APPENDIX B TO SUMMARY REPORT: AN EVALUATION OF PUBLISHED MEASURES OF DIABETES SELF-CARE VARIABLES

Kent H. Marquis, John E. Ware, Jr., Roger Johnston, Susan Marquis, Marie Michnich, Barbara Rose, Judith Stein, Clairice Veit

A Rand Note
prepared for the
CENTER FOR DISEASE CONTROL, U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
The research reported herein was performed pursuant to Contract No. 200-77-0722 with the Bureau of Health Education, Center for Disease Control, U.S. Department of Health, Education, and Welfare.

The Rand Publications Series: The Report is the principal publication documenting and transmitting Rand’s major research findings and final research results. The Rand Note reports other outputs of sponsored research for general distribution. Publications of The Rand Corporation do not necessarily reflect the opinions or policies of the sponsors of Rand research.
APENDIX B TO SUMMARY REPORT: AN EVALUATION OF PUBLISHED MEASURES OF DIABETES SELF-CARE VARIABLES

Kent H. Marquis, John E. Ware, Jr., Roger Johnston, Susan Marquis, Marie Michnich, Barbara Rose, Judith Stein, Clairice Veit

A Rand Note
prepared for the
CENTER FOR DISEASE CONTROL, U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
PREFACE

To assist in the evaluation of diabetic patient education, the Bureau of Health Education of the Center for Disease Control, U.S. Department of Health, Education, and Welfare, sought to identify and develop reliable measures of patient (a) performance (behavior) in managing the disease, (b) knowledge of treatment tasks, and (c) attitudes toward carrying them out. This research effort was conducted by The Rand Corporation under Contract 200-77-0722.

This is one of three appendix volumes under the contract which support the final Summary Report:

Measures of Diabetic Patient Knowledge, Attitudes, and Behavior Regarding Self-Care: Summary Report, Kent H. Marquis and John E. Ware, June 1979 (R-2480-HEW).

The appendix volumes are as follows:

- Appendix A to Summary Report: Variables Judged Important in Compliance with Diabetes Treatment Regimens, Kent H. Marquis and John E. Ware, June 1979 (N-1151-HEW).
SUMMARY

This appendix identifies and evaluates published measures of diabetic self-management variables using secondary data sources. Attention is focused on the patient behavioral, knowledge, and attitudinal areas that an expert panel judged to be "very important" or "important" in achieving good management of diabetes (See Appendix A).

Comprehensive automated and manual searches of literature databases and indexes were undertaken. Copies of possibly relevant research materials were obtained, scanned, and briefly abstracted. Topics for which measurement evaluation information was available were assigned to staff members. Each staff member critically reviewed the published research in his or her assigned area and prepared an assessment of measurement quality.

The approach to assessing measurement quality is to estimate measurement bias and measurement error variance for existing measures of the important self-care variables.

Measurement bias is the difference between the measured mean and the true mean; measurement error variance is the difference between the measured variance and the true variance. These parameters were chosen because their effects on commonly calculated statistics can be easily interpreted.

We found less than half of the important self-care variables in the measurement literature. The quality of only a few of these measures could be evaluated and none of these evaluations involved diabetic populations.

In the nutrition area, the 24-hour recall method of assessing daily calorie, protein, carbohydrate, and fat intake was found to be minimally biased but subject to large amounts of measurement error variance. Time-sampling bias and error variance appeared to be relatively small: that is, the extrapolation of nutrition data from one day to longer time periods appeared fairly accurate. The literature is not conclusive on whether seven-day recall or seven-day diary measurement methods can improve upon the precision of the 24-hour
recall estimates. A comprehensive diet history interview approach is expensive and apparently very sensitive to interviewer biases. A short dietary questionnaire looks promising because it is inexpensive to use and may produce about the same quality of data as the 24-hour recall interview. No existing self-report measures appear to obtain very precise data (because they all contain substantial amounts of measurement error variance), but several contain little or no systematic response bias.

Self-report measures of oral medication compliance appear severely biased. All were obtained by doctors or by representatives of the patient's medical care facility. Apparently, patients do not like to admit to the doctor or his proxy that they have not been following the doctor's orders. Physician ratings of pill-taking compliance are not very informative. Physicians' average ratings are often very close to the overall compliance rates as measured by a pill count or urine test technique. However, they do no better than chance at predicting the compliance of individual patients. None of the pill-taking compliance studies included samples of diabetic patients.

Self-reports of visiting the doctor or dentist contain very low levels of response bias. Measurement error variance, however, appears to be moderate or high. There is no evidence that extending the length of the recall period reduces response error variance. Again, none of the research had samples of diabetic patients. There are no measures available to evaluate in the remaining important behavior, knowledge, and attitude areas. The next phase of this project will attempt to develop the needed questionnaire measures and begin to evaluate their reliability and validity with diabetic patients.
# CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREFACE</td>
<td>iii</td>
</tr>
<tr>
<td>SUMMARY</td>
<td>v</td>
</tr>
<tr>
<td>Section</td>
<td></td>
</tr>
<tr>
<td>I. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>II. METHODS</td>
<td>2</td>
</tr>
<tr>
<td>The Literature Search</td>
<td>2</td>
</tr>
<tr>
<td>Measurement Quality Estimation</td>
<td>11</td>
</tr>
<tr>
<td>Ease of Administration</td>
<td>20</td>
</tr>
<tr>
<td>III. NUTRITION BEHAVIOR</td>
<td>22</td>
</tr>
<tr>
<td>Dietary Assessment Methods</td>
<td>23</td>
</tr>
<tr>
<td>Measurement Evaluation</td>
<td>25</td>
</tr>
<tr>
<td>Conclusions</td>
<td>38</td>
</tr>
<tr>
<td>IV. ORAL MEDICATION</td>
<td>40</td>
</tr>
<tr>
<td>Takes Appropriate Dosage of Oral Medication</td>
<td>40</td>
</tr>
<tr>
<td>Obtains Prescriptions for Oral Medication</td>
<td>51</td>
</tr>
<tr>
<td>V. OBTAINING MEDICAL CARE</td>
<td>52</td>
</tr>
<tr>
<td>Doctor Visits</td>
<td>53</td>
</tr>
<tr>
<td>Visits to Specialists</td>
<td>61</td>
</tr>
<tr>
<td>Appointment Keeping Behavior</td>
<td>63</td>
</tr>
<tr>
<td>Measures of Other Behavior Items</td>
<td>64</td>
</tr>
<tr>
<td>VI. OTHER SELF-MANAGEMENT BEHAVIOR</td>
<td>65</td>
</tr>
<tr>
<td>Comprehensive Studies of Diabetic Patient Behavior</td>
<td>65</td>
</tr>
<tr>
<td>Urine Testing</td>
<td>66</td>
</tr>
<tr>
<td>Insulin Administration Behavior</td>
<td>69</td>
</tr>
<tr>
<td>Safety Behavior</td>
<td>70</td>
</tr>
<tr>
<td>Foot Care</td>
<td>71</td>
</tr>
<tr>
<td>VII. KNOWLEDGE TESTS</td>
<td>73</td>
</tr>
<tr>
<td>Content and Structure of Knowledge Tests</td>
<td>74</td>
</tr>
<tr>
<td>Data-Gathering Methods</td>
<td>77</td>
</tr>
<tr>
<td>Return Rates</td>
<td>78</td>
</tr>
<tr>
<td>Descriptive Statistics for Test Scores</td>
<td>78</td>
</tr>
<tr>
<td>Other Psychometric Criteria</td>
<td>79</td>
</tr>
<tr>
<td>VIII. ATTITUDE ITEMS</td>
<td>80</td>
</tr>
<tr>
<td>IX. CONCLUSIONS</td>
<td>81</td>
</tr>
<tr>
<td>Evaluation of Specific Measures</td>
<td>82</td>
</tr>
<tr>
<td>Further Measurement Research and Development</td>
<td>87</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>89</td>
</tr>
</tbody>
</table>
I. INTRODUCTION

This appendix has two purposes: (1) to identify published measures of important diabetic self-care variables and (2) to evaluate the quality of these measures using data in the research literature.

Existing measures have been located for some but not all of the variables judged important by the Delphi panel (Appendix A) and it was possible to evaluate the quality of only a small number of these measures. The results of the quality evaluations are fairly encouraging: most (but not all) self-report measures studied appear unbiased or not importantly biased; that is, the measured mean and the criterion mean are usually very similar. For measures that do contain an important amount of response bias, there are modifications that can be made to existing measurement procedures that may reduce the bias. All evaluated measures potentially contain moderate to large amounts of error variance (or "noise"). Ways of possibly reducing this category of error (and, thereby, increasing estimation precision) are discussed.

The first part of this appendix describes the major methodologies used in this phase of the research: the literature search procedures and the methods of evaluating measurement validity and reliability from published research.

The remainder of the appendix is organized around the general classes of self-care variables. Evaluations of measures of behavior in the areas of nutrition, oral medication, and obtaining health care are discussed in Sections III-V. Section VI discusses behavior measures that were found but could not be evaluated. Sections VII and VII address the existing measures of diabetic patient knowledge and attitudes. Section IX discusses the current measurement state of the art and the large range of uncertainty that still faces researchers wanting to select good measures of diabetic self-care behavior, attitudes, and knowledge.
II. METHODS

In this section the methods used to identify and evaluate existing measures of important variables are discussed. The first task was to find the relevant published literature. The automated and manual search strategies and the scanning and abstracting procedures are described and evaluated here. The second task was to evaluate the quality of existing measures. The second part of this section discusses the definitions, assumptions, and estimation procedures used to prepare the evaluations reported later in this appendix.

THE LITERATURE SEARCH

A first step in carrying out the research was to conduct a comprehensive review of the available literature to identify and evaluate the quality of available measures of diabetic self-care material potentially relevant to measurement procedures and evaluation of measurement quality. Additional methods of obtaining relevant source materials were used as well. The computerized data bases used are described below along with the keywords and keyword strategies employed for the automated searches, the manual data bases and other sources searched, the scanning process, and the critical review process performed by the staff of the project. In addition, there is a discussion of possible limitations of the literature search.

Computerized Search

Bibliographic Data Bases. Rand has over sixty machine-readable bibliographic data bases available to researchers. With the help of Rand research librarian Shirley Lee, the most useful data bases were selected. They covered the fields of medicine, science, food and nutrition, psychology, sociology, and the other behavioral and social sciences. The following is a brief description (supplied by research librarian Barbara Quint) of each selected data base:
AGRICOLA: provides comprehensive worldwide coverage of journal and monographic literature in the field of agriculture and related subjects. These include rural sociology, agricultural chemistry, food and nutrition, and agricultural economics.

CATLINE: contains full bibliographic data for all material catalogued by the National Library of Medicine; it covers books, proceedings, technical reports, new journals, serial publications, audiovisual materials, and monographs.

CDI (Comprehensive Dissertation Index): is a definitive subject, title, and author index to 99 percent of all American doctoral dissertations. Canadian and foreign dissertation coverage is increasing.

MEDLINE (MEDLARS On-Line) (Backfile): consists of citations to articles from more than 3000 journals in the medical and health fields. It is the most comprehensive index to medical journal literature in the world. Backfile contains earlier Medline material.

NTIS (National Technical Information Service): is a broad, cross-disciplinary file of government-contracted research reports from over 240 agencies (including NASA, ERDA, DOD, HEW, Commerce, etc.) plus other federal government studies and a miscellany of material. As the major unclassified reference tool to contract research, it is one of the most frequently used and useful files available.

PSYCHABS (Psychological Abstracts): indexes and abstracts the worldwide literature in psychology and related behavioral sciences. It covers books and reports as well as over 900 periodicals.

SCISEARCH: Includes all the citations contained in Science Citation Index, which has the broadest coverage available of scientific journal literature. Over 2600 periodical titles range from acoustics to zoology and include medicine, behavioral sciences, physics, etc.
Social SCISEARCH: the Social Sciences Citation Index offers the broadest coverage of social science journals of any single source. It includes over 1000 journals indexed completely and over 2200 indexed selectively. It indexes both by authors and cited references and by authors who cited the source. It includes keyword indexing of article titles from citing sources. The journals cover the range of the social sciences from anthropology to sociology.

Sociological Abstracts: covers over 1200 journals and other serial publications as well as monographs in the field of sociology and related disciplines including mass communications, demography, policy and planning, etc.

Keywords and Keyword Strategies. In a computerized data base, a keyword may be found in assigned subject headings, the title of the publication, or embedded in the abstract. Table 1 lists the keywords and the data bases with which they were used. Most chronic conditions involve important self-care activities, so the search was not limited to diabetes-related materials. Many of the behavior, knowledge, and attitude measures of interest have been tested only with nondiabetic patients.

Strategies for combining the keywords were then devised to avoid obtaining citations of articles not on the subject. Keywords such as "patient" or "medical" used alone produced thousands of citations, the vast majority of which were of no interest to this project. Using the technique of chaining keywords with "and" and "or," publications more directly on the topics of interest could be uncovered. An example of an interactive on-line session performed by a Rand research librarian is shown in Table 2. Other sessions employed similar keyword strategies.

The results of the general search indicated that published measurement evaluations existed in three relevant areas: nutrition, pill taking, and obtaining medical care. Additional, in-depth searches
<table>
<thead>
<tr>
<th>Keyword</th>
<th>MEDLINE (Medlars) 74-77</th>
<th>NTIS 64-77</th>
<th>PSYCHABS 64-77</th>
<th>SCISEARCH 74-77</th>
<th>SOCIAL SCISEARCH</th>
<th>CIDI 1861-1977</th>
<th>SOCIOABST, 63-77</th>
<th>CATLINE 65-77</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adherence</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Compliance</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Self-care</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Behavior</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Non-compliance</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Diabetes</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dropouts</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combination of</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>above with:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluation or</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test or Method or</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Measurement</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEDICAL +</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compliance</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Noncompliance</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Therapeutic + Compliance</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Behavior therapy</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chronic disease</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes mellitus</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attitudes to health</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Authors Search:</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taylor or Sackett</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>or Becker or Cordis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient or Health</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>or Medical or Care or</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Behavior</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Table 2**

**AN EXAMPLE SEARCH SPECIFICATION USING MEDLINE**

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: PATIENT$1 AND EDUCAT$3</td>
<td>Requests articles containing references to both patients and education. The truncation &quot;$#&quot; allows any combination of letters up to the number specified after the root; e.g., &quot;educate,&quot; &quot;educating,&quot; and &quot;education&quot; will be picked up with the term &quot;EDUCAT$3&quot;.</td>
</tr>
<tr>
<td>RESULT</td>
<td>774</td>
</tr>
<tr>
<td>774 articles were retrieved, an unwieldy number which undoubtedly includes many irrelevant citations.</td>
<td></td>
</tr>
<tr>
<td>2: SELF ADJ CARE</td>
<td>6 articles cited that mention the topic &quot;self care&quot;.</td>
</tr>
<tr>
<td>RESULT</td>
<td>6</td>
</tr>
<tr>
<td>3: (1 or 2) and DIABET$3)</td>
<td>Requests articles in either of the first two categories that also contain information on &quot;diabetes,&quot; &quot;diabetics,&quot; or use the word &quot;diabetic.&quot;</td>
</tr>
<tr>
<td>RESULT</td>
<td>27</td>
</tr>
<tr>
<td>27 articles are found. This is a manageable number. These 27 citations will be printed out and copies of articles that have not yet been incorporated in the literature search will be obtained.</td>
<td></td>
</tr>
</tbody>
</table>
were conducted in each of these areas, using the data bases and key-
words shown in Table 3. The data bases chosen for these particular
searches were judged to be the most likely sources for information
on these more specialized topics. Blank areas in the charts indicate
subject combinations that would have been too general, not useful
enough to warrant an expensive search, or redundant and therefore not
a cost-effective use of data base.

In addition to the above computerized searches, the UCLA-based
"Databank of Program Evaluations" (DOPE) was used as well. Search
topics used were: "weight problems," "self-help behavior," "attitudes
and values," "opinions," and "knowledge and learning ability".

Validity of Machine Searches: A further search using synonyms
for the major keywords was undertaken to ascertain that all relevant
topics were accessed during the original machine searches. MEDLINE
was used for the validation search since it was the most fruitful
source of materials for this study and is quite inexpensive to use.
The thesaurus for MEDLINE was consulted for keyword synonyms and pos-
sible alternative selections.

A strategy of first eliminating all previously cited references
and calling for all articles that were linked with the three subject
areas in Table 3 produced 63 new citations. Most were not relevant,
several had already been identified by other procedures, and six
proved to be relevant, previously undetected articles. Most of
these had been inserted into the data base after we had conducted the
original searches.

Many of the most useful articles, especially those on visit re-
porting and nutrition measurement evaluation, were published in the
1950s and 1960s, years not included in the machine searches. Private
literature collections of the principal investigators or their col-
leagues proved more useful than the mechanized searches in these par-
ticular areas.
### Table 3
SECOND ROUND (SPECIALIZED) SEARCH STRATEGIES

<table>
<thead>
<tr>
<th></th>
<th>MEDLINE (Medlars) 72-77</th>
<th>PSYCHABS 64-77</th>
<th>AGRICOLA 70-77</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIET (Dietary) +</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Compliance</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Regimen</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Therapy</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Behavior</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Measurement</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PRESCRIPTIONS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Filling</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compliance</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>APPOINTMENT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypertension</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arthritis</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Manual Searches

Library Data Bases. Manual searches were performed by Rand research librarian Shirley Lee on Index Medicus (1964–77), Medical Socioeconomic Sources (1971–77), and Nutrition Abstracts (1971–77). All "patient" headings were searched in Index Medicus (such as "patient compliance," "patient attitude") as well as "behavior therapy," "chronic disease," "diabetes mellitus," and "attitude to health." All "patient" headings as well as "attitudes to health" were searched in Medical Socioeconomic Sources. "Diet," "dietary," and "weight control" headings were searched in Nutrition Abstracts.


Other Sources. There were several other ways in which materials were obtained for this report. The Annotated Bibliography on Compliance of Patients with Therapeutic Regimens, (Haynes and Taylor, 1977) was searched. Unpublished doctoral dissertation materials of Marie Michnich, one of the researchers on this project, and her large bibliography on compliance amassed during her own research on compliance were used as well. Many review articles on compliance have been published and the large bibliographies of these articles cited many primary source materials. In addition, as primary articles were scanned and/or abstracted by the research staff, further pertinent referenced studies were obtained.

Scanning and Abstracting

The liberal keyword strategy used in the machine searches to avoid missing important references produced hundreds of citations, many of which were not applicable to the topic explored by this project. Staff members read the machine-generated citations and abstracts and obtained copies of all articles that could possibly deal with
diabetes and other chronic diseases, empirical measurement of compliance, patient self-care, knowledge, or education. The staff, in some cases, had to rely solely on the titles when deciding whether an article was relevant or not. Works by previously identified key authors were ordered as well. During the scanning, the staff read several hundred articles and summarized them on forms created specifically for this project. The most relevant studies concerned with measurement of the behavior categories judged important by the Delphi panel were examined in more detail. Each staff member prepared a summary of empirically measured, self-care behavior discovered in the literature, including information about measurement procedures, population characteristics, and available evaluations of the reliability and validity of the measures employed.

Critical Review

After the initial scanning phase of the literature review, specific topics for which measurement evaluation information was available were assigned to staff members for more detailed abstraction and interpretation. Such topics included compliance with oral medication prescriptions, prescription filling, measurement of attitude and knowledge, doctor's visits and appointment keeping, dietary compliance, as well as existing measures for diabetic care such as insulin injection, urine testing, safety measures, foot care, and other residual items. At this point, the researchers critically evaluated the quality of the studies and identified sources and magnitudes of measurement bias and error variance. Items of interest included the measurement procedure used, how measures were scored, patient demographic characteristics, numbers of subjects (originally and at the end of the study), and the delivery setting. In those cases where it was possible, estimates of bias and error variance were derived. The results of this review are the basis of the information reported in later sections of this appendix.
Limitations of the Literature Search

The most important weakness of any literature search is not knowing of the existence of an important publication. Omissions can result from several kinds of problems: the data bases do not cover all relevant journals, the keyword strategies are not comprehensive enough, or other restrictions are placed on the search.

The data bases searched are extremely comprehensive in their coverage of relevant sources. Their main limitation is that often they fail to include research published in the past (e.g., more than 10 years ago). Much of the material cited in the appendix was published many years ago; however, there was no apparent way of assessing how many of the "classics" were missed. The comprehensiveness of the keyword strategies did appear to be good because the validity test indicated a very low false negative rate. The only major other restriction placed on the searches was to exclude articles published in a language other than English.

In summary, the searches conducted were especially comprehensive. The search validity study indicated that the marginal productivity of additional searches with other keywords was very low. The only potentially important weakness of the search results was the incompleteness of the data bases' coverage of publications more than 10 years old. However, many important articles published over 10 years ago were cited in other studies and then obtained by the staff.

MEASUREMENT QUALITY ESTIMATION

The second research objective was to evaluate the reliability and validity of existing measures. The evaluation methodology and the reasons for selecting it are explained next.

Measurement reliability and validity are defined in terms of a simplified "working" model of measurement bias and measurement error variance. Derivations are presented to show how these terms in the model can be used to anticipate the effects of measurement error on statistics commonly calculated in evaluation research. The evaluation researcher, then, can identify the kind(s) of statistics
and sample sizes to be used in his study and use the parameter estimates reported later in making his own measurement decisions.

The applied estimation is based primarily on criterion validity data. Some criterion validity data may contain errors of nonobservation (e.g., sampling) and errors due to processing (e.g., scoring) in addition to measurement errors. The various sources of bias and error variance are explained and the methods used to estimate them briefly reviewed.

The final part of the discussion mentions some of the limitations of this approach.

**Working Model**

A simplified model of measurement error is used. It specifies parameters that, with assumptions, can be estimated from the published literature. When the term measurement is used, it refers to actual observations. Measurement sources of error are distinguished from other error sources in this section.

The simplified model employs assumptions about what is contained in an observation for a particular person, \( i \):

\[
M_i = T_i + B_i + e_i
\]

where \( M_i \) = The measured (observed) value for person \( i \) on the target measure

\( T_i \) = The person's true value on the characteristic of interest

\( B_i \) = The person's response bias which is assumed to be uncorrelated with \( T \) or \( e \)

\( e_i \) = Person \( i \)'s random response error which is uncorrelated with \( T \) and \( B \) and has an expected mean value of zero over independent replications of \( M_i \).

The major evaluation approach is criterion validity. It consists of comparing the target measurement, \( M \), to criterion measurements, \( C \). The working model for a person's score on the criterion is
\[ C_i = T_i + e_{ci} \]

where \( T_i \) = The true score for person \( i \)

\( e_{ci} \) = The random error for individual \( i \) in the criterion value.

\[ \text{Cov} (T, e_{ci}) = 0, \text{Cov} (e_{ci}, e_i) = 0, \text{Cov} (e_{ci}, \overline{T}) = 0 \]

and \( \text{E} (e_{ci}) = 0 \).

**Basic Estimation Strategy**

A set of observations from individuals is to be evaluated for bias and error variance.

Bias is defined to be the expected value (the population average value) of the difference between measured scores and true scores. Letting a bar (-) over a variable denote the population mean,

\[ \text{Bias} = \overline{B} = \overline{M} - \overline{T}. \]

Because the criterion error has an expected value of zero, \( \overline{T} = \overline{C} \), and the bias can be rewritten as

\[ \overline{B} = \overline{M} - \overline{C}. \]

The estimate of bias based on a sample is the difference between the sample mean and criterion mean (proportion, etc.).

Letting \( \text{Var} M \) denote \( \text{E} (M - \text{E}(M))^2 \), and letting \( \text{Var} T \) denote \( \text{E}(T - \text{E}(T))^2 \), etc., and using the previous assumptions, then

\[ \text{Var} M = \text{Var} T + \text{Var} B + \text{Var} e \]

\[ \text{Var} C = \text{Var} T + \text{Var} e_{ci} \]

\[ \text{Cov} (M, C) = \text{Var} T. \]

The measurement error variance, then, is

\[ \text{Var} B + \text{Var} e = \text{Var} M - \text{Cov} (M, C). \]
Following classical test theory, the "reliability" of a measure is defined as the ratio of the population variance of true score values (Var T) to the population variance of measured values (Var M). Letting IR denote the index of reliability,

\[ IR = \frac{\text{Var } T}{\text{Var } M}. \]

One minus the index of reliability is the proportion of error variance in the measured variance.

The literature often reports the correlation coefficient between the target and criterion measures. The correlation coefficient between ratio of the covariance of the criterion and target measure to the geometric mean of the two variances:

\[ r = \frac{\text{Cov } (M,C)}{\sqrt{\text{Var } M \text{ Var } C}}. \]

Recalling from above that the covariance of the target and criterion is equal to the true score variance, the correlation may be rewritten as:

\[ r = \frac{\text{Var } T}{\sqrt{\text{Var } M \text{ Var } C}}. \]

If the target and criterion contain equal amounts of error variance (Var \( E_C = \text{Var } B + \text{Var } e \)), then Var M = Var C and

\[ r = \frac{\text{Var } T}{\text{Var } M} = IR. \]

If the criterion is error free, i.e., Var \( e_C = 0 \), then

\[ r^2 = \frac{\text{Var } T^2}{\text{Var } M \text{ Var } T} = \frac{\text{Var } T}{\text{Var } M} = IR. \]

*The correlation coefficients in the literature are calculated from samples and so are estimates of the ratio; however, for convenience we express them in terms of the population parameters.
That is, assuming that the criterion error variance is less than or equal to the measurement error variance in the target measure, the index of reliability is bounded by $r^2$ and $r$; the proportion of error variance in the target measure is bounded by $1 - r$ and $1 - r^2$.

**Estimation for Dichotomous Variables**

Many of the published criterion validity data employ dichotomous variables (e.g., yes/no or 0,1). The basic estimation strategy used in this case is explained next.

Dichotomous criterion validity data may be thought of in terms of the $2 \times 2$ table portraying their cross-classifications.

<table>
<thead>
<tr>
<th>Criterion Value</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Target Measure</strong></td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td><strong>Value</strong></td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td></td>
<td>$a + c$</td>
<td>$b + d$</td>
</tr>
</tbody>
</table>

The rows are the target measure values (e.g., responses from patients) and the columns are the criterion values. $a + b + c + d = 1$, so the cell entries are proportions of the total number of observations. The measured mean of the variable (the proportion of persons having the value 1) is $a + b$; the mean of the criterion is $a + c$.

The bias in the target measurement is:

$$
egin{align*}
\bar{B} = (a + b) - (a + c) = b - c.
\end{align*}
$$

A positive difference reflects an estimated net overstatement of criterion-defined truth in the target measurements and a negative difference indicates a net understatement bias.

The product moment correlation, $\phi$, for dichotomous variables is

$$
\phi = \frac{(ad - bc)}{\sqrt{(a + b) (c + d) (a + c) (b + d)}}.
$$
Phi has the same interpretation as the correlation coefficient, $r$, discussed above.

**Relationships of Measurement Bias and Error Variance to Statistics**

Measurement properties are described in terms of the two parameters $\bar{B}$ (bias) and IR (reliability index) because these parameters are easily interpreted with respect to their effect on commonly calculated statistics. These interpretations are sketched briefly here using previously defined notation and assumptions.

When the target measure is used to estimate a sample mean, $\bar{m}$, then

$$\bar{m} = \frac{1}{n} \sum_{i=1}^{n} M_i \quad (n = 1 \ldots i \ldots n), \quad E(\bar{m}) = \bar{M} = \tau + \bar{B}.$$ 

The measured mean is biased by the amount $\bar{B}$.

The variance of the estimated mean of the measured values is higher than the variance of a sample mean of a perfectly measured attribute. The variance of the measured mean is:†

$$\text{Var} \bar{m} = (\text{Var} \tau + \text{Var} B + \text{Var} e)/n.$$ 

The variance of the sample mean of a perfectly measured attribute is:

$$\text{Var} \bar{\tau} = \text{Var} \tau/n.$$ 

Hence the precision in estimating the mean is decreased by measurement error. The effect is the same as reducing sample size by $n(\text{Var} \tau/\text{Var} \bar{M}) = n \text{IR}$.

When the target measure is used to obtain the difference between two means (e.g., a between-group comparison or a before-after design) the measurement errors have somewhat different effects:

*The lower case $\bar{m}$ is used to denote the sample estimate of $\bar{M}$.

†Assuming a simple random sample and that measurement errors are not correlated across individuals.
\[ E(\bar{m}_1 - \bar{m}_2) = \bar{M}_1 - \bar{M}_2 = (\bar{T}_1 - \bar{T}_2) + (\bar{B}_1 - \bar{B}_2). \]

If \( \bar{B}_1 = \bar{B}_2 \), then \( \bar{m}_1 - \bar{m}_2 \) is an unbiased estimate of the difference between the two means.

The effect of measurement errors on the variance of the mean difference depends, in part, on the covariance of the biases, Cov \((B_1, B_2)\):

\[
\text{Var}(m_1 - m_2) = [\text{Var}(T_1 - T_2) + \text{Var}B_1 + \text{Var}B_2 - 2 \text{Cov}(B_1, B_2)]
+ \frac{\text{Var}e_1 + \text{Var}e_2}{n}.
\]

The measured variance of the difference is affected by noncovarying error variances and precision is decreased. The effective sample size is

\[
n [\text{Var}(T_1 - T_2)/\text{Var}(M_1 - M_2)] = n \text{IR}_{M_1 - M_2}
\]

where \( \text{IR}_{M_1 - M_2} \) is the reliability of the measured difference.

The target measure, \( M \), correlation with another variable, \( Z \), that is perfectly measured is a biased version of the true correlation:

\[
r_{m,z} = \frac{\text{Cov}(M,Z)}{\sqrt{\text{Var}M \text{Var}Z}}
= \frac{\text{Cov}(T,Z)}{\sqrt{\text{Var}M/\text{Var}T} \text{Var}T \text{Var}Z}
= r_{t,z} \sqrt{\text{Var}T/\text{Var}M} = r_{t,z} \sqrt{\text{IR}}.
\]

The calculated coefficient is attenuated by \( 1 - \sqrt{\text{IR}} \). It is unaffected, however, by \( \bar{B} \).

When the target measure is used as a dependent variable in a regression, the calculated slope is unbiased but the intercept is affected by the \( \bar{B} \).

Define the true relationship between \( D \) and \( Z \) as:
D = k + ΔZ + u.

Z is perfectly measured. When the slope, Δ, and intercept, k, are estimated using imperfect measures, M, of the variable D:

\[
\text{Slope} = Δ = \frac{\text{Cov}(D,Z)}{\text{Var} Z} = Δ, \quad \text{but}
\]

\[
\text{Intercept} = k = D - Δ \overline{Z} + \overline{M}.
\]

The above discussion is meant to illustrate how information about measurement errors can be used in applied contexts and why the following sections attempt to estimate the measurement parameters \( \overline{M} \) and the index of reliability for the target measures. The illustrations assume simple models and ignore several other types of bias found in the measurement literature. A more complete discussion may be found in Marquis and Marquis (1977). In developing measures for research use, the other forms of bias and their effects need to be considered.

**Sources of Error**

In evaluating the quality of measurements in this appendix, primary attention is focused on errors in target observations (e.g., how close does what a patient reports come to some criterion value). The total error in a research study, however, comes from many sources. Often, error from one source is misinterpreted as error from another. In evaluating measurement (observation) error, discussions in this appendix must sometimes consider (and attempt to estimate) the other errors. A schematic representation is given below of the major sources and types of errors along with some examples of each.

In the nutrition literature, the two sources of error most frequently confused are time sampling and observation. The fact that a particular day's food intake is not exactly like intake on other days is sometimes incorrectly referred to as an observation error (measurement bias or measurement error). It is important to keep the two sources of error separate because remedies for each are potentially different.
In large sample criterion validity studies, such as those used for reporting patient visits to the health care system, processing errors are potentially important. The clerical task of matching the right visit record to the right survey report is cumbersome and possibly error prone. Errors introduced in this way can cause inflated estimates of the measurement error parameters.

The doctor visit studies (and others) sometimes suffer from a sampling bias. Only certain kinds of events are sampled, others are unobserved. The nonobservation bias (sometimes called a design bias) can produce important distortions in interpretations of measurement effects.

**Limitations**

Simplified models of measurement error are used partly to accommodate the limited types of data and estimates reported in the literature. The reader should be aware that several important types of
measurement error are not represented in our models because they cannot be estimated with available data. The main omission is the type of response bias that is correlated with the true value of the variable Cov (T, B) ≠ 0. An example would be patients who do not comply and lie on their self-reports, while patients who comply tell the truth. Other types of bias not represented in the model include that correlated with characteristics of the patients (e.g., older people are more prone to forgetting or physicians rate people with more education as more likely to follow a treatment regimen) and bias correlated with an experimental treatment. Each of these biases has somewhat different effects on calculated statistics. Ideally, information about them would be available for any measure considered for use in research.

An assumption is made that true score variance is the same in both target and criterion measures. Violations of this assumption result in overestimates of error variance when using estimates based on the simplified model. While we do not believe the assumption is wrong, there is no empirical evidence to support it either.

EASE OF ADMINISTRATION

The original intent was to create a quantitative comparison of similar methods of measuring the same underlying trait on ease of administration. Two steps would be necessary to accomplish this: specifying the variables that influence ease of administration and expressing their values in common units such as dollars.

This elaborate approach has not been adopted because the cost estimates needed to standardize disparate pieces of information cannot be obtained. Precision "costs" of missing data vary with the statistics to be calculated and the sample sizes being used. Dollar costs of using special interviewers and coders (e.g., nutritionists) are a function of whether they are already on a research project's staff.

The approach adopted here is to mention considerations that influence ease of administration so that a potential user of a measurement procedure can begin to evaluate the tradeoff between cost and
quality for his own research context. The description of measures includes information about needed skilled personnel to obtain or score the patient data. It also includes available information about patient cooperation rates such as refusals to participate and attrition from the research.

Finally, quality evaluations cannot be made for a large number of measures so that very few choices among measures of the same underlying trait are presented. In some cases where a choice is presented, the complex, difficult-to-administer measure yields a lower quality measurement than cheaper procedures. The choice, in this case, can be made without a formal consideration of administration costs. Other choices are often between unbiased procedures, with the cheapest approach yielding less precision. By conducting simple pretests in his own research setting, the potential researcher can generate his own information about the tradeoff between measurement and sample size effects on the precision of his intended estimates.
III. NUTRITION BEHAVIOR

Adherence to a prescribed diet is considered a basic component of the diabetic treatment regimen, although the content and level of adherence required often vary widely. Regardless of the dietary philosophy, assessment of food intake with some particular or general focus is a primary element of almost every evaluation of a diabetic treatment program. The following nutrition behavior items were selected by the expert panel as important in the patient's management of his diabetes:

- Achieves and maintains ideal body weight (Metropolitan Life table value ± 10 percent).
- Has and uses food exchange list (insulin-dependent diabetic).
- Changes eating time to counteract effects of exercise (insulin-dependent diabetic).
- Eats prescribed ratio of protein to carbohydrate to fat daily (insulin-dependent diabetic).
- Changes eating time to prevent development of hypoglycemia.
- Measures portions with measuring cup, measuring spoons, and scales or learns to judge portion amounts and weights accurately (within 10 percent) (insulin-dependent diabetic).
- Plans alcohol intake with reference to other meal content (insulin-dependent diabetic).
- Eats prescribed meals at prescribed times (insulin-dependent diabetic).

With the exception of the first item concerning body weight (the measurement of which requires a table of ideal weights and an accurate scale), some type of dietary assessment procedure is required. Fortunately, the literature on dietary assessment methods is large, with several comparative studies to aid in the description and evaluation of the major techniques available.
The goal of this section is to summarize the dietary assessment methods in general use, to present an analysis of the measurement characteristics of alternative methods, to evaluate the relative ease of administering different methods, and to provide recommendations for further tests on existing measures and/or improvements in existing techniques.

**DIETARY ASSESSMENT METHODS**

There are three major approaches that have been used to collect *individual* food intake information: the food record or diary, recall of past intake (typically a day or a week), and the diet history. A number of other methodologies have been used to assess food intake of families, such as food inventories and expenditure records. However, because the diet habits and requirements of diabetics may differ from other members of the family, only data collection methods that focus on the food consumption of the individual are reviewed. The basic features of the three approaches are described below.

**The Food Record or Diary**

In this method, the subject concurrently records the amount of food eaten each day during the survey period. Recording may be in grams (requiring weighing), in measured household units (cups, teaspoons, etc.), or in estimated household units.

The weighed food record *is considered by many workers to be the most accurate technique for assessing diet (e.g., Becker et al., 1960; Pekkarinen, 1970). However, the weighed record requires extensive cooperation among the subjects and usually close supervision by nutritionists (Marr, 1971, Pekkarinen, 1970). Hence, it is both a costly technique and not appropriate for wide application. For these reasons, the weighed food record is not evaluated; it is used here as a criterion measure against which to evaluate other data collection techniques.

*There are two types of weighing techniques: the precise weighing method and the weighed inventory method. These techniques are described in Marr (1971).
Food diaries recorded in household units (e.g., cups) require less supervision and no special equipment. The advantage of the food record compared to a recall interview or diet history is that it does not rely on the subject's ability to remember past intake. However, the food diary does require more participation by the subject than the interview methods; it may change diet habits (Hawthorne effect) and hence not reflect normal consumption.

Diet Recall

In the diet recall interview, the subject reports the amount of foods consumed during some prior prespecified time period. The most common recall period is 24 hours; however, some investigators have questioned respondents about a full week's consumption in one interview. Quantities may be estimated in household units, by the use of food models, or a combination of these techniques.

The diet recall interview usually asks about each meal, beginning with the most recent. An alternative approach, called the list-recall procedure, structures the questioning around food items rather than meals (Hankin et al., 1975).

The disadvantages of the diet recall are its reliance on the subject's memory and the possibility that food consumption during a short period may not be typical of usual diet habits. The level of training of interviewers required for the diet recall method is not clear from the literature. All of the evaluation studies of the recall interview employed nutritionists as interviewers, so it is not possible to assess the effect of using less specialized personnel as interviewers. If nutritionists are required to obtain data of acceptable quality, the diet recall method may prove costly.

Balogh et al. (1968, 1971) devised a modified diet recall questionnaire that they feel can be administered by personnel other than nutritionists. The questionnaire asks about frequency of food consumption over the past week, using a list-recall procedure, and the size of the usual portion. Thus, it reflects elements of both the diet recall and diet history (discussed below). In field tests of the questionnaire, nurses and nutritionists have been used as interviewers.

Food models are pictures or physical replicas that illustrate different quantities of a particular food.
Diet History

The diet history method was developed by Burke (1947). The diet history differs from the diet recall in that questions are asked about general food habits over a long period of time rather than specific consumption during a day or week. On the surface, the diet history would appear to be a superior method for assessing usual diet practice.

The method developed by Burke employs several different approaches to obtain usual dietary intake. In the first part, the overall pattern of eating is asked, e.g., "What do you usually have for breakfast?" In the second, the "cross-check", questions are asked about specific foods, such as likes and dislikes, use, etc. The third part consists of a menu for the prior three days.

The diet history takes about an hour or more to administer (Burk and Pao, 1976) and requires highly trained nutritionists as interviewers.

MEASUREMENT EVALUATION

In the remainder of this section, the evidence on the biases and amount of error variance in self-report measures of diet and food consumption is reviewed. Biases are assessed by the mean difference between self-report and criterion measures; error variance is estimated by the correspondence between individual values of self-report and criterion measures. The criterion measure is, where possible, a weighed record because of its acknowledged superiority by practitioners in the field. In cases where a comparison of a self-report technique and a weighed record is unavailable, two different self-report techniques are compared.

Two kinds of information are needed for an accurate estimate of diet habits: the kinds and amounts of food consumed and an accurate determination of their nutrient values. The evaluation focuses on the reporting of kinds and amounts consumed and bypasses issues of the accuracy of food conversion tables. If precise, unbiased measures of kinds and amounts of food eaten are identified, future investigators may select (or develop) conversion tables (or other methods) appropriate
to season and local conditions.

Comparisons are made of nutrient values obtained from food table conversions of both the self-report and weighed record. These comparisons contain any "coder" biases (effects of different nutritionists in calculating nutrient values) but they are probably small. Browe et al. (1966) compared calculations of nutrient intake on ten questionnaires by three nutritionists and found no differences in the mean values. No information about coder reliability is available.

Twenty-four-Hour Diet Recall: Bias

The 24-hour recall technique is acceptable to most respondents and appropriate for large-scale research studies.*

Two types of biases are of concern in evaluating the 24-hour diet recall: memory bias and time-sampling bias (see Fig. 1). Memory bias is usually thought of as a "forgetting" bias; that is, subjects will not remember all of the foods eaten and quantities will be underreported. Time-sampling bias arises because one 24-hour period may not adequately represent "usual" or average diet habits. For example, if diet patterns differ between weekdays and weekends, a 24-hour recall would reflect only one of these consumption patterns.

The studies reviewed next suggest that both the time-sampling bias and the response bias in measures of food eaten obtained by 24-hour recall are small.

Memory bias can be evaluated by comparing a 24-hour diet recall to a weighed record of the same meals. Table 4 summarizes the results of six studies which presented such comparisons. The general pattern of results suggests some underreporting of the recalled diet, but the amounts of underreporting are usually small. Unfortunately, none of the studies sampled diabetic populations.

*For example, The National Health and Nutrition Examination Survey (Miller, 1973).
Table 4
PERCENTAGE DIFFERENCE BETWEEN NUTRIENT VALUES
OBTAINED FROM WEIGHED DIET VALUES AND
SHORT-PERIOD RECALL INTERVIEWS

<table>
<thead>
<tr>
<th>Study</th>
<th>Observations</th>
<th>Number of Meals</th>
<th>Percentage Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Calories</td>
</tr>
<tr>
<td>Samuelson (1970)</td>
<td>99 children</td>
<td>School lunch</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>and adolescents</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Greger and Etnyre (1978)</td>
<td>17 female</td>
<td>Two 24-hour</td>
<td>-5</td>
</tr>
<tr>
<td></td>
<td>adolescents</td>
<td>periods</td>
<td></td>
</tr>
<tr>
<td>Thomson (1958)</td>
<td>20 women</td>
<td>24-hour period</td>
<td>-17</td>
</tr>
<tr>
<td>Bransby et al. (1948)</td>
<td>49 children</td>
<td>Three 24-hour</td>
<td>-2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>periods</td>
<td></td>
</tr>
<tr>
<td>Meredith et al. (1951)</td>
<td>94 children</td>
<td>School lunch</td>
<td>-5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Madden et al. (1976)</td>
<td>76 elderly</td>
<td>Noon meal</td>
<td>-10</td>
</tr>
</tbody>
</table>

NOTE: Studies compare same meals.
Confirmation of this finding is contained in a study by Linusson et al. (1974).* Recalled consumption over 24 hours by 86 hospitalized women was compared to weighed records kept by the nutritionists. In 13 of 14 food groups, the amount of food intake reported by the women was smaller than the recorded value. The largest errors were in three food groups (salads, soups, and sweets); the women underestimated by 50 percent the quantities of these foods. However, for the remaining ten food groups, the underreporting was small, averaging six percent.

The largest memory bias was detected in the Thomson study (1958) (Table 4). This study's methodology deviates in several important respects from the other six. The women were participating in a self-weighed record study. During the weighed survey week, they were asked to recall the food consumed during the previous 24 hours. The recalled diet was compared to the women's weighed record. In the other studies, however, served portions of leftovers were weighed by nutritionists. Secondly, the other studies evaluated recall of an institutional meal or meals. Institutional meals may be more standardized and hence easier to recall than normal household meals. Whether the larger deviation in the Thomson study reflects biases in a self-weighed record, a real forgetting bias in self-reports, or a small sample effect, cannot be conclusively determined from existing literature. However, another approach to evaluating the 24-hour recall method also suggests that memory bias in the recall diet is small. Bransby et al. (1948) compared the responses of children to three successive 24-hour recall interviews with the three-day diet of a comparable group of 22 children obtained from a weighed record kept by the mothers. Calories reported in the interview were ten percent higher than the weighed record, protein and carbohydrates were 13 percent higher in the interview, and fats four percent higher. However, as Peikkarinen (1970) notes, any recall bias in the recalled diets is confounded with the potential Hawthorne effect, that is, any effect that keeping a weighed record has on behavior. Since the

*Unlike the studies summarized in Table 4, Linusson et al. compared the recalled and weighed measures in food groups rather than the overall diet.
sign of the observed effects is opposite to those obtained in the studies reported earlier, it is likely that most of the difference is due to a change in meal preparation among women who weighed the child's food (the Hawthorne effect).

The total effect of memory bias and time-sampling bias in the 24-hour recall diet is assessed by comparing the interview measures to a weighed record maintained over a longer time period. Blake et al. (1962) compared a 24-hour recall diet obtained from 36 adults to a seven-day weighed record. Mean protein and calories obtained from the interview were smaller than the weighed record, but the differences were only about three percent. Since this difference is the same magnitude as the differences obtained in comparing a recall diet and a weighed record of the same meals, the Blake results suggest virtually no time-sampling bias. Unfortunately, no details on the Blake methods or the type of meals inquired about are available.

Ohlson et al. (1950) compared the three-day diet of 13 women obtained by three 24-hour recall interviews* to measures obtained from a ten-day weighed record maintained subsequent to the interviews. This study does not directly address the time-sampling bias of a single recall interview, but does provide information about how well a few repetitions approximate more long-term patterns and about the memory bias. The results showed a 12 to 14 percent higher calorie and protein content of the recalled diet than the weighed diet. Note that these comparisons include any Hawthorne effect of the weighed record as well as memory bias and time-sampling bias. Further, they are of the same magnitude of difference as found in the Bransby et al. study, reported earlier, which contained potential Hawthorne effects. Again, since few other evaluations of the 24-hour recall find a positive recall bias, we attribute the results of the Bransby et al. and Ohlson et al. studies to the Hawthorne bias in the weighed record rather than recall bias in the interview.

*It is not reported whether the recall interviews were continuous in time.
Several investigators (Madden et al. 1976; Linusson et al. 1974; and Young et al. 1952) have presented analyses of 24-hour diet recall which they interpret (and which others have cited) as evidence of a bias which is called the "average man bias" by some measurement specialists. That is, persons with a large food intake underreport their food consumption whereas persons with low food intake overreport their consumption.

The evidence cited as proof of the average man bias in 24-hour diet recall is the result of regressions of weighed record values on 24-hour recall measures. In the regressions, the estimated coefficient on weighed record values is less than one. However, as is commonly known, this result will be obtained if the weighed record values contain random measurement error. That is, even if the coefficient from a regression of true diet on the recalled measure equals one, random error variance in the independent variable will produce an estimated coefficient less than one. It is suggested that this statistical regression toward the mean phenomenon accounts for the findings of the authors mentioned above rather than a real average man bias in 24-hour recall.

In summary, although it is not possible to accurately estimate the time-sampling bias and recall bias in 24-hour recall interviews on the basis of existing literature, the evidence tends to suggest that these biases are small.

The only study specifically evaluating the 24-hour diet recall in a diabetic population is Watkins et al. (1967a). They obtained home interviews that included a 24-hour recall of food intake with 60 insulin-dependent, adult diabetic patients. They report an "unsatisfactory correspondence" between caloric intake estimates and records of patient weight and conclude that the 24-hour recall could not be representative of the subject's daily diet. The magnitude and direction of the discrepancies are stated so we do not know if they observed an average bias or offsetting disagreements that are characteristic of random measurement and sampling error. They substituted an analysis of responses about the regularity and spacing of meals but did not evaluate the quality of these data with respect to
meeting the estimation objectives.

24-Hour Diet Recall: Error Variance

There are also two important components to the error variance in the 24-hour diet recall measures: response-error variance and time-sampling variance. Response-error variance refers to the variation of deviations of the reported diet from the actual diet for that 24 hours; the time-sampling variance is due to the deviation of the particular days' diet from usual diet habits.

The proportion of response-error variance in the diet recall measures can be estimated by observing the correlation of the recalled reports with a weighed record for the same meals. (See Section II for an explanation of the interpretation of the correlation coefficient in terms of error variance.) Analyses of variance of self-reported diet data over longer time periods are used to estimate time-sampling variance.

The study by Linusson et al. (1974) discussed earlier reported correlations that ranged between 0.3 and 0.7 between amounts obtained from the recall interview and the weighed record in the 14 groups of food. Emmons and Hayes (1973) studied dietary interviews obtained from children in four grades with weighed records of their consumption of food during school lunch. The correlation of calories obtained under the alternate methods ranged from 0.23 to 0.77; for protein the correlations were 0.05 to 0.82. The correlations, however, tended to increase with grade, suggesting that very young children may be poor subjects for diet interviews. Madden et al. (1976) report a correlation of 0.35 between caloric values obtained by interview and weighed records in the lunches consumed by the elderly subjects; protein values correlated 0.28. Unfortunately, the Madden et al. study did not contain adequate weighing procedures and these correlations may reflect more than typical error in the weighed record. Estimates reported later suggest that the proportion of true score variance in 24-hour recall measurements is about 50 percent (implying \( r = 0.5 \) if the weighed measures have error variances similar to those of the recall measures). This is in the middle of the distribution.
implied by the above studies ($r = 0.3$ to $r = 0.8$) and represents a reasonable "working" estimate for purposes of evaluation.

Two evaluation studies of dietary measures present information about day-to-day variation in diet: the time-sampling variance. Hankin et al. (1967) present an analysis of variance of the daily caloric, protein, fat, and carbohydrate values of diets of 93 Japanese-American men living in San Francisco obtained from a seven-day measured record. Young et al. (1953) report a similar analysis for 18 adults from a 28-day measured record. Neither of these studies found time to be a significant source of variation in diets, and time-sampling variance accounted for less than one percent of the total measurement variance in nutrient contents.* It should be emphasized that the findings of minimal time-sampling variation may not generalize to current eating habits. With the growth of convenience food products, fast food service, and the increase in the number of meals eaten away from home, current diets may show greater daily and weekly variation than did diets 10 and 20 years ago.

The analyses of variance presented by Young et al. and Hankin et al. may be used to estimate response-error variance in a daily report. Based on Hankin et al. (1967) data, response-error variance accounted for between 50 and 70 percent of total variance in calorie, protein, fat, and carbohydrate measures. The Young et al. (1953) data suggest quite similar findings; error variance was about 50 percent of measured variance for both protein and calorie data.† The analysis of the daily reports from the two diet records tends to

*Young et al. found a significant person-by-day interaction for 2 of 10 nutrients. We have treated this as error variance in our calculations.

†These results are estimated from mean squared errors reported in Hankin and Young and using formulas for the expected value of the mean squared error in Cronbach et al. (1972). A study by Balogh et al. (1971) also using an analysis of variance, reports quite different findings; they conclude that error variance is only 10 percent of measured variance; however, they do not present their calculations in detail.
confirm the earlier conclusion based on the comparison of recall and weighed records: self-report measures of one-day food intake contain 50 percent or more error variance.

In summary, 24-hour recall measures of food consumption appear to contain large amounts of response-error variance that will decrease the precision with which parameters are estimated and attenuate estimated relationships among diet habits and other characteristics. One approach to reducing the response-error variance is to take several repeat 24-hour recalls and use the average quantities reported on the repeated interviews. For example, if response errors across repetitions are independent, the average of two measures would contain only one-half the error variance in a single measure. Ohlson et al. (1950) reported correlations between the average nutrient value of three 24-hour recall interviews and a ten-day weighed record. The correlation for calories and protein were 0.78 and 0.66 respectively. The sample was 13 women between ages 50 and 77. The Ohlson findings might be compared to the results of Madden et al. (1976) of 0.28 and 0.35 for a similar age group reported above. (Recall, however, that we suspect the Madden et al. finding to contain a large amount of weighed record error.)*

However, less expensive solutions may be possible, such as extending the recall period or using a diet record or diary. An evaluation of these procedures follows. Do the alternative methods obtain smaller error variances and, if yes, is there an increase in bias?

Seven-Day Recall Interview

The seven-day recall interview has two potential advantages over the 24-hour recall method: (1) time-sampling errors and biases may be smaller and (2) response variance may be smaller in the seven-day average than in a report of a single day's diet. However, memory biases and memory-error variance may increase as a result of the longer recall period.

*In interpreting these results it must be remembered that errors in the weighed record are also reduced by averaging. The difference between the Ohlson and Madden correlations may result from a reduction in both recall-error variance and weighing-error variance in the Ohlson procedure.
Two studies provide some evidence about bias in the seven-day recall interview. Hankin et al. (1975) compared food consumption in a seven-day recall interview with measures obtained from a seven-day record maintained for the same week. The subjects were 53 males. They presented differences in amounts consumed in 33 food groups as measured by the two methods. In 23 of the food groups the recall yielded less than the record. The average discrepancy across all food groups was an 18 percent underreporting in the interview.

However, Adelson (1960) found no consistent differences between the seven-day recall of food intake by 59 men and a seven-day weighed record kept for the subsequent week. The recalled amounts consumed were higher than the weighed record in five of eight food groups and less in the remaining three. Differences were less than 10 percent. Caloric value of the recalled diet was 2 percent less than the weighed record, protein 2 percent higher in the interview than the weighed results, and fat 4 percent less. In the Adelson study, wives participated in the recall interview. Adelson considered this to be an important aspect of accurate recall. In the Hankin study, wives did not assist in the recall and their absence could be a factor in the underreporting results.

There is no information in the literature to assess whether average diet measured in a seven-day recall interview contains less error variance (time-sampling and response-error variance) than the single diet obtained from a 24-hour measure.

**Seven-Day Diet Record**

A seven-day diet record (reporting in household measures) has the advantages of reducing time-sampling bias and variance and, in addition, reduces reliance on memory. However, the process of recording food intake may alter eating habits and hence not represent usual behavior. Further, it may be more difficult to achieve high levels of cooperation using a diary panel than could be obtained from a single or even repeat interviews.

As noted in the discussion about 24-hour diet recall, the findings of two studies do suggest that participation in recording a
weighed diet for seven days affected diet patterns. However, this result does not necessarily generalize to a diary in which food intake is entered in estimated household units so that no special effort is required at meal preparation time.

Young et al. (1952) provide information about the size of the Hawthorne effect in using the diary procedure. They compared information obtained from a seven-day diet record with data collection in a 24-hour diet recall interview. The recall interview covered a day before the seven-day period during which the record was kept. Subjects were 166 men and women age 12-34. There were no detectable differences in the nutrient values of the diets obtained under the alternative method.*

It is possible that the measured record would affect eating habits only if the subject is also the person who prepared the meals. In the Young et al. (1952) study, 28 subjects were pregnant females aged 16-34, whereas the remaining subjects were high school and college students. The responses of the 28 female subjects also did not evidence significant differences in the two methods.

There is very little information about response rates in the literature on seven-day diet measurement. Marr (1971) presents information on cooperation rates in five record studies, ranging from 45 to 95 percent. However, these figures indicate cooperation among subjects who agreed to participate in the study. Burk and Pao (1976) report that among 55 industrial workers willing to be interviewed, only nine cooperated by keeping seven-day records.

The Diet History

The diet history is the third major approach used in collecting information about diet habits. Are its measurement characteristics

---

*Trulson (1954) also compared responses for a seven-day diet diary to responses obtained from recall interviews with children age 7-12. No consistent differences were found; however, it is not reported whether the time periods covered in the diet record and interview overlap, so this may not be a true test of the Hawthorne effect.
more favorable than those of other dietary assessment methods? The conclusion reached is that they are not. Table 5 presents a summary of studies which have compared information obtained from the diet history with a variety of other methods. There are large differences between the diet history method and the other, relatively bias-free measures. Further, although most of the studies obtain substantially higher nutrient values in diets measured using the history measures, in three of the studies the diet history yielded lower nutrient values than the criterion. It is possible that differences in findings between studies are due to differences in characteristics of the persons studied. However, Young et al. (1952) feel that the diet history is subject to large interviewer effects which explain differences in evaluation results from study to study.

Because of potentially large biases, because it requires highly specialized nutritionists as interviewers, and because it is a long (1 hour or more) interview, this measurement procedure should not be considered for evaluating diabetic compliance.

The Balogh et al. Dietary Questionnaire

The dietary questionnaire developed by Balogh et al., which was described earlier, has received only limited testing on a specialized population group. Thus, it requires additional evaluation before more general use is recommended. However, the method is mentioned here because preliminary testing of the questionnaire has indicated that it is quite valid and reliable; administration time requires only about 15 minutes and may not require interviewers with training in nutrition.

Two tests of data obtained from the questionnaire have been reported; both tests used male civil service employees who were taking part in the Israel Ischemic Heart Disease Project.

Fourteen men were administered the short questionnaire followed by a weighed record four months later (Balogh et al. 1968). Seventy-one men responded to the diet questionnaire and also participated in eight or more 24-hour recall interviews (Balogh et al. 1971). Calorie and nutrient values obtained for the Balogh dietary questionnaire
### Table 5
PERCENTAGE DIFFERENCE BETWEEN NUTRIENT VALUES OBTAINED FROM DIETARY HISTORY AND CRITERION MEASURES

<table>
<thead>
<tr>
<th>Study</th>
<th>Observations</th>
<th>Criterion</th>
<th>Calories</th>
<th>Protein</th>
<th>Fat</th>
<th>Carbohydrates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trulson (1954)</td>
<td>47 children</td>
<td>Average of 3 or more 24-hour recalls</td>
<td>--</td>
<td>6</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Trulson (1954)</td>
<td>47 children</td>
<td>7-day record</td>
<td>--</td>
<td>1</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Young et al. (1952)</td>
<td>63 children</td>
<td>7-day record</td>
<td>28</td>
<td>22</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Young et al. (1952)</td>
<td>104 children</td>
<td>7-day record</td>
<td>25</td>
<td>27</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Young et al. (1952)</td>
<td>77 adolescents</td>
<td>7-day record</td>
<td>-9</td>
<td>-3</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Young et al. (1952)</td>
<td>68 young adults</td>
<td>7-day record</td>
<td>-7</td>
<td>1</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Young et al. (1952)</td>
<td>49 women</td>
<td>7-day record</td>
<td>10</td>
<td>20</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Young et al. (1952)</td>
<td>129 men</td>
<td>7-day record</td>
<td>12</td>
<td>8</td>
<td>21</td>
<td>--</td>
</tr>
<tr>
<td>Huenemann and Turner (1942)</td>
<td>21 children</td>
<td>10-14-day record</td>
<td>11</td>
<td>5</td>
<td>21</td>
<td>4</td>
</tr>
<tr>
<td>Van den Berg and Mayer (1954)</td>
<td>35 women</td>
<td>1-day record</td>
<td>33</td>
<td>17</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td>Balogh et al. (1968)</td>
<td>14 men</td>
<td>7-day weighed</td>
<td>-9</td>
<td>-12</td>
<td>+13</td>
<td>--</td>
</tr>
</tbody>
</table>
were consistently less than the weighed diet or recalled diet; differences, however, were generally less than 10 percent. Correlations between the dietary questionnaire and the weighed record were 0.75 or greater for five nutrient measures and correlations of 11 nutrient values between the recalled diet and measures from the diet questionnaires exceeded 0.55. It should be noted that the correlations between the short dietary questionnaire and the seven-day weighed record are of the same magnitude as found by Ohlson in comparing the average diet measured in three repeat 24-hour recall interviews with a weighed record. It is tempting to conclude from this that the Balogh diet questionnaire contains considerably less measurement error variance than other methods; however, evaluation of the responses to the questionnaire from other population groups are necessary before such generalizations can be made with confidence.

CONCLUSIONS

With the exception of the diet history, existing measures of food intake appear to contain minimal bias. The evidence on bias in seven-day recall interviews is, however, mixed. The participation of wives (who are likely to prepare meals) may reduce the underreporting bias. This hypothesis should be further tested.

Recalled measures about a single day's diet contain substantial amounts of error variance, 50 percent or more. Repeated measurement is one technique for improving measurement reliability but it is a costly solution. Questioning respondents about a longer period of time may provide more reliable information about a typical day's diet than a single 24-hour recall interview; however, we were unable to ascertain the improvement, if any, from existing literature. The diet questionnaire developed by Balogh et al. has not received enough testing to make confident conclusions about the technique; preliminary evaluations, however, suggest that it may yield more reliable measures than other existing measures.

Several important questions relating to the design of data collection about adherence to a prescribed diet among diabetics were not addressed in the literature. No discussion was found about special considerations or potential biases in collected diet data from
diabetics or others on prescribed dietary regimens (for example, reporting food intakes that have been prescribed rather than what was actually eaten). Secondly, nutritionists have been used as interviewers in most of the evaluation studies. Whether or not the use of less specialized personnel would influence data quality is unknown. Third, with the exception of the diary technique, there has been only one attempt at collecting diet information in a self-administered questionnaire.* The possibilities of the self-administered questionnaire need to be investigated further. These issues comprise an agenda for future research in methods of collecting information about diabetic compliance with prescribed diet practices.

Finally, measures have not been developed for several behavior variables rated important by the expert panel. Three omissions concern the special diet planning and preparation behavior of insulin-dependent diabetics: whether he has and uses a food exchange list, whether food portions are measured or judged accurately, and whether the day’s intake calculations include any alcohol taken by the patient. The existing literature also omits any mention of measures of whether the insulin-dependent diabetic adjusts the timing of his eating to prevent development of hypoglycemia.

---

*Browe et al. (1966) compared responses of 29 men to a self-administered questionnaire with the responses to a dietary history. Nutrient values estimated from the self-administered questionnaire were consistently lower than those from the dietary history. However, the diet history cannot be considered an unbiased criterion, so little knowledge can be gained from the Browe et al. study.
IV. ORAL MEDICATION

Two categories of oral medication behavior were rated important in phase one of the project: (a) Takes appropriate dosage of oral medication and (b) obtains (fills) prescription for oral medication. The quality of existing measures of these items is assessed in this section.

Assessment of the quality of self-reports and physician ratings as measures of compliance includes estimates of bias and estimates of the relationship between target measures and criterion measures based on urine testing or tablet counts.

The literature analysis strongly supports the conclusion of a significant overreporting bias in self-reports of taking medication. Measurement error variances tend to be moderately large or very large based on low to moderate correlations between the self-reports and the criterion measures. There is almost no evidence of a main-effect bias in the physician ratings and a strong suggestion that all of the measured rating variance is error variance. Physician ratings appear to reflect only their knowledge of average compliance rates, not the pill-taking behavior of individual patients.

The available studies relevant to obtaining prescriptions for oral medication provide sufficient information to assess the quality of self-reports as a measure of prescription-filling behavior. They suggest that self-reports are not perfect.

TAKES APPROPRIATE DOSAGE OF ORAL MEDICATION

This literature review concentrates on criterion validity studies. The target measures of compliance with instructions to take oral medication are physician ratings and patient self-reports. The question being addressed is how well do these target measures reflect actual compliance. In these studies, actual compliance is defined by criterion measures obtained from urine testing and tablet counts.
Criterion Measures

The physician rating and the self-report measures are evaluated with respect to the criterion measures used. Therefore, the evaluations necessarily depend in part on the goodness of these criteria for oral medication taking.

Urine Testing. There is general consensus in the literature reviewed that urine testing is objective, reliable, and produces rare (if any) false positives or false negatives (Charney et al., 1967; Preston and Miller, 1964; Hecht, 1974; Gordis et al., 1969). Some limitations of this procedure as an indicator of oral medication behavior have been noted. In general, the urine test result reflects medication taken only in the 12 to 36 hours before the test (Moulding et al., 1970). The procedure indicates the presence or absence of the drug (or a tracer element) in the urine but fails to provide measurement of the consumption level (Boyd et al., 1974).

A particular urine test may be an imperfect measure of a patient's general pill-taking behavior because the test day is only a sample of all days on which the patient is instructed to take the pills. The day (time) sampling errors may be of two kinds, sampling bias and sampling error variance. There is a time-sampling bias, for example, if the patient takes a different amount of medication on urine test days than on other days. There is time-sampling error variance that affects estimation precision, if there is variation from day-to-day in pill-taking behavior. As mentioned in an earlier section, the error variance can be composed of both bias variance and random-sampling error variance.

Tablet Counts. Tablet count measures may provide more information about the patient's usual drug consumption than do urine tests because they can reflect pill-taking behavior over longer periods of time.

Tablet counts have a major limitation: missing tablets may not have been ingested. Moulding et al. (1970) have developed a medication monitor, a dispenser equipped with photographic film to record regularity of tablet removal from the dispenser. This monitor is useful in detecting irregularities such as removal of several doses
at one time or failure to remove the daily doses, but the measurement procedure still requires the assumption that removed medication has been ingested and requires interpretation of failures to remove doses (e.g., did patient remove a small supply to take regularly during vacation?).

**Self-Reports**

Three published studies compare self-reports with urine analysis (Charney et al., 1967; Gordis et al., 1979; Preston and Miller, 1964). Tablet count is the criterion measure in two studies (Rickels and Briscoe, 1970; Arnhold et al., 1970). Hecht (1974) uses a criterion that involves urine analysis and/or tablet count.

Self-reports were obtained during interviews in all of these studies. The interviews were conducted during regular clinic visits for Preston and Miller (1964), Rickels and Briscoe (1970), and Gordis et al. (1969). Home interviews, conducted by representatives of the medical care facility, were used by Arnhold et al. (1970), Charney et al. (1967), and Hecht (1974).

These studies represent a variety of patient populations and they each use a different definition in categorizing patients as compliers or noncompliers. In the Charney et al. (1967) study, the patients were children for whom oral penicillin was prescribed by seven private practice pediatricians. The self-report compliance group was composed of those children whose mothers claimed that their child had used at least four of the five ounces of penicillin prescribed. Children with rheumatic fever histories on daily oral penicillin were the clinic outpatients studied by Gordis et al. (1969). Compliers were those who reported in 75 percent or more of their clinic interviews that they had taken their penicillin that day. The Preston and Miller (1964) patients were being treated with tuberculosis drugs on an outpatient basis. A physician asked the patients when they had last taken their medication; those reporting having taken their medication in the last one to three hours before the urine sample was collected were defined as self-reported compliers. Hecht (1974) also
studied adult outpatients receiving tuberculosis chemotherapy; those patients who admitted making any medication errors were considered noncompliant on the self-report measure. Based on verbal reports solicited by the physician, neurotic outpatients taking psychotropic drugs in the Rickels and Briscoe (1970) study were placed in one of three categories of compliance: no dosage deviation, 50 percent or less deviation, more than 50 percent dosage deviation. In the Arnhold et al. (1970) study, mothers were asked if they forgot to give their children antibiotics prescribed by pediatricians in a prepaid group practice. The three categories of compliance used were: never forget, occasionally forget, and often forget.

Estimation of Measurement Quality. The data obtained in each of the six studies described above were organized into a 2 x 2 (self-report x criterion measure) contingency table, as shown in Fig. 2.

<table>
<thead>
<tr>
<th>Criterion measure indicates compliance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target measure (self-report or physician rating)</td>
</tr>
<tr>
<td>indicates compliance.</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

(a + b) = p'
(a + c) = p

Fig. 2—Contingency table. Cell entries represent proportions.

The rows of the table reflect the compliance rate based on the target measure (self-report or physician rating). The columns represent the compliance rate derived from the criterion measure (urine analysis or tablet count). Bias effects were evaluated by testing the obtained p' – p difference against the hypothesis that p' = p (McNemar, 1969). A significant difference in the positive direction signifies an overreporting bias; a significant negative difference indicates an underreporting bias. Cells b and c contain the proportion of self-reports that were in disagreement with the criterion measure. In addition to the assessment of bias, the correlation (phi coefficient) between the predictive and criterion measures was computed to provide an indication of the possible range of measurement error variance in the target measure.
**Results.** Table 6 presents the findings of the studies of self-reporting that have been reviewed. The first column presents the sample size in each of the studies. The next two columns show the self-report compliance rate and the criterion measure compliance rate. The difference between the column values (not shown) is an estimate of the bias in the self-reports. The statistical significance of the difference is tested and the statistic is given in the column headed Test of Bias. These results strongly support the conclusion of a significant overreporting bias in self-reports.

The next to last column in Table 6 shows the correlation (phi coefficient) between the patient (or parent) and the criterion measure. Four of the six studies* find correlations of about 0.5; the others are lower. If the target and criterion data have equal error variances, then about half of the measured variance in each is error variance. A more conservative approach is to assume that all the error variance is in the self-reports (namely, the criterion data are perfect). Under the conservative assumption, the percent of measured variance is \((1 - \phi^2)\). If \(\phi\) is 0.5, then the upper limit of percentage of error variance in the self-reports is 75 percent. We have been unable to find studies in the literature (e.g., test-retest estimates of random measurement error variance) that would help to refine our estimates further.

The amount of bias in these self-reports would argue against their use as indicators of compliance. If costs are not an obstacle, a tracer element could be added to diabetic oral medication and urine tests for it used to assess short-term compliance. However, this type of measure is likely to contain a Hawthorne effect bias: compliance may be better just before the patient knows he will be tested than at other times. Without the bias, the measure is still subject to time-sampling variation. On the other hand, pill counts may reduce time-sampling effects but still require frequent visits by the

*Five out of nine correlations.
<table>
<thead>
<tr>
<th>Study</th>
<th>No. of Patients</th>
<th>Proportion of Compliers (p' = (a + b))</th>
<th>Proportion of Compliers Measurement (p = a + c)</th>
<th>Test of Bias (Z-score)</th>
<th>Correlation Between Target and Criterion Measures (Phi Coefficient)</th>
<th>Percent Disagreement (b + c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-report by proxy (mother) vs urine analysis (Charney et al., 1967)</td>
<td>26</td>
<td>.85</td>
<td>.62</td>
<td>*(a)</td>
<td>0.54**</td>
<td>23%</td>
</tr>
<tr>
<td>Self-report vs urine analysis (Gordis et al., 1969)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>o Proxy (mother)</td>
<td>43</td>
<td>.73</td>
<td>.42</td>
<td>3.47**</td>
<td>0.52**</td>
<td>31%</td>
</tr>
<tr>
<td>o Child (self-report)</td>
<td>103</td>
<td>.69</td>
<td>.33</td>
<td>6.08**</td>
<td>0.46**</td>
<td>36%</td>
</tr>
<tr>
<td>Self-report vs urine analysis (Preston and Miller, 1964)</td>
<td>25</td>
<td>.96</td>
<td>.72</td>
<td>*(a)</td>
<td>0.33</td>
<td>24%</td>
</tr>
<tr>
<td>Self-report vs urine analysis and/or pill count (Hecht, 1974)</td>
<td>47</td>
<td>.79</td>
<td>.51</td>
<td>3.18**</td>
<td>0.55**</td>
<td>26%</td>
</tr>
<tr>
<td>Self-report vs tablet count (Rickels and Briscoe, 1970);</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>o Compliance = no deviation</td>
<td>301</td>
<td>.75</td>
<td>.22</td>
<td>12.26**</td>
<td>0.25**</td>
<td>55%</td>
</tr>
<tr>
<td>o Compliance = &lt;50% deviation</td>
<td>301</td>
<td>.96</td>
<td>.88</td>
<td>4.60</td>
<td>0.51**</td>
<td>9%</td>
</tr>
<tr>
<td>o Never missed</td>
<td>66</td>
<td>.27</td>
<td>.77</td>
<td>-5.03**</td>
<td>-0.10</td>
<td>66%</td>
</tr>
<tr>
<td>o Never or occasionally missed</td>
<td>66</td>
<td>.92</td>
<td>.77</td>
<td>2.12**</td>
<td>0.01</td>
<td>27%</td>
</tr>
</tbody>
</table>

*p < .05.

**p < .01.

*Computed probability directly using binomial expansion.

**Computed using normal curve approximation of the binomial expansion with correction for continuity.

P*Computed using normal curve approximation of binomial expansion.
patient to the clinic so that the count data may be obtained. This lengthens the research study period (increasing costs) and introduces problems of missed appointments and attrition; the former may be correlated with true compliance (e.g., those who have not complied may be more likely to miss appointments or otherwise decide not to continue their participation in the research). A patient who does not want to reveal his noncompliance can easily remove pills from the bottle before they are counted.

The above suggests that better self-report measures should be developed, specifically; measures that minimize the overreporting bias and are inexpensive to use. We do not know whether such measures can be developed. Our recommendation is to make the attempt based on the hypothesis that the patient overreporting bias is due to the "demand" characteristics of the data collection environment. The self-reports mentioned above were obtained from patients (or parents) by medical personnel. Patients may have special reasons for not wanting to admit noncompliance in this situation.* By changing the context of the data collection situation and assuring the patient (parent) that his answers will not be given to his health care providers and that truthful answers are needed for good medical research, the quality of new measures may be improved.

Definitions of Compliance. The definitions used to categorize patients as compliant or noncompliant can have a major influence on the results of any study. These definitions vary greatly from study to study. The compliance classifications used for self-reports are described above. For tablet count, as an additional example, compliance has been variously defined as taking 90 percent or more of the prescribed medication, as taking 66 percent or more of the prescribed medication, no more than one pill in error, no errors, etc. For multiple urine tests, compliance definitions may differ

*Note that in the measurements evaluated elsewhere in this appendix, an overreporting response bias is not observed. These respondents are probably not motivated to lie either because they are not supposed to be following a treatment regimen or because the data are collected by an impartial agency that will not reveal them to health care providers.
in the proportion of positive tests required. This issue presents limitations for across-studies comparisons. The definition of compliance can influence the results of a study and a change in the definition can reverse the conclusions drawn from a study. This is evidenced by the last study presented in Table 6, Arnhold et al. (1970). Note that when compliance is defined as never missing a dosage, only 27 percent of the sample reported themselves as compliers. But when compliance is broadened to include occasionally missing a dosage, 92 percent of the sample become self-reported compliers. In the first instance, a significant underreporting rate is obtained; and in the second, a significant overreporting rate is shown. This may indicate either that the question is ambiguous (respondents interpret "occasionally" differently) or that the pill count is insensitive to occasionally missed doses. Table 6 also indicates that a change of compliance definition for the Rickels and Briscoe (1970) study influenced the magnitude of the overreporting bias, but in this case the conclusions regarding the direction of the bias were not changed.

Physician Ratings

An alternative to patient self-report medication measures is the physician's rating of patient compliance. The quality of physician rating measures is examined here.

Two of the studies using physician ratings employ tablet counts as the criterion measure (Mushlin and Appel, 1977; Moulding et al., 1970), and two studies use urine tests to measure "actual" compliance (Charney et al., 1967; Preston and Miller, 1964).

Mushlin and Appel (1977) and Moulding et al. (1970) had physicians predict the percentage of the amount of prescribed medication that the patient would actually take. The estimated complier category was defined to contain those patients predicted to take 90 percent or more of their medication. Both studies involved adult outpatients. The Moulding et al. (1970) patients were being treated with tuberculosis drugs; the Mushlin and Appel (1977) study involved followups on discharged hospital patients with various illnesses and medications.
The physicians participating in the Charney et al. (1967) and the Preston and Miller (1964) studies were asked to place patients in one of two categories based on whether or not the patient was judged to be taking medication as prescribed. Charney's patients were children for whom oral penicillin was prescribed. Preston and Miller studied tuberculosis outpatients.

The diversity of methods used to obtain physician ratings, definitions of compliance, and patient populations makes across-studies comparisons difficult. However, the conclusions that do emerge from an examination of the findings of these studies are presented in Table 7. Except for the residents' data in the Mushlin and Appel (1977) study, there is no evidence for a main effect bias in the physician ratings.

Even though the doctor may correctly estimate the average compliance rate, his prediction for specific individuals contains no additional "information." When comparing physician estimates with the patient's performance (Table 7), the correlations are insufficient to predict the performance of individual patients. The last column in Table 7 indicates, generally, over one-third of the judgments are in error.

In support of the above, Caron and Roth (1968) studied hospital ward patients on an antacid regimen for peptic ulcer. They used a line-mark procedure to obtain physician estimates; the ends of the line were marked 0 percent and 100 percent and the physician placed a mark on the line to indicate his prediction of medication consumption. The criterion measure used was a tablet count. Twenty-seven physicians rated an average of 20 patients each. The median of the 27 correlations between physician rating and tablet count was -0.01. They reported an overreporting bias in physician ratings and concluded from the correlations an inability to discriminate good from poor compliers.
<table>
<thead>
<tr>
<th>Study</th>
<th>No. of Patients</th>
<th>Proportion of Compliers (p* = (a + b))</th>
<th>Proportion of Compliers Criterion Measurement (p = a + c)</th>
<th>Test of Bias (Z-score)</th>
<th>Correlation Between Target and Criterion Measures (Phi Coefficient)</th>
<th>Percent Disagreement (b + c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physician rating vs tablet count (Mushlin and Appel, 1977)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interns</td>
<td>73</td>
<td>.68</td>
<td>.73</td>
<td>-0.54&lt;sup&gt;a&lt;/sup&gt;</td>
<td>-0.02</td>
<td>43%</td>
</tr>
<tr>
<td>Residents</td>
<td>73</td>
<td>.42</td>
<td>.73</td>
<td>-3.67&lt;sup&gt;*a&lt;/sup&gt;</td>
<td>0.09</td>
<td>49%</td>
</tr>
<tr>
<td>Physician rating vs tablet count (Moulding et al., 1970)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>102</td>
<td>.76</td>
<td>.69</td>
<td>1.37&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.17</td>
<td>33%</td>
</tr>
<tr>
<td>Physician rating vs urine analysis (Charney et al., 1967)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>153</td>
<td>.68</td>
<td>.64</td>
<td>0.90&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.11</td>
<td>40%</td>
</tr>
<tr>
<td>Physician rating vs urine analysis (Preston and Miller, 1964)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>.80</td>
<td>.72</td>
<td>Not significant</td>
<td>0.58&lt;sup&gt;*&lt;/sup&gt;</td>
<td>16%</td>
</tr>
</tbody>
</table>

* p < .01.
<sup>a</sup> Computed using normal curve approximation of the binomial expansion.
<sup>b</sup> Computed probability directly using binomial expansion.
A possible conclusion from the above group of studies is that physician ratings reflect their knowledge only of average compliance rates (possibly from reading other research results). If true, physician ratings may be insensitive to treatment group differences and should not be used in certain kinds of evaluation designs.

The psychological literature reflects a similar problem of judgmental accuracy; also, in this literature there is little empirical support for a relationship between ability to predict individual behavior and the rater's training, experience, or confidence (Wiggins, 1973; Brody, 1972). The compliance literature suggests also that physician estimates do not improve with greater clinical experience (Caron and Roth, 1968; Mushlin and Appel, 1977; Davis, 1966). Caron and Roth (1968) found nonsignificant correlations between accuracy and a number of different variables including number of patients judged and amount of medication prescribed. Confidence was also investigated by Caron and Roth, and when judgments were considered for only the cases in which the physicians indicated greater confidence, the predictions were still no more accurate than chance selections.

Of the studies reviewed, only Caron and Roth (1968) assessed test-retest reliability. Nine physicians provided two sets of ratings about two weeks apart. The median correlation was reported at 0.65. This suggests a fair degree of consistency in a physician's rating of a patient over time (i.e., the ratings are not completely random) but the consistency is unrelated to the criterion definition of compliance.

It seems reasonable to conclude that the physician ratings do not predict the criterion measures of compliance—urine tests and tablet counts. If ratings consist mostly of random error or are based on characteristics other than compliance and these

---

*For example, the physician may base his rating on personal, psychological, demographic, or medical characteristics of the patient and these characteristics may be unrelated to criterion-defined compliance. A different possibility is that the test-retest correlation underestimates random measurement variance in the first measure because physicians partially recall their original rating and furnish similar ratings on the retest to appear consistent.
characteristics vary across treatments or programs, physician ratings should not be used in comparative program or treatment assessments.

OBTAINS PRESCRIPTIONS FOR ORAL MEDICATION

Only two studies were found that evaluate the quality of measures of obtaining prescribed medications. Parry et al. (1971) compare self-reports with prescription records. The question that they address is: "Does a respondent who is known to have filled one or more prescriptions for sedatives, stimulants, tranquilizers, or antibiotics affirm or deny that fact?" Referring to the contingency table in Fig. 2, this study only provides entries for the cells labelled a and c. The frequency of cell c is greater than zero, * indicating that the self-reports and pharmacy records sometimes disagree. It is impossible to tell whether this indicates bias or error variance or its probable source (self-report, record, or something else).

Hammel et al. (1961) compare self-reports and pharmacy records. They interview patients known to have not filled their prescriptions when local pharmacy records indicated otherwise. The number of persons filling prescriptions is not mentioned but is presumed to be over 500. † The compliance reporting bias, therefore, is very small and possibly nonexistent if records are in error. (There was at least one case of record error).

Full-design criterion validity studies will be needed to obtain the desired estimates of the measurement error parameters. Neither of the studies cited above provides sufficient information to assess the quality of self-report measures of prescription-filling behavior.

* The proportion of apparent respondent denials ranges from 15 percent for tranquilizers to 36 percent for antibiotics.
† The total number of prescriptions in the research was 745.
V. OBTAINING MEDICAL CARE

The goal of this section is to evaluate published measures of obtaining medical care. The Delphi panel rated a number of behavior items as important or very important for diabetics. These are (1) contacts the physician for new (or persistent) symptoms, poor control, and secondary consequences; (2) visits an ophthalmologist as prescribed; (3) visits the physician as prescribed; (4) visits a podiatrist as prescribed; (5) visits a dietician/nutritionist as prescribed; (6) visits a dentist as prescribed; (7) informs all providers of diabetic condition.

Previous research has not used measures of most of these medical care behavior categories. Only one class of measure, self-reports of physician contact behavior, can be evaluated for quality. This evaluation is the primary focus here. Unfortunately, again, none of the cited research sampled diabetic populations.

The predominant impression among health care researchers is that self-reports of visiting a physician contain an underreporting bias, possibly resulting from forgetting. The studies supporting this impression are based on incomplete criterion validity designs. We have found another group of studies that use a different version of an incomplete criterion validity design and find an overreporting bias in self-reports. A third group of studies obtained estimates of overreporting and underreporting for the same subjects. One found the rates to be roughly comparable in magnitude and the author of the other encountered a major problem in the criterion data. These studies are not definitive but suggest that what others have interpreted as a self-reporting bias may be partly or entirely random-measurement error variance. This hypothesis is supported by the results of a final group of studies that use complete criterion validity designs. Self-reports of whether or not the patient has seen a physician during a specified time period appear unbiased in these studies. Estimated error variance is moderate to high, suggesting that use of these
measures in small-sample research studies will present problems of estimation precision. Reports of visits to dentists and specialists are discussed briefly at the end of the section.

DOCTOR VISITS

Measurement evaluations of self-reports of contacts with the medical care system are available only in the survey research literature. The common assumption among survey researchers is that self-reports are negatively biased; that is, not all contacts are reported.* The evidence underlying the underreporting assumption is reviewed next.

Underreports

Loewenstein (1969) drew a sample of 311 names of persons known to have received care at health departments during a seven-month period and was able to interview 260 (84 percent) of them. Interview questions were designed to elicit the number of visits per sample person to each source of outpatient medical care. For this analysis, the patient data were scored according to whether any visit to the health department clinics during the seven-month period was reported.

Balamuth et al. (1965) drew a stratified random sample of members of a prepaid insurance plan in New York. Interviews, conducted in 1388 households, asked about doctor visits for 3937 sample persons. The interview information about visits was compared to information in the records of the prepaid plan's facilities. Comparisons were made of the number of visits during the two-week period prior to the interview and whether or not the sample person received outpatient treatment at a clinic during the 12 months before the interview.

Cannell and Fowler (1963) selected a sample of outpatient visits from records of a prepaid health plan in Detroit, Michigan. Interviews

*For example, Cannell states "...health events are more likely to be underreported than overreported." (Cannell et al., 1977).
were conducted with about 586 persons to obtain a report of the number of doctor visits made by each person during the two-week period before the interview. The interview and record information were compared on a visit-by-visit basis.

All three studies accepted proxy responses for children and for adults not present when the interviewer called. Each study reports an estimate of survey underreporting bias defined as the percent of known visits not reported in the survey:

\[
\text{Percent survey underreporting} = \frac{\text{Number of survey-reported persons or visits not matched to record reports}}{\text{Total number of record reports of visits or persons}}
\]

The underreport estimates obtained from the three studies are shown in Table 8. They range from 19 percent to 52 percent. It is on the basis of findings such as these that researchers suspect self-report data about contacts with the medical system contain a moderately large, negative response bias.

Referring to the table in Section II, it can be seen that these studies obtain good estimates only of the a and c cells. Other studies were found that provide estimates of cells a and b. These are discussed next.

Overreporting

Andersen et al. (1975) report the results of a record check of doctor visits reported by respondents in a national, cross-section sample. An interview inquiring about use of doctor services in the preceding year was administered to 3880 families comprising 11,022 individuals. If a doctor visit was reported to have occurred, physicians were contacted in an attempt to verify that the utilization actually took place. Record and interview information were compared. On the basis of the record check data, they present an adjusted estimate of per capita physician visits which is 13 percent lower than the survey
Table 8
SURVEY UNDERREPORTING IN THREE STUDIES

<table>
<thead>
<tr>
<th>Study</th>
<th>Length of Reporting Period</th>
<th>Information Sought</th>
<th>Percent Survey Underreport</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loewenstein (1969)</td>
<td>12 months</td>
<td>Did sample person make any visit to selected health department clinics?</td>
<td>52%</td>
</tr>
<tr>
<td>Balamuth et al. (1965)</td>
<td>2 weeks</td>
<td>Any outpatient visit to a prepaid plan clinic?</td>
<td>36%</td>
</tr>
<tr>
<td></td>
<td>12 months</td>
<td>Any outpatient visit to a prepaid plan clinic?</td>
<td>19%</td>
</tr>
<tr>
<td>Cannell and Fowler (1963)</td>
<td>2 weeks</td>
<td>Number of outpatient visits to a prepaid plan clinic.</td>
<td>23%</td>
</tr>
</tbody>
</table>

report. This suggests that at least 13 percent of survey-reported visits were not confirmed by the record.

A study by Loewenstein (1969) used a record check design similar to Andersen's. An initial interview was administered to 1540 families in the Washington Heights Health Center District of New York City. Respondents were asked to report doctors whom they had consulted during the previous year. Mentioned doctors then were mailed a questionnaire asking about treatment provided to the individual. *

*Questionnaires were mailed to 43 percent of the doctor-person pairs identified in the interview. Reasons for not attempting to contact the doctor included respondent refusal to sign permission form, doctor not in Manhattan, and a decision to send a maximum of six names to each doctor. Sixty-five percent of the mailed questionnaires were completed. Thus, verification was completed on 28 percent of the person-provider pairs identified in the interview.
In 14 percent of the returned questionnaires, the provider indicated that the person had not been treated during the year. Loewenstein felt that this overstates the overreporting because some questionnaires may have been sent to the wrong doctor (because of incomplete addresses, similar doctor names, etc.). Including only the cases in which the doctor reported that the individual had been a patient at some time in the past, 9 percent of the treatments (single or multiple visit) reported in the interview could not be verified in the doctors' records.

The problem with interpreting overreport rates or underreport rates as indicators of survey bias is that they may not reflect the entire range of possible errors. In terms of the 2 x 2 contingency table shown in Section II, the survey underreport rate is

\[
\text{Underreport} = \frac{c}{a + c}
\]

and the survey overreport rate is

\[
\text{Overreport} = \frac{b}{a + b}
\]

The definition of survey bias, \( \overline{B} = b - c \), requires estimates or assumptions about the values of both b and c. The underreporting studies obtain frequencies for cell c. They must assume cell b is zero (or less than c) when implying that underreporting indicates a negative response bias. Similarly, the overreport studies obtain frequencies for cell b and must make assumptions about cell c to conclude that their overreporting rates imply a net positive bias in survey reports.

There are several studies that provide estimates for cells a, b, and c. There are additional studies that empirically estimate the whole matrix. The former group is discussed next.

**Studies Estimating Cells a, b, and c**

Some criterion studies collect data sufficient to estimate a full validation matrix if the person is the unit of analysis. However, they choose to present data using the visit as the unit of analysis, precluding a meaningful estimate of cell d. Two studies of this kind are described here.
Sudman et al. (1974) drew a sample of names from the membership lists of two health maintenance organizations, one in Chicago and the other in Marshfield, Wisconsin. Reporting visits over a three-month period was by either a diary or three recall interviews. Some respondents were paid, some were given a summary report of their medical expenses during the study, and some were offered nothing. Three-fourths of the households who cooperated in the data collection phase of the study granted permission to the researchers to use their medical records for a criterion validity study. The interview and diary reports were compared to record information on a visit-by-visit basis with the following results:

<table>
<thead>
<tr>
<th>Visit reported in diary or interview?</th>
<th>Yes</th>
<th>No</th>
<th>1047</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>795</td>
<td>252</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>164</td>
<td></td>
<td>959</td>
</tr>
</tbody>
</table>

The underreporting rate, 164/959, is 17 percent; the overreporting rate, 252/1047, is 24 percent. Because the record and survey data contain different estimates of the visit rates and the frequency in cell d is unknown, it is not possible to infer net reporting bias from these data. The study demonstrates, however, that high over and underreporting rates can be obtained from the same sample if the research employs a criterion validity design capable of detecting both kinds of errors.

Feather (1972) reports the results of a criterion validity study carried out in Saskatchewan, Canada. A random sample of persons eligible for the provincial health insurance program in three areas of Saskatchewan were included in interview. Adults furnished information about themselves and their children concerning their visits to physicians during the two weeks before the interview. These reports were compared to computerized insurance records. Unfortunately, the units of analysis in the survey and in the insurance records differ. Visits are recorded in the interview but "fee submissions," which
can cover a number of visits, are in the records. The results of the visit-by-fee submission match are:

<table>
<thead>
<tr>
<th>Visit (survey)</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>328</td>
<td>274</td>
</tr>
<tr>
<td></td>
<td>52</td>
<td></td>
</tr>
</tbody>
</table>

The 14 percent underreport rate is not too dissimilar from under-report rates cited above. The survey overreport rate of 46 percent is higher than that obtained in most studies and is undoubtedly inflated by the lack of uniform single-visit reporting in the records.

Studies Estimating the Full 2 x 2 Matrix

Studies that estimate all four cells of the 2 x 2 matrix are termed full-design studies. A full design is necessary to obtain an estimate of net bias and error variance in self-reports of doctor visit utilization.

Unfortunately, an intensive search of the literature turned up only three studies that used a full design for measurement evaluation. Cartwright (1963) had general practitioners within an English suburban community to keep detailed records of their consultations during a specified time period. A sample of approximately 2000 persons were interviewed about their visits to general practitioners during a four-week reference period. Self-report and record information were compared.

The Madow study (1973) interviewed approximately 5000 enrollees in a prepaid health plan about doctor visits over the past 12 months. These responses were compared to record data abstracted from information provided by physicians.

The full-design Loewenstein (1969) study also involved records from a prepaid health plan. If, in the larger sample study, a respondent said he belonged to the prepaid health plan (and used its clinics), he was asked how many clinic visits he made during the past 12 months. Efforts were made to locate the relevant clinic
records (successful in 96 of 106 cases). The record and self-report information were then compared.

These studies were reanalyzed by scoring the data dichotomously to reflect whether or not any visit was made. The findings are summarized in Table 9 and discussed below.

Bias effects in self-reports of doctor utilization are estimated by looking at the significance of the difference between under and overreporting rates (McNemar, 1969). A negative difference between these two rates indicates an overreporting bias; a positive difference indicates a significant underreporting bias. The fourth and fifth columns of Table 9 give the percent of underreporting and overreporting, respectively; Z-score values in column 6 are tests of the significance of the difference between these two types of reports (i.e., the bias) under the assumption of independence of observations. None of the tests for bias were significant. Data from the Cartwright (1963) study produced the largest Z-score; for respondents in this study, the tendency was towards overreporting but this effect was not significant.

The amount of agreement between self-reports and medical records is indicated from the magnitude of the correlation between these two measures given in column 7 or Table 9. Correlation coefficients were significant ($p < 0.01$) for all three studies. One way to interpret the magnitude of the coefficient is in terms of goodness of prediction, often assessed in terms of variance accounted for ($\phi^2$). The variance accounted for in the medical records by self-reports is 17 percent, 40 percent, and 64 percent for the Loewenstein, Madow, and Cartwright studies, respectively. This is also reflected in the last column, which shows the percent of respondent reports which did not agree with medical records. Percent disagreement would be expected to decrease as the measure of strength of association increases, which is the case (compare columns 6 and 7).

The correlation between records and self-reports is highest in the Cartwright study. The reporting period was four months compared

* The phi coefficient was used; this reflects the product moment correlation between two dichotomously scored variables.
Table 9
OVERREPORT AND UNDERREPORT RATES, CORRELATION, AND PERCENT DISAGREEMENT IN THREE FULL-DESIGN CRITERION VALIDITY STUDIES OF SELF-REPORT PHYSICIAN VISITS

<table>
<thead>
<tr>
<th>Study</th>
<th>Reference Period</th>
<th>N</th>
<th>Percent Under-reporting</th>
<th>Percent Over-reporting</th>
<th>Test of Bias Z-scores</th>
<th>Correlation Between Target and Criterion Measures</th>
<th>Percent Disagreement (b+c)/(a+b+c+d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loewenstein</td>
<td>12 months</td>
<td>96</td>
<td>11%</td>
<td>14%</td>
<td>-0.41</td>
<td>0.41*</td>
<td>25%</td>
</tr>
<tr>
<td>(1969)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Madow</td>
<td>12 months</td>
<td>5027</td>
<td>6%</td>
<td>5%</td>
<td>0.63</td>
<td>0.63*</td>
<td>12%</td>
</tr>
<tr>
<td>(1973)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cartwright</td>
<td>4 weeks</td>
<td>2040</td>
<td>3%</td>
<td>4%</td>
<td>0.80</td>
<td>0.80*</td>
<td>7%</td>
</tr>
<tr>
<td>(1963)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NOTE: Self-report and record data have been scored dichotomously.
* Significant at the .01 level.
to the 12-month reporting period used in the other two studies. This
should not be interpreted as evidence of forgetting error on the
amount of response error variance. In the Cartwright study, the doc-
tors maintained special records for the four-week period intended
specifically for research purposes. It is likely, therefore, that
both record and match errors in the Cartwright study are lower than
in the other two studies. Random record and match errors attenuate
the correlation coefficient.

The three studies, employing appropriate criterion validity de-
signs, indicate that self-reports of visiting a physician or medical
care facility in a prespecified period of time are unbiased. The
estimates of measurement error variance vary considerably across the
three studies. The study that designed a record keeping system
specifically for research purposes suggests that the response error
variance is between 20 and 36 percent of total measured variance.

VISITS TO SPECIALISTS

The preceding discussion has addressed the quality of survey
reports about all types of doctor visits. Some of the measures we
would like to evaluate concern visits to medical specialists. The
only survey reports about visits to a particular type of provider
which have been evaluated are measures of dental visits. However,
one study examined the accuracy with which respondents are able to
identify the types of medical specialists and practitioners they con-
sult. This study along with the evaluation of the reporting of
total dentist visits provides some information about the quality of
data that would be obtained if respondents were asked about visits
to specific medical specialists. The evaluation of survey reports
of dental visits and types of medical specialists is discussed below.

Two studies have compared survey reports of the number of dental
visits with information in dentists' records (Newman and Anderson,
1972, Marquis et al., 1976). Both of these studies obtained an
initial interview with persons in households inquiring about the
names of dentists visited in the previous year and the number of
times each dentist was visited. All dentists reported as having
treated any family member during the year were contacted. The
named provider was sent a form for each individual in the family and
asked to report how many times that individual had been treated dur-
ing the year. In the Marquis et al. study, respondents were also
asked to name a "usual" dentist. The "usual" dentist was included in
the record check study even if no family member reported receiving
treatment. The criterion validity design in both of these studies
underrepresented false negative survey reports (incorrect reports of
no dental service use), because a survey report of no dental utiliza-
tion could be verified only by checking with the dentists whom the
respondents mentioned. Newman and Anderson (1972) report that no
record of treatment was found for 10 percent of the persons who re-
ported having seen a dentist in the survey. However, they do not
report how many instances of underreporting were detected. Hence,
the difference was not significant. The sign of the difference was
to be expected because not all survey underreports could be detected
in the record check design used. The correlation between the survey
reported visits and the dental records was 0.75, suggesting that
survey error variance may range between 25 and 50 percent of total
measured self-report variance. This suggests that self-reports of
visits to dental specialists are also unbiased and contain no more
error variance than reports of physician visits.

Respondents appear to know and report medical provider specialty
information fairly well. The National Health Interview survey asked
each respondent if he had visited each of 10 kinds of medical specialist
in the past 12 months and, if yes, to give the name and address
of the specialist. Hannaford (1966) compared these reports to medi-
cal roster information. Eighty-two percent of the specialists named
by respondents were found in the medical rosters. Of the matched
cases, respondents correctly reported the specialty of 88 percent.
The rate of correct identification did not vary substantially across
the 10 classifications of specialty used in the study. The design does not tell us how often a true specialist contact went unreported in the interview. The data do indicate that when the respondent reported visiting a particular kind of specialist, his classification of the doctor's specialty was correct between 72* and 88 percent of the time. This is a fairly high rate of accuracy and we suspect diabetics who make routine visits to specialists may be even more accurate. Taken together these two studies do not suggest that there will be unique measurement problems in obtaining unbiased self-reports of visits to specialists.

APPOINTMENT KEEPING BEHAVIOR

Twenty-seven reports of studies that measured appointment keeping behavior were reviewed. Only one of these articles (Donabedian and Rosenfeld, 1964) collected the data in an interview with the patient. They contacted 82 patients who had been discharged from a hospital three months previously and interviewed them about compliance with the discharging physician's recommendations for follow-up treatment. However, patients' responses were not verified with records and so the quality of the collected data cannot be evaluated.

The most frequent method used to obtain information on appointment keeping behavior is for a member of the study team to collect and abstract the physician's appointment schedule and attendance records. Examples of studies using this procedure include Alpert (1964), Becker et al. (1977), Hulka et al. (1975 a, b), Meller and Anderson (1976), Stine et al. (1968), and Weinerman et al. (1965). The two other methods of collecting data which were reported in the literature are: (1) a telephone interview with the physician inquiring about the appointment keeping behavior of patients (Mushlin and Appel, 1977) and (2) visits to the medical facility at the time of a scheduled appointment to determine if the patient kept the appointment (Mushlin and Appel, 1977; Rockart and Hofmann, 1969). None of the investigations examined the quality of the doctor records and/or the abstraction procedures.

* Assumes all unmatched cases were incorrect.
Although the quality of data collected in studies measuring appointment keeping behavior is unknown, some problems in using doctor records to measure an individual's behavior can be pointed out. To obtain an overall measure of a patient's appointment keeping behavior by examining doctor records, the relevant physician and other medical personnel have to be identified. Reviewing the records of all providers in a geographic area for each individual under study would be time-consuming and prohibitively expensive. However, if the doctors are identified by an initial interview with respondents, the records of providers not mentioned in the interview will be missed. Noncompliance with a physician's referral is likely to be missed using either approach.

In summary, most measurement of appointment keeping behavior has relied on doctor records. Although there is no evaluation of the collected data, the use of provider records poses practical problems when the desired measure is one of an individual's overall compliance.

**MEASURES OF OTHER BEHAVIOR ITEMS**

We have been unable either to identify or to evaluate published measures of two important medical care items:

- Contacts the physician for new (or persistent) symptoms, poor control and secondary consequences;
- Informs all [medical care] providers of diabetic condition.

Measures of these kinds of behavior need to be developed if a comprehensive assessment of the patient's self-management is to be made. The data quality problems likely to be encountered in constructing good quality measures in these areas are unknown.
VI. OTHER SELF-MANAGEMENT BEHAVIOR

Research on the measurement of self-care behavior in many other important areas is infrequent and, perhaps surprisingly, few attempts at formal measurement of actual behavior have been made for diabetic populations. As a result, little is known about the potential quality and practicality of measuring everyday patient management activities. Nor, at present, does a published pool of question items (or other measurement procedures) exist with which to construct measures of many elements of important self-care behavior. The behavior measurement state-of-the-art in urine testing, insulin medication, safety measures, and foot care is reviewed here.

COMPREHENSIVE STUDIES OF DIABETIC PATIENT BEHAVIOR

The published literature contains only two or three studies that attempt to assess a wide range of diabetic self-care behavior. Only a few of their measures were evaluated for quality. Many other studies focus on measures of diabetic patient knowledge, attitudes, and skills (which may be prerequisites to actual behavior), but stop short of obtaining data about what the patient actually does. A brief description of the behavioral studies is given here. Their measurement contributions are discussed in the content-oriented sections that follow.

Gabrielle and Marble (1949) obtained responses from 86 of 116 insulin-dependent diabetic boys (ages 3-16 years) on a questionnaire "filled out under the [summer camp] physician's supervision." Questions covered past behavior in the home. The items are not given and the reasons for the low response rate in the "captive" sample are not discussed.

Watkins et al. (1967a,b) interviewed 162 adult diabetic patients drawn from clinic and private practice populations. Seventy-one percent were taking insulin and the remainder took oral hypoglycemic medication. The public health nurse and medical student interviewers obtained a response rate of about 90 percent. Patients answered
questions about their knowledge and their disease-management behavior at home. Patient demonstrations of urine testing and insulin administration were obtained and scored. The questions and protocols used are unpublished.

Hulka et al. (1975a,b; 1976) obtained interviews with between 234 and 242* adult diabetic patients and (123 patients, aged 50-75, with a diagnosis of congestive heart failure). Among the diabetics, 233 were dependent upon external insulin medication, 164 used oral medication, and 56 were on prescribed treatment regimens that excluded anti-hyperglycemic medication.

Patients were interviewed in their homes. The questionnaire and other measurement procedures used are not yet published. The items to which a patient was asked to respond depended upon (a) whether the patient's physician said the patient was "instructed" in a particular area and (b) whether the patient said the physician instructed him or her in the area. Patients who passed this double screening procedure for selected areas were asked for certain demonstrations. These included showing medication, showing a diabetic identification, and demonstrating their urine testing procedure.

Eighty-four percent of the patients asked to participate agreed to do so. Varying percentages of missing data are present for questionnaire and demonstration items.

URINE TESTING

Behavioral items about urine testing judged important by the Delphi panel were:

- Accurately matches test color to chart.
- Tests urine at correct time in relation to meals.
- Interprets urine test results correctly; carries out additional ketone test(s) if urine test results so require.
- Times test reaction for correct length of time before reading result.

*There is a slight variation in reported diabetic sample sizes within and across the published reports. The remainder of this discussion relies on data published in Hulka et al. (1975b).
o Uses medicine dropper properly (if using Clinitest tablets).

o Stores urine test materials appropriately so potency is maintained.

o Records an adequate sample of urine test results accurately.

A later suggestion of one of the panel members was to add:

o Takes appropriate action on the basis of test results (e.g., varies dosage of insulin). This is similar to the item "varies dosage of insulin correctly according to urine test results," which is discussed later.

The literature refers to both self-report and demonstration measurement methods used to assess patient urine testing behavior. Unfortunately, none of the measurement procedures are specifically given nor can the quality of any of the measures be evaluated.

Stone (1961) used an unspecified method (not a questionnaire "because it seemed to evoke answers that were superficial and rigid and gave a false impression of accuracy") to obtain self-reports from adult clinic patients about their frequency of urine testing and their behavior if the test readings were greater than 14. The Gabrielle and Marble (1949) questionnaire covered frequency of urine testing at home and the color of the test reaction. The items used are not given.

There are three reports of urine testing demonstration measurement. Shenfield and Steel (1977) asked 100 patients to test their urine in the clinic after it had been tested by a nurse. They compared the patient's and nurse's test result and at least informally noted gross procedural errors made by patients. They concluded that this measurement technique would not detect many procedural errors, especially when the true sugar concentrations were zero or extremely high. In a second study they asked patients to test urine known to contain 3/4 to 1/000 ml glucose and informally noted procedural errors. The kinds of errors recorded included:
- Imperfect use of the dropper.
- Failure to time the test correctly.
- Difficulty reading color chart due to poor eyesight

Hulka et al. (1975b) asked selected patients to test a standardized dextrose solution during a household interview. The authors state that the interviewer recorded whether or not the patient performed and read the test correctly. The observation protocol, however, is not in the published report. There has been very little evaluation of the measurement quality of this particular procedure. Because the demonstration was administered only to patients whose doctors said they were "instructed" in urine testing and who also said they received instruction, missing data rates are large (34 percent of the eligible sample according to physician data). Interviewers failed to conduct the test with three patients and one patient refused. Patients who normally test their urine were not asked to do so if the physician data indicated the patient was not given instruction in urine testing. This error rate is unknown. If the research objective is to measure the quality of the patient's urine testing skill, the most important measurement lesson to be learned from the above is not to use overly restrictive or error-prone screening questions to determine who is asked to demonstrate urine testing procedures. A screening item such as "Do you know how to test urine for sugar?" might reduce apparent errors of nonobservation considerably.

Watkins et al. (1967a) obtained patient demonstrations of urine testing during a home interview. Details of the procedure are unavailable. Interviewers made inferences about:

- How often the patient collects urine
- Correct procedure (second voided specimen, other)
- Reads results according to directions
- Regulates insulin according to test results
- Records results for use by the physician
None of the studies using patient demonstrations asked the patient for a report of usual testing procedures. Thus, the research does not furnish insight into the potential criterion validity of self-reports when the criterion is a demonstration.

**INSULIN ADMINISTRATION BEHAVIOR**

The following items were thought to be especially important by the Delphi panel:

- Accurately measures intended dose of insulin
- Varies dosage of insulin correctly according to urine test results

We also added:

- Takes insulin at appropriate times

Three studies attempt to assess patient insulin administration behavior. All use self-reports, one contains a demonstration allowing a criterion validity test of the self-report but failed to publish the validity results in sufficient detail. None of the self-report question items are published.

Gabrielle and Marble (1949) included items about whether campers knew how to administer insulin and whether they got supplementary insulin (at home) if the urine tests were bad. The quality of the answers is unknown.

Hulka et al. (1975b) had their nurse interviewers ask patients to show their medication, say what it was for, report the physician's scheduling prescription, and whether or not the patient was following the prescription. It is not clear from published reports whether there was additional questioning about varying dosage or when the medication was usually injected. The quality of the self-reports was not evaluated.
Watkins et al. (1967) obtained self-reports from 115 insulin-dependent patients about their current dosage and then asked each to measure out that dose. Thirty-two of 115 patients did not measure the amount they reported they were taking. Unfortunately, the direction and magnitude of the discrepancies is not given in the published report. It is not known whether the self-reports are biased and, if so, by how much, or, if not, what amount of imprecision is contained in the verbal answers.

None of the studies attempted to measure the time schedule followed by the patient for administering insulin.

SAFETY BEHAVIOR

Of the safety behavior items rated by the Delphi panel, the literature refers to measures of only two of them: carrying identification and carrying emergency carbohydrates. The quality of self-reports of the former has been ascertained by asking the patient to show the identification.

The top rated items are:

- Eats in response to hypoglycemic symptoms.
- Informs relevant other persons (family, friends, coworkers, school personnel) about symptoms of and treatment for hypoglycemia (if at risk for hypoglycemia).
- Carries source of glucose (emergency carbohydrate) (if insulin-dependent).
- Reports severe hypoglycemic episode (causing change in consciousness) to physician promptly.
- Reports moderate or greater levels of urinary ketones to physician.
- Carries diabetic identification (if at risk for hypoglycemia).
- Correctly varies dosage of insulin according to symptoms of hypoglycemia.
Gabrielle and Marble (1949) ascertained whether their camper-patients carried some form of carbohydrate while at home. They do not report exactly how this information was obtained from the children nor its measurement quality.

Hulka et al. (1975b) evaluated patient reports of carrying something that identified him as a diabetic. Patients whom the doctor told to carry the I.D. were asked whether the doctor did recommend this. Forty-six of 89 said yes. Forty of the 46 were willing to show it to the interviewer. This is an incomplete criterion validity assessment design so the nature of the measurement error in the self-report question is unknown. The answers are not perfect. Future studies might use designs that ask the self-report question of all respondents and do not make the request for a demonstration contingent on the answer to the question.

Measures for the remainder of the safety items, all given very high importance ratings by the panel, need to be developed and evaluated.

FOOT CARE

After considerable debate, the Delphi panel concluded that diabetic hygiene requirements were not much different from those for the general population. The one exception was in foot care. For the subset of diabetics with existing neuropathy or vascular impairments, the following item was judged important:

- Inspects feet daily, cuts nails correctly, and wears shoes that fit well.

Although many studies have measured knowledge of foot care requirements, none appears to have assessed actual behavior.

This review reveals the absence of published measures of many important things a diabetic may do for himself to help manage his diabetes. Many of the unmeasured behavior categories are those
rated most important by the expert panel. Techniques to obtain high-quality assessments in these areas need to be developed because evaluations of diabetic self-care that do not address these important variables could be seriously incomplete.
VII. KNOWLEDGE TESTS

Although knowledge tests have been frequently used in published studies of diabetic patients, parents of patients, and providers, no psychometric evaluations of these instruments were identified in the published literature. The majority of published items pertain to informational categories that were rated low in importance by our expert panel (e.g., pathophysiology, prevalence, epidemiology). Some important informational categories were frequently represented by published test items. However, published information about these items is limited, without exception, to difficulty levels (percent of correct responses). This information, particularly in the case of items fielded in more than one population (e.g., Etzwiler's Diabetes Knowledge Test), will be helpful in selecting and writing relevant items for further study. However, essential information regarding the discriminatory power of items (e.g., item-total correlations) and reliability must be obtained during the next phase of the project.

Considerable emphasis has been placed on assessing diabetic patient knowledge, based largely on the assumption that accurate information about diabetes and an understanding of how to recognize and control diabetic symptoms are prerequisites to successful management and control of the disease. Knowledge tests have been used for the following purposes: (1) to increase understanding of the problems involved in patient management and control of diabetic symptoms (Bowen et al., 1961; Watkins et al., 1967a,b; Williams, 1967; Partridge et al., 1972; Bohdan and Jans, 1977; Morley et al., 1977; and Myers, 1977); (2) to evaluate the success of specific treatment or patient education programs (McDonald, 1968; Etzwiler and Robb, 1972; Hulka et al., 1975a,b; Salzer, 1975; Laugharne and Steiner, 1977; and Suren, 1977); (3) to describe the level of knowledge for diabetic patients and their parents (Beaser, 1956; Etzwiler and Sines, 1962; Collier and Etzwiler, 1971) and knowledge levels of the general U.S. population (Holland, 19680; (4) to assess the level of knowledge for providers (nurses and dieticians) involved in patient education (Etzwiler, 1967); and (5) to link knowledge
to patient sociodemographic and other characteristics (Simon et al., 1976).

One review (Graber et al., 1977) of diabetic education programs and evaluation instruments (including knowledge tests) was identified. Williams (1976) has also summarized what diabetic patients need to know and some of the literature on levels of knowledge among diabetic patient populations.

CONTENT AND STRUCTURE OF KNOWLEDGE TESTS

Knowledge tests have been fielded most often by Etzwiler and his colleagues (Etzwiler and Sines, 1962; Etzwiler, 1967; Collier and Etzwiler, 1971; Etzwiler and Robb, 1972). The latter three studies employed 34 or 35 item versions of Etzwiler's Diabetes Knowledge Test, which was the only instrument identified as having been used in more than one population. This test contains multiple-choice items covering five topics, as follows: general knowledge about diabetes, insulin effects, levels of control, symptoms, urine testing, and nutrition. Comprehensive tests (i.e., those covering three or more categories of information) have also been developed by Bowen and others (1961), Williams and others (1967), and Simon and others (1976).

Knowledge tests have most often been of the multiple-choice type; however, true-false tests (e.g., Williams et al., 1967) and open-ended (e.g., essay type, fill-in-the-blanks) tests have been employed (e.g., Hulkà et al., 1975a,b; Suren, 1977). Williams and others (1967) fielded all three kinds of tests in the same study; however, content areas were not comparable across formats. Suren (1977) described multiple-choice and open-ended questions and was the only investigator identified who used visual aids in his test; data for Suren's test have not yet been reported.

Although numerous publications pertaining to the assessment of diabetic knowledge were identified, only five reported actual items used in knowledge assessment. No publications presented psychometric evaluations (e.g., item discrimination statistics, reliability estimates of knowledge tests used). The published information about knowledge tests is very limited and permitted only the following to be
accomplished: (1) identification of items that appear (in terms of manifest content) to measure knowledge categories judged important by the expert panel; (2) a summary of the kinds of studies that have employed knowledge measures; (3) a description of the different data-gathering methods that have been used to assess knowledge; (4) some sense of the distribution of correct answers for knowledge items and test scores; and (5) some insight into other methodological considerations (return rates, administration times, etc.).

Publications that included detailed information about test items were reviewed carefully to assess content validity in relation to the conclusions of our expert panel. This analysis revealed that the majority of published items deal with specific areas of knowledge that were not rated high in importance by our expert panel. For example, knowledge tests have often focused on the pathophysiology of diabetes and pharmacologic mechanisms (Collier et al., 1971; Simon et al., 1976; Suren, 1977; and Tribble and Hollenberg, 1977); information about etiology, epidemiology, and prevalence of diabetes (Collier and Etzwiler, 1971; Simon et al., 1976), and prognosis and complications (Collier and Etzwiler, 1971; Simon, 1976), which were rated low in importance.

The relationship between knowledge items rated important or very important by the expert panel and items identified in the published literature is summarized in Table 10. The categories identified in the first column are those rated important, as previously reported (Marquis and Ware, 1977). Entries in the second column indicate whether each important knowledge category appeared to have been covered by one or more items in a given investigator's instrument. (It is important to note that published items varied considerably in quality and that at least some of the items referred to in Table 10 are not likely to satisfy item analysis criteria or reliability and validity requirements). Some important items (e.g., recognizing hypoglycemic symptoms, reading urine test results) were covered by one or more items in several published instruments. Five of the other important items (e.g., measuring insulin dosage, recording urine test results, storing test materials) were not covered in any of the identified tests and eight items (e.g., varying insulin with test results, losing weight, varying eating
<table>
<thead>
<tr>
<th>Important Categories/Items</th>
<th>Published Knowledge Tests(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Safety Measures</strong></td>
<td></td>
</tr>
<tr>
<td>Recognizing hypoglycemic symptoms</td>
<td>SI, ET, SU, TR</td>
</tr>
<tr>
<td>Contacting physician</td>
<td>SI, HU</td>
</tr>
<tr>
<td><strong>Medication</strong></td>
<td></td>
</tr>
<tr>
<td>Obtain oral medication</td>
<td></td>
</tr>
<tr>
<td>Take prescribed dosage</td>
<td>ET, HU</td>
</tr>
<tr>
<td>Measure insulin dosage</td>
<td>BE, ET, SU</td>
</tr>
<tr>
<td>Follow injection schedule</td>
<td></td>
</tr>
<tr>
<td>Vary insulin with urine tests</td>
<td>ET</td>
</tr>
<tr>
<td>Vary insulin with symptoms</td>
<td>BE, SI, ET, HU, SU</td>
</tr>
<tr>
<td><strong>Urine Testing</strong></td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>SI</td>
</tr>
<tr>
<td>Conduct urine test</td>
<td>SI, ET, HU, SU, TR</td>
</tr>
<tr>
<td>Read and interpret urine test</td>
<td>BE, ET, SU</td>
</tr>
<tr>
<td>Record urine test results</td>
<td></td>
</tr>
<tr>
<td>Store test materials</td>
<td></td>
</tr>
<tr>
<td><strong>Diet and Nutrition</strong></td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>BE, SI, ET, TR</td>
</tr>
<tr>
<td>Lose weight</td>
<td>HU</td>
</tr>
<tr>
<td>Maintain ideal weight</td>
<td></td>
</tr>
<tr>
<td>Follow eating schedule</td>
<td>SI, HU</td>
</tr>
<tr>
<td>Change eating schedule with symptoms</td>
<td>SI, ET</td>
</tr>
<tr>
<td>Change eating schedule with exercise</td>
<td>SI</td>
</tr>
<tr>
<td>Use of food exchange list</td>
<td>BE, SI, ET, SU</td>
</tr>
<tr>
<td>Measure or judge food proportions</td>
<td>SI</td>
</tr>
<tr>
<td><strong>Obtaining Medical Care</strong></td>
<td></td>
</tr>
<tr>
<td>Contacting providers</td>
<td>SI</td>
</tr>
<tr>
<td><strong>Hygiene</strong></td>
<td></td>
</tr>
<tr>
<td>Inspect feet</td>
<td>SI, ET, HU, SU</td>
</tr>
<tr>
<td>Cut toe nails</td>
<td>BE</td>
</tr>
<tr>
<td>Select well-fitting shoes</td>
<td>SI</td>
</tr>
</tbody>
</table>

schedules with exercise) were covered by only one test. Thus, considerable test item development will be necessary to adequately represent all of the information items rated high in importance by our expert panel. This is particularly true in view of the fact that the item categories in Table 10 do not cover all of the knowledge items that are implied by the behavior items rated high in importance by the expert panel.

Other instruments that appeared (from published descriptions of content areas) to cover important areas but for which items were unavailable may also be useful if copies can be obtained (e.g., Williams et al., 1967; Holland, 1968; Partridge et al., 1972).

Items developed for use in studies of other patient populations may also be useful in constructing diabetic knowledge tests. For example, the following were identified: interviewer question guides (Leary et al., 1971) focusing on carrying identification; use of medication in response to symptoms; a medical quiz developed by McKercher and Rucker (1977); a knowledge measure used by Parkin and others (1976) to study patient compliance; and an evaluation questionnaire including items pertaining to diet, which was designed for patients with myocardial infarction (Rahe et al., 1975).

**DATA-GATHERING METHODS**

In those studies where data-gathering methods were described, data were gathered using self-administered questionnaires most often (e.g., Etzwiler, 1967; Collier and Etzwiler, 1971; Salzer, 1975; Simon et al., 1976; Suren, 1977).

In evaluating information regarding methods used to construct tests and to gather knowledge data from patients, parents, and providers, it is important to keep in mind that important methodological details often were not presented in published reports. For example, information regarding the structure of items (e.g., forced-choice, open-ended) often was not given. The content of items was presented in only six of the published reports that were identified; however, some others summarized content categories. Other important methodological details were omitted more often than not, including return rates, data-gathering methods.
(e.g., interviewer or self-administered), and scoring methods. Personal
interviews (most often by nurses) have been employed by others (Bowen
et al., 1961; Andersen et al., 1963; McDonald, 1968; Hulka et al.,
1975a,b).

RETURN RATES

Information on the extent of returned and usable knowledge test
instruments has only rarely been reported. Watkins and others (1967)
were able to complete 60 of 75 (80 percent) home contacts for purposes
of testing knowledge in clinic patients. Approximately 81 percent (65
of 80) of patients who participated in a health education program for
diabetics were successfully tested by mail in a study reported by
Salzer (1975); however, only 34 of 80 could be tested a year later.
Finally, Simon and Stewart (1976) received completed tests from 74 per-
cent of the clinic patients participating in their study of knowledge
level.

DESCRIPTIVE STATISTICS FOR TEST SCORES

The most frequently reported statistical information about knowl-
edge test scores has been the percentage of correct and incorrect re-
sponses to items. These score distributions reflect item difficulty,
which is one of two important criteria in a psychometric evaluation of
knowledge tests. Generally speaking, it is desirable (for purposes of
maximum discrimination among subjects) to construct test items so that
they have approximately 50 percent of subjects (in the population of
interest) answer the item correctly. However, items described in the
published literature do not appear to have been analyzed with this cri-
terion in mind. Rather, the effect of item difficulty on conclusions
about knowledge level appears to have been ignored. In this regard,
some investigators (e.g., Andersen et al., 1963; Pratt et al., 1957)
have concluded that knowledge levels among diabetic patients are poor
without reporting data. Other results suggest that test performance
differs for patients, parents, and providers, namely, higher for par-
ents (Etzwiler, 1967) and that scores vary tremendously across informa-
tion categories (Bowen et al., 1961; Simon and Stewart, 1976).
OTHER PSYCHOMETRIC CRITERIA

Information regarding the internal consistency of test items (e.g., internal consistency reliability coefficients or item-total correlations) were not reported in any of the publications identified. Thus, it is not possible to determine (from published information) whether items in the same knowledge category can be combined meaningfully for scoring purposes and it is not possible to determine how much information we can expect to obtain with published test-items.

Although validity (i.e., what the tests measured) was never mentioned in published reports, sparse information relevant to validity was identified. Expert opinion was used to determine the appropriateness of items (i.e., face validity) in the test described by Simon and Stewart (1976) and it was shown that knowledge test scores were sensitive to educational interventions (Bowen et al., 1961; Etzwiler and Robb, 1972).

Other correlational results suggest that knowledge instruments were valid: positive associations with age (Simon and Stewart, 1976), greater for diabetics who have experienced serious symptoms such as coma or shock (Simon et al., 1976), and positive associations with adequacy of self-management (Watkins et al., 1967).
VIII. ATTITUDE ITEMS

Attitudes inferred from behavior rated important by the expert panel have only rarely been mentioned in the published literature and instruments used in the few studies of attitude have not been published. In part, this reflects the infrequent consideration of behavior-specific attitudes in studies of diabetic self-care and compliance behavior.

If instruments can be obtained, some published studies may prove useful in developing instruments to measure attitude items rated important by the expert panel (see Marquis and Ware, 1977). Khurana and White (1970) studied attitude toward diabetes for patients and their parents and attitude toward insulin and hypoglycemia. Attitude toward the disease, the regime in general, and toward diabetic patients was studied by Bowen and others (1961). Attitudes toward weight control and taking insulin were studied by Anderson and others (1963).

However, psychometric evaluations of attitude measures were not identified and nearly all of the behavior-specific attitudes rated high in importance by the expert panel were not mentioned in published reports. Thus, the development and validation of attitude instruments must begin with the construction of untested items during the next phase of the research.
The state of the art in measuring important diabetic self-care variables is not far advanced. A panel of experts identified the most important things a diabetic should do to manage his disease, yet measures for a large number of these variables have not even been developed for use in compliance research. Of the published measures identified, knowledge of their measurement quality is available only for a very few, and most of those did not focus on diabetics. Development work can proceed based, in part, on what has been learned about existing measures from this literature review.

The main conclusion to emerge about measurement quality is that the majority of measures that could be evaluated contain little or no measurement bias (e.g., forgetting, Hawthorne effects, time-sampling bias), but moderate to large amounts of error variance. These measures should only be used in large sample research studies. Attention should be paid to the effects of large measurement error variance on statistics such as the correlation coefficient because the attenuating effects of large amounts of error variance cannot be mitigated by increasing the size of the research sample.

There is a major caveat to the conclusion about lack of response bias. Measures with little or no bias were obtained either from general population samples or from persons not following a treatment regimen. Interviewers were not doctors, nurses, medical students, or other persons associated with a patient’s primary health care source. Self-report measures that were found to be biased (pill-taking compliance) were obtained by the patient’s doctor or someone who could be perceived as a proxy for the doctor. Respondents (or their children) were following a medical treatment regimen. This response effect is termed a compliance bias, and it may have been produced by the particular circumstances of the interview about compliance. The interview was conducted by someone who gave the orders to be complied with (or his agent). An additional hypothesis is that the compliance bias can be minimized by collecting data in a situation that does not contain
this particular set of demand characteristics (e.g., using a neutral interviewer and convincing the respondent that his or her answers are confidential). This hypothesis needs to be tested before patient compliance research, based on self-reports, continues much further.

A final general conclusion is that the time-sampling bias and error variance appears to be a very negligible part of total data error in the evaluated measures. Implications for further measurement development appear to be to devote fewer resources to correcting time-sampling effects and more to identifying and reducing other sources of error variance.

EVALUATION OF SPECIFIC MEASURES

The literature was used to evaluate several existing measures of self-care behavior. The results of these evaluations are summarized in Tables 11 and 12.

The quality of several self-report methods of measuring nutrient intake was assessed. Table 11 gives an overview of the results. The 24-hour recall interview, used successfully in large-scale nutrition research, received initial attention. Measurements obtained by it appear to contain little or no response or time-sampling biases, little time-sampling error variance, but large amounts of response error variance. In theory, an average of measures from repeated 24-hour recall interviews should contain a lower proportion of response error variance. No empirical evidence was found to support or refute the theoretical prediction. Interviews or diaries covering successive days (seven-day recall and seven-day record) might yield more precise estimates of usual intakes, but the published results are not definitive. The task of recalling eating details over an entire seven-day period might produce forgetfulness. One study supports the forgetting hypothesis. Another study had wives help in the recall and did not find forgetting effects. The seven-day record may also reduce error variance but we could not find empirical evidence for this. The requirement of keeping a food record for so long a time may produce a Hawthorne effect (the measurement task may induce the subject to change his usual eating habits). No Hawthorne effect was found in research designed to detect
<table>
<thead>
<tr>
<th>Measurement Method</th>
<th>Measurement Hypotheses</th>
<th>Potential Advantages</th>
<th>Potential Disadvantages</th>
<th>Measurement Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>24-hour recall</td>
<td>Feasibility and</td>
<td>Memory bias</td>
<td>Very small or none</td>
<td>Used successfully in</td>
</tr>
<tr>
<td>interview</td>
<td>compliance rates</td>
<td>Time-sampling bias</td>
<td>Very small or none</td>
<td>The Health &amp; Nutrition</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Response error variance</td>
<td>50-75 percent</td>
<td>Examination Study</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Time-sampling variance</td>
<td>Small but no recent</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Average man bias</td>
<td>estimates</td>
<td></td>
</tr>
<tr>
<td>Replicated</td>
<td>Reduce response</td>
<td>Increases expense</td>
<td>No data</td>
<td></td>
</tr>
<tr>
<td>24-hour recall</td>
<td>and time-sampling</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>interviews</td>
<td>variance</td>
<td>Increases attrition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seven-day</td>
<td>Reduces time-sampling</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>recall</td>
<td>and response variance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>interview</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seven-day</td>
<td>Less memory bias</td>
<td></td>
<td></td>
<td>No data</td>
</tr>
<tr>
<td>record</td>
<td>Reduces error</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(diary)</td>
<td>variance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hawthorne bias</td>
<td></td>
<td>Not found by two</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>studies that looked</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>for it</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Increases attrition</td>
<td></td>
<td>Cooperation varies</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>from very low to very</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>high</td>
<td></td>
</tr>
<tr>
<td>Diet history</td>
<td>Maximum possible</td>
<td>Expensive to use</td>
<td>Possibly very large</td>
<td></td>
</tr>
<tr>
<td>interview</td>
<td>accuracy</td>
<td></td>
<td>interviewer effects</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Requires nutritionist/</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>interviewer with spe-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>cial training and</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>more than an hour</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>per respondent or</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>interrogation time</td>
<td></td>
</tr>
<tr>
<td>Dietary</td>
<td>Inexpensive to use</td>
<td>Generally poor</td>
<td>Limited research</td>
<td></td>
</tr>
<tr>
<td>questionnaire</td>
<td></td>
<td>measurement quality</td>
<td>suggests bias and</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>error variances</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>similar to 24-hour</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>recall</td>
<td></td>
</tr>
</tbody>
</table>
Table 12
EVALUATION OF OTHER MEASURES OF SELF-CARE BEHAVIOR

<table>
<thead>
<tr>
<th>Measurement Topic and Method</th>
<th>Measurement Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-reports of oral medication dosage compliance</td>
<td>Overreporting bias. Sensitive to scoring rules. 50–75 percent error variance</td>
</tr>
<tr>
<td>Physician ratings of oral medication dosage compliance</td>
<td>Error variance may be 100 percent of total measured variance.</td>
</tr>
<tr>
<td>Self-reports of having the medication prescription filled</td>
<td>Some measurement error but type, source, and magnitude not yet known.</td>
</tr>
<tr>
<td>Self-reports of visiting any physician (or number of visits)</td>
<td>No self-report bias. Measurement error variance from 20–84 percent of total measured variance.</td>
</tr>
<tr>
<td>Self-reports of number of dentist visits</td>
<td>Little or no response bias. Error variance between 25–50 percent of total measured variance.</td>
</tr>
<tr>
<td>Self-report classification of type of specialist visited</td>
<td>Between 72 and 88 percent of the reported classifications are accurate.</td>
</tr>
</tbody>
</table>
it in seven-day record measures.

The extensive diet history measurement approach might elicit better quality responses because it is comprehensive, includes cross-checks, and is administered by a skilled nutritionist. But published evaluations of the diet history technique do not support this expectation. Viewing the evaluation studies together, it appears that diet history data are very sensitive to interviewer effects. Other measurement approaches such as the dietary questionnaire look somewhat promising as inexpensive ways of obtaining measurements, but, again, do not appear to be able to solve the problems of imprecision. Cost-effective diet assessment techniques have yet to be developed.

Measures of compliance with medication instructions and self-reports of obtaining medical care were also evaluated in this phase of our project. A summary of the findings of measurement quality is given in Table 12.

The self-reports of oral medication dosage contain the over-reporting bias referred to above as the compliance bias. Substituting physician ratings of dosage compliance removes the measurement bias but also removes the true score variation. The tested physicians could do no better than chance at predicting whether or not an individual will follow instructions for oral medication dosage. Patient reports of filling prescriptions for oral medication have not been fully tested by published measurement studies. Existing data indicate the self-reports are not completely accurate but the characteristics of the inaccuracy cannot be inferred.

Self-reported doctor visits, contrary to the belief of some measurement experts, contain little or no average response bias. The measurement error variance could be as low as 20 percent or as high as 84 percent of total measured variance. Self-reported dental visits show similar measurement characteristics but have not been studied so extensively. According to one study, people do a reasonably accurate job in reporting the types of medical specialists they consult. This particular research does not tell us how often the patient reports (or thinks) a specialist is a general practitioner, however.
Nonself-Report Measures

Is it possible to substitute nonself-report measures in studies of patient self-care? While the final answer depends on one's particular research goals, elements of an answer can be sketched. They involve whether existing nonself-report measures cover the topic domain of interest, whether their use introduces a design bias, how their particular measurement characteristics affect statistics of interest, and whether they are more or less easy to obtain than self-reports.

On the first point, nonself-report measures of private phenomena such as attitudes, beliefs, opinions, and knowledge are rare. Given the lack of a valid general theory that links private phenomena to directly observable behavior, we must rely on measures that involve introspection. Efforts to increase a patient's ability to introspect accurately are likely to have a bigger short-term payoff for measurement quality than efforts to develop nonself-report measures of private phenomena.

Written records of patient behavior often exist already and are a tempting source of information that could substitute for self-reports. These sources include medical appointment records, medical charts, pharmacy prescription ledgers, and health insurance records. If the patient is the unit of analysis, however, he is likely to obtain medical services from more than one source. Accurate measures of such actions as obtaining medication and obtaining all prescribed medical care depend on locating all record entries for the patient (or an unbiased sample of entries). Reviewing records of only one or two sources of care (e.g., the records of a health maintenance organization) is a potentially biased sampling procedure. Obtaining an unbiased sample of records of all sources of medical care in a geographic area for each individual under study is very expensive. Short cuts, such as checking records only of sources that the patient said he used, risks not detecting the care received (or prescribed and not obtained) from unreported sources. Either of these short-cut approaches potentially introduces a design bias that threatens the accuracy of conclusions drawn from the research data.

A third problem is that nonself-report measurements may contain
biases and error variance also. While it is comforting to assume that the physician's word on a written record is completely accurate and comprehensive, or that "objective" measures based on skilled observations of behavior, urine tests, chemical analysis, or physical measurement contain no error, these assumptions are unwarranted. A recent study by the Institute of Medicine (1977) points out potentially high clerical error rates in obtaining patient measures from medical records. Nonself-reports can yield biased results to the extent that the patient is aware of the current or impending measure and changes his behavior as a result. This is the Hawthorne effect bias.

Using objective measurement procedures sometimes poses practical problems. Some require skilled personnel to administer. Others require special equipment to analyze. Most require patient cooperation of some form. When measures of rare events are needed such as whether the patient recognized the early stages of potentially severe hypoglycemia and what he did to counteract it, there are often no practical alternatives to self-reports. Additional practical considerations enter when a "usual" behavior pattern is to be measured. Often, objective measures must be repeated over an unbiased sample of occasions within the time period of interest. Repeated measurement designs are expensive and risk introducing Hawthorne biases. One-time self-reports are a convenient way to obtain retrospective reports of usual behavior (or repeated behavior from which the usual pattern is inferred) but are subject to memory effects.

This discussion is not meant to imply that objective nonself-report measures are undesirable, but that these measures have a range of problems of their own that need to be considered in planning the measurement design of any research or evaluation study.

FURTHER MEASUREMENT RESEARCH AND DEVELOPMENT

Appendix A identified what a diabetic should know and to to take proper care of himself. Assessment of this entire range of important variables appears necessary for a meaningful evaluation of diabetic patient education programs. The preceding sections of this appendix have shown that measures of many of the important variables have not
been developed and that the quality of most existing measures is unknown. A large amount of research will be necessary to create and evaluate the full range of measures necessary for a comprehensive assessment of diabetic patient self-management. The measurement evaluations conducted point to the significant problems to be addressed in further work and indicate that some other kinds of potential problems can receive less attention. The literature review has provided a reasonably solid foundation for future measurement development work.
REFERENCES


Blake, Elaine C., and J.V.G.A. Durnin, "Dietary Values from a 24-hour Recall Compared to a 7-day Survey on Elderly People," papers read at the 154th meeting of the Nutrition Society at the Medical School, St. Mary's Hospital, London, W.2, on December 1, 1962.


Canneill, C. F., and F. J. Fowler, A Study of the Reporting of Visits to Doctors in the National Health Survey, Institute for Social Research, University of Michigan, October 1963 (mimeo).


Loewenstein, R., Two Approaches to Health Interview Surveys, Columbia University, School of Public Health and Administrative Medicine, New York, 1969.


Preston, D. F. and F. L. Miller, "The Tuberculosis Outpatient's Defec-
tion from Therapy," The American Journal of the Medical Sciences,

Questionnaire for Postmyocardial Infarction Patients," Heart and

Rickels, K. and E. Briscoe, "Assessment of Dosage Deviation in Outpatient

Rockart, F. J. and P. B. Hoffman, "Physician and Patient Behavior Under
Different Scheduling Systems in a Hospital Outpatient Department,"

Salzer, Joan E., "Classes to Improve Diabetic Self Care," American

Samuelson, G., "An Epidemiological Study of Child Health and Nutrition
in a Northern Swedish County," Nutrition and Metabolism, Vol. 12,

Shenfield, G. M. and J. M. Steel, "Urinary Glucose Testing by Diabetic

Simon, John W., and Michael M. Stewart, "Assessing Patient Knowledge
About Diabetes," The Mount Sinai Journal of Medicine, Vol. 43,

Stine, O. C., C. Chaugui, C. Jimenez, and W. C. Oppel, "Broken Appoint-
ments at a Comprehensive Clinic for Children," Medical Care, Vol. 6,

Stone, D. B., "A Study of the Incidence and Causes of Poor Control in
Patients with Diabetes Mellitus," American Journal of Medical Science,

Sudman, S., W. Wallace, and R. Ferber, The Cost-Effectiveness of Using
the Diary as an Instrument for Collecting Health Data in Household
Surveys, Survey Research Laboratory, University of Illinois, 1974.

Suren, Jean F., "Education of the Culturally and Educationally Deprived
Diabetic," Nursing Clinics of North America, Vol. 12, No. 3, September

12, 1958, pp. 446-461.

Tribble, Norma M., and Emily E. Hollenberg, "The Impact of a Quality
Assurance Program on Diabetes Education," Nursing Clinics of North


