in the United States and the distributions of the population with respect to physical, physiological, and psychological characteristics and (2) analysis of relationships among the various measurements without reference to an explicit finite universe of persons.

SERIES 12. **Data From the Institutionalized Population Surveys**—Discontinued in 1975. Reports from these surveys are included in Series 13.

SERIES 13. **Data on Health Resources Utilization**—Statistics on the utilization of health manpower and facilities providing long-term care, ambulatory care, hospital care, and family planning services.

SERIES 14. **Data on Health Resources: Manpower and Facilities**—Statistics on the numbers, geographic distribution, and characteristics of health resources including physicians, dentists, nurses, other health occupations, hospitals, nursing homes, and outpatient facilities.

SERIES 15. **Data From Special Surveys**—Statistics on health and health-related topics collected in special surveys that are not a part of the continuing data systems of the National Center for Health Statistics.

SERIES 16. **Compilations of Advance Data From Vital and Health Statistics**—These reports provide early release of data from the National Center for Health Statistics' health and demographic surveys. Many of these releases are followed by detailed reports in the Vital and Health Statistics Series.

SERIES 20. **Data on Mortality**—Various statistics on mortality other than as included in regular annual or monthly reports. Special analyses by cause of death, age, and other demographic variables; geographic and time series analyses; and statistics on characteristics of deaths not available from the vital records based on sample surveys of those records.

SERIES 21. **Data on Natality, Marriage, and Divorce**—Various statistics on natality, marriage, and divorce other than those included in regular annual or monthly reports. Special analyses by demographic variables; geographic and time series analyses; studies of fertility; and statistics on characteristics of births not available from the vital records based on sample surveys of those records.

SERIES 22. **Data From the National Mortality and Natality Surveys**—Discontinued in 1975. Reports from these sample surveys based on vital records are included in Series 20 and 21, respectively.

SERIES 23. **Data From the National Survey of Family Growth**—Statistics on fertility, family formation and dissolution, family planning, and related maternal and infant health topics derived from a periodic survey of a nationwide probability sample of women 15–44 years of age.

SERIES 24. **Compilations of Data on Natality, Mortality, Marriage, Divorce, and Induced Terminations of Pregnancy**—Advance reports of births, deaths, marriages, and divorces are based on final data from the National Vital Statistics System and are published annually as supplements to the Monthly Vital Statistics Report (MVSR). These reports are followed by the publication of detailed data in Vital Statistics of the United States annual volumes. Other reports including induced terminations of pregnancy issued periodically as supplements to the MVSR provide selected findings based on data from the National Vital Statistics System and may be followed by detailed reports in the Vital and Health Statistics Series.

For answers to questions about this report or for a list of titles of reports published in these series, contact:

Scientific and Technical Information Branch
National Center for Health Statistics
Centers for Disease Control
Public Health Service
6525 Belcrest Road, Room 1064
Hyattsville, Md. 20782

301-436-8500

DEPARTMENT OF
HEALTH & HUMAN SERVICES

Public Health Service
Centers for Disease Control
National Center for Health Statistics
6525 Belcrest Road
Hyattsville, Maryland 20782

DHHS Publication No. (PHS) 92-1079, Series 6, No. 4