Nutrition and Health, 2001, Vol. 15, pp. 17–28
0260-1060/01 $10

ARE RELIABILITY, REPRODUCIBILITY AND VALIDITY THE CORRECT TERMS TO ASSESS THE CORRECTNESS OF DIETARY STUDIES?

GLORIA JOACHIM*

University of British Columbia School of Nursing, Canada

ABSTRACT

Nutritional studies often use the terms reliability, reproducibility and validity to indicate the correctness of the study. These terms do not appear to have a universal meaning to all researchers. The components of a dietary study are the input, the data collection instrument and the compiled data. Frequently the data collection questionnaire/tool/instrument is tested for reliability, reproducibility or validity.

The data collection questionnaire/tool/instrument is simply a structure, a vehicle for gathering data. An argument is presented that demonstrates the reasons that such a structure cannot be tested for reliability, reproducibility or validity. The logical approach to the use of the terms reliability, reproducibility and validity is presented. Reliability refers to the input component of the study, reproducibility may or may not lead to strengthening the study and validity refers to the truthfulness of the database generated. Validity must be derived from reliable and reproducible data.

Frequently, nutritional studies use the terms reliability, reproducibility and validity of dietary questionnaires as a measure of the correctness of the study. These terms are used by some interchangeably and by others as distinct and different entities. At issue is how these terms could have different meanings to some researchers while being interchangeable to others, and how these terms are related to dietary questionnaires. Most often, researchers test the degree of reliability, reproducibility and validity of the dietary questionnaire/tool/instrument and assume that this reading determines the degree of correctness of dietary studies. At first glance, this approach appears reasonable, however, a systematic and logical analysis raises serious questions regarding this method of thinking.

*Address for correspondence: Gloria Joachim, Associate Professor, University of British Columbia school of Nursing, T-201 2211 Wesbrook Mall, Vancouver, British Columbia Canada V6T 2B5. Phone: (604) 822-7463, email: joachim@nursing.ubc.ca

Willett (1998), made fundamental observations regarding dietary studies. Willett made a clear distinction between the terms data and database. The structure of the database is the format, or list, of questions used to collect data. The data are the information collected from subjects, and the database is the total information collected using the structure of the questionnaire. Willett also noted that dietary studies have three components. The components are the input, the data collection questionnaire/tool/instrument and the compiled data/ database. Willett uses the terms reliability, reproducibility and validity as logical measures of the correctness of the data.

The purpose of this paper is to show that 1) the terms reliability, reproducibility and validity have distinct and different logical meanings when dealing with dietary studies, 2) when evaluating dietary studies, the approach must change from testing the questionnaire to examining the study as a whole, and 3) a common denominator or universal language must be established and used when dealing with the correctness of dietary studies.

LITERATURE REVIEW

Dietary researchers use terms that describe the research and results. Key words include reliability, reproducibility, validity and the notion of truth values. Much research time and money are spent trying to test the absolute correctness of dietary research. When absolute correctness cannot be shown, elaborate methods including measurements of reliability, reproducibility and validity are introduced by researchers.

Woods and Catanzaro (1988) define reliable instruments as those whose measurements are reproducible. Talbot (1995) defines reliability as the ability of an instrument to consistently measure what it aims to measure. Reliability refers to the stability of the instrument and measurement process when it is applied under standard conditions (Loether & McTavish, 1993). In the case of dietary studies, one might wonder about the meaning and relevance of standard conditions. The American Psychological Association (1985) defines validity as the appropriateness and usefulness of inferences made from the instrument. Talbot (1995) notes that a reliable instrument may not be valid.

Some researchers confuse absolute value with perceived value and make calculations between measures to derive a perceived value. Terms such as variance are used. Woods and Catanzaro (1988) compare the Likert scale measurements with thermometer scale measurements and state that the truth derived from the Likert scale is equivalent to that derived from a thermometer. In reality, the thermometer instrument measurement is an empirical and absolute measurement while the Likert scale, or any similar scale, is an observed value and not an absolute measurement.

The correctness of a single measurement cannot be trusted (Talbot, 1995). It could be correct by chance alone. Therefore, a true experimental design provides the strongest proof of relationship and causality (Talbot, 1995). Since most dietary studies are not true experimental studies and are nearly always conducted in a field setting, their correctness is more difficult to assess. In these studies, the notion of true values lies where O (observed) = t (true scores) + e (error score) (Talbot, 1995).

Assignment of Reliability and Validity to Questionnaires

In an attempt to prove that the results of a dietary study are true and correct, researchers have calculated the rate of change between instruments to derive the reliability and validity of an instrument. Can an instrument, or structure, really be reliable and valid? The reproducibility and validity of a 142-item food frequency questionnaire was assessed in a population of people in New Zealand (Metcalf, Swinburn, Scragg & Dryson, 1997). Reproducibility was higher among European people than the Maori and Island people. The results of the food frequency instrument were compared with three-day diet records and validity was judged good for most nutrients.

An eight-item questionnaire designed to assess calcium intake was developed and compared with a 52-item questionnaire to assess the reliability and validity of the shorter one (Angbratt & Moller, 1999). The short questionnaire was found to elicit similar answers to the longer one. Because of the scores, the shorter instrument was judged to be suitable to be used in place of the longer one. In Japan, a food frequency questionnaire was simplified by reducing the number of questions. The validity of the new questionnaire was assessed by comparing its results with one-day diet records (Takatsuka, Kurisu, Nagata, Owaki, Kawakami, & Shimizu, 1997). A life span history questionnaire of dairy food was tested for reliability and validity using a test re-test method for reliability and compared with two existing surveys for validity (Swan, Spitler, Whiddon, Evans, & Wells, 1997). Although both the reliability coefficient and the validity coefficient were 0.64, the instrument was judged as promising as a global assessment of lifespan dairy intake. In all of these assessments of reliability and validity, it was the structure, or the instrument/tool that was tested and judged.

In a study of reproducibility and validity among children, results of a questionnaire differed between a group of fourth and fifth graders and a group of sixth and seventh graders. Although the instrument used was the same, results differed with the age of the study population (Field, Peterson, & Gormaker *et al.*, 1999). The results of this study introduce a logical conflict. Although the instrument/questionnaire/tool was consistent, the data changed with the age of the population. The authors conclude that the instrument must be used with caution when dealing with a young population. This represents a clear contradiction of the notion that the reliability or validity of the questionnaire itself can be determined.

Expanding usual assumptions about the influence of time (season) on food consumption

It has long been known that food consumption varies with time (Subar, Frey, Harlan, & Kahle, 1994, Joachim, 1997, Beaton, 1994, Hankin, 1992). While the consumption of staple foods such as bread and eggs resembles a linear function, the consumption of seasonal foods resembles a parabola (Joachim, 1998). The seasonal factor for fruit, for example, peaks in July and August, declines to a seasonal low in December and January and then increases to a summer peak (Food Market Commentary, 1991). In North America, three types of seasonal food consumption patterns have been identified (Joachim, 1997). These patterns are food use 1) related to cycles of nature 2) according to person created or commercial conditions such as Christmas and 3) due to changes in the supply and demand causing trends (Joachim, 1997). An example of a change in food trend is the increased consumption of chicken following the Surgeon General's report (Liebman, 1992). Food market development refers to activities that increase the value of some foods for consumers (Kohls & Uhl, 1990). Willett stated that when economic resources are limited, food intake and income are related (Willett, 1990). The food lists of most food frequency questionnaires contain both seasonal and staple foods. Therefore, there are generally two consumption graphs in a single questionnaire. Some researchers have tried to minimize the issue of types of foods consumed and their effects on data collected by citing nutrient equivalencies. Although energy intake is stable over time (Tarasuk & Beaton, 1991), the intake of some nutrients varies with time (Tarasuk & Beaton, 1992). People alter the types of foods that they eat in different seasons to achieve the same caloric intake. This supports the significance of examining types of foods consumed when considering the relationship between diet and disease.

An interesting question regarding reported data involves the meaning of zero consumption. Does reported zero consumption about a food mean that a person never ate the food, or does reported zero consumption mean that the food was unavailable when the question was asked? Not all reported zeros are equal (Joachim, 1999).

Willett's Pragmatic Approach to Evaluating the Correctness of Dietary Studies

Willett made several fundamental observations about the effects of season on data collection that have often been overlooked by nutritional researchers. Willett noted that during some months of the year fruit and leafy green vegetables are not available to a specific population in Bangladesh, and had data regarding these foods been collected at that point in time, a single assessment of intake would have been misleading regarding the usual carotene intake of the subjects (Willett, 1990). Willett demonstrated the importance of the person, place and time model observing that asking the same question in two seasons would yield two different answers. Either answer could be correct depending upon the intent of the researcher and the research question. No amount of calculation between the two measurements makes either reading incorrect. In this case, to calculate a reliability coefficient between the two assessment points makes little sense. Furthermore, two readings when leafy green vegetables were unavailable could yield the same answer and both stand incorrect if the intent was to assess the usual carotene intake of the subjects.

Willett (1998) summarized his views regarding the importance of the accuracy of dietary studies saying that assessing food consumption by a population over a long period of time is the only way to obtain a true measurement of food consumption. Measuring food consumption using a single measurement or a few measurements would not necessarily give a true picture.

Willett (1998) notes that diet consumed is a function of the variables person, place and time—a standard epidemiological model. Willett further sets conditions under which reproducibility will occur when all variables remain constant. He describes the correctness of the data and not the correctness of the questionnaire (Willett, 1998).

DISCUSSION

A review of the literature regarding dietary studies, shows a wide range of approaches to demonstrate the correctness of the outcome/instrument/questionnaire/data. Many tests and calculations are performed. Words such as reliability, reproducibility and validity are used. Are studies more correct as a result of these tests? In the literature, the use of the terms reliability, reproducibility and validity is sometimes interchangeable and sometimes not. In a scientific analysis, it is impossible that the words have the both the same and different meanings. Confusion among the terms makes it difficult to compare studies.

Willett's Bangladesh example raises a clear logical problem. The intent was to study usual carotene intake. Willett noted that foods containing carotene were available only in a single season. If two or three assessments had gathered data when foods containing carotene were not available, the data would have been wrong and meaningless and yet reproducible in the same season. If two or three readings were done when leafy green vegetables were available, the data would have been reproducible and yet meaningless. If two or three assessments had been performed when carotene was available and one when carotene was not available, the results would have been diluted. The data would not have been reproducible and any comparison of the two readings would have made little sense regarding correctness. In this case, zero consumption would mean that foods with the nutrient carotene were not available at that point in time.

Willett (1998) makes it clear that data collected for dietary studies varies depending upon the study population, (person), the time of data collection (time) and the geographical area of data collection (place). In other words, the data collected are a function of several variables. Data are f [x1,x2,x3.......... xn] where x1 could be person, x2 could be place and x3 could be time or factors of person, place and time could be a grouping of variables. Willett also defines the process of data collection (Willett, 1998). From a large population, a sample is derived. Using a questionnaire that contains many questions, the sample population is asked to reply to the questionnaire and the data collected are organized into a database. While Willett made important observations about dietary studies, he did not outline the rules by which dietary studies should be evaluated.

Reliability, reproducibility and validity are logical operators. A logical operator is a term to which a truth value can be assigned. A truth value involves making a decision about whether something is true or false. When a mathematical test is performed, the results of that test can be judged to be true or false. An example is $x + y = z$ where x, y, and z could be numbers or expressions. We can assign to z the value T (true) or F (false) if the answer is correct or incorrect. Accordingly, reliability, reproducibility and validity are logical operators. To each of the terms, T (true) or F (false) could be assigned. Logically, it makes little sense to assign degrees of correctness. Mathematical correctness can be calculated but that does not change the logical correctness. This paper demonstrates that mathematical calculations can be made to produce mathematical answers, but the only true and correct way to evaluate nutritional studies is by judging them according to a test of logical correctness.

The Input

Normally, the researcher defines the subject for study. Using the definition for study, the sample population is selected. Selection of the sample population is the first potential problem for the input data. Does the sample population properly reflect the goal of the researcher as defined in the original idea for study? Was the method for collecting dietary data the correct method to achieve the objective of the study? These two distinctions must be dealt with before any data collection takes place.

Person

Dietary intake is a direct function of the subjects that contribute the data. Food intake depends on socioeconomic status, age, education, and ethnicity. Each one of these variables can be reflected in the outcome of the study. There is a greater consumption of fruit and vegetables among people with more income and education. These people can afford to eat seasonal foods out of season.

Place
The geographic location of data collection could influence the results of a dietary study. Although researchers may argue that improved methods of transporting food reduces the effects of seasonality and food availability, the cost of food distribution is reflected in food prices. Storage, fuel, and manpower affect the price of foods and therefore who is able to consume them. In the sunbelt of the United States, the prices and availability of fruit and vegetables, make consumers in the southern states choose different foods in different seasons from consumers in the northern states. This is an important point with implications for multicite studies.

Time
The time of data collection and the length of the study affect the ability to collect data that accurately reflect consumption over time. While seasonal foods are often thought to be associated with time of the year, seasonality is also related to trends (Joachim, 1997). The issue of time is particularly crucial when data collection involves inquiring about many seasonal foods. When the questionnaire contains many seasonal foods, the reproducibility of the output will depend greatly on when the data were collected.

To summarize the problems associated with input, it is becomes clear that the input is a function of many variables and issues. The breakdown to person, place and time is done for convenience purposes only. It is easier to deal with three groups of variables rather than to deal with individual variables independently.

The Questionnaire

The second component described by Willett is the questionnaire (Willett, (1998). The input is the source of the data and the questionnaire is the structure of the database to be constructed. The questionnaire is a form that generally consists of demographic data, a list of foods to be tested, and information such as the frequency of consumption of a particular food. Each person that contributes data is a record in the database. Each question represents a field. The name of a person is a field consisting of approximately 15 bytes. The type of data is alpha. Birthdate is a date field and consists of eight bytes. For each food, there will be a description of the food (could be 15 bytes) followed by subfields of frequency consumed per day (two numerical bytes) followed by per week (two numerical bytes) followed by per month (three numerical bytes). The questionnaire is a tool used as a format to collect data that will be stored in the database. The questionnaire is the blank form issued to the study subjects to be completed and returned to the investigator. The only impact that the questionnaire has on the results is the clarity and accuracy of the information recorded and entered as data into the database. When designing a questionnaire, the ratio between seasonally consumed food and regularly consumed food must be considered since questionnaires that contain a combination of regularly and seasonally consumed foods will likely demonstrate problems with reproducibility (Joachim, 1998).

Analysis of the Data

Willett makes a clear distinction between data and the database. Data collected from the sample population are referred to as raw data. Much time and effort must be allocated to ensure that the data are cleaned in terms of typographical errors and ambiguous entry into the database. The cleaning process is essential when dealing with the logical correctness of the outcome of dietary studies. Cleaned data form a workable database. Using the database as the source for the analysis, certain comparative calculations of observations can be made.

Currently, the terms reliability, reproducibility and validity are used. The three terms are logical operators that can be assigned the values true or false. To date, reliability usually describes the ability of an instrument to consistently measure what it aims to measure (Talbot, 1995). Reproducibility refers to whether or not the data can be reproduced. Validity has been defined to mean the extent to which an instrument measures what it aims to measure (Talbot, 1995). Some researchers test the reliability of the questionnaire. Some researchers assume that if the data are reproducible, the questionnaire is valid. Logically, there cannot be different interpretations of these terms. Logically, the conclusions of a study can be either true or false.

Logical Operators—Reliability, Reproducibility and Validity

When constructing a truth table using the above three terms, the following will appear. Assigning a logical T (true) value to reliability will occur under the following conditions: 1) consideration was given to person, place and time as affecting factors of the input. Willett's Bangladesh example illustrates the significance of time in the data collection 2) if any of the variables of person, place and time was not considered, then the F (false) must be assigned to the reliability part of the data collected. The emphasis of assigning true or false to the input is subject to the intent of the study. Referring to Willett's work, if the intent would have been to measure the intake of leafy green vegetables in the summer and the measurement occurred in the summer, then the reliability test equals T (true). If the intent was to show total consumption of leafy green vegetables over the year and the measurements would have taken place when leafy vegetables were not available, then the reliability factor would have been F (false). At the same time, if the intent was to measure the intake of leafy green vegetables in the fall and the question was asked in the fall, the reliability factor is T (true).

When assigning the truth value to reproducibility, we can see that the data could be reproduced when the reliability test is assigned a true (T) value or a false (F) value. In other words, data can be reproduced twice—when the data are deemed to be reliable and when the data are deemed unreliable.

The designation of reproducible by itself does not give information as to the validity of the outcome of the study. Logically, it is necessary to analyze under what conditions a study is deemed valid. It can be said that a study is valid if and only if the data are reliable and reproducible. It is incorrect to assume that data are valid only because the data are reproducible. It is also incorrect to say that data are valid only because the data are reliable. We cannot say that data are valid if and only if the data are reproducible, nor can we say that the data are valid if and only if the data are reliable. Data that are not deemed reliable could be reproducible, and hence the entire study would not be valid. This observation summarizes under what conditions we can assign a truth value to a study and designate it as valid.

Another issue that must be raised is the issue of the use of two comparison tests in order to determine validity. Many researchers increased the number of tests from 2 to three or more in order to assign truth value to a study and test whether or not a study is valid. Increasing the number of studies alone does not necessarily increase the measure of validity of a study. It could, but this could be by chance alone.

At no time in discussions of reliability, reproducibility and validity does the word questionnaire appear. The questionnaire is a too] and a format to collect data. The questionnaire is a structure that cannot be assigned a true or false value. The questionnaire cannot be reproduced, it can only be reprinted. Therefore, assigning a truth value to the questionnaire cannot withstand any logical test. At best, it can be said that the questionnaire captures the intent of the researcher or it does not. When studying protein intake, for example, does the questionnaire contain all possible protein sources? Using this distinction makes it easier to review dietary studies. This distinction makes it possible to compare dietary studies using standardized terminology.

Mathematical Approach to a Logical Problem

Reliability

It has been shown that reliable data are f(x1, x2, x3……xn)

where	xl, x2 x3 ... is a group of variables known as person
and	x4, x5, x6 is a group of variables known as place
and	x7, x8, x9 is a group of variables known as time.

When the correct assumptions regarding person, place and time, were made by the researcher, the data are reliable.

Reproducibility
Data collected are f(x1, x2, x3 xn).

Therefore, f(person, place and time)

A study could be deemed reproducible if all variables (x1, x2, x3......xn) are pegged. Any change in any of the variables will appear as a degree of change in the reproducibility of a study. There is no correlation between the reproducibility and validity of the data.

Validity
To test the validity of a study, the following occurs:

Assigning truth value to reliability:

Test I (tl) could be true (T) or false (F)
Test 2 (t2) could be true (T) or false (F)

Assigning truth value to reproducibility:

True when reliability test is T/T and F/F
False when reliability test is F/T and T/F

Assigning truth value to validity:

If and only if t1 = true
And t2 = true
Therefore, reproducibility = T (true)
Therefore, validity = T (true)

This mathematical approach proves that 1) validity derives from true data and not necessarily from reproducible data and 2) validity is derived from reproducible data but reproducibility does not necessarily make a study valid.

CONCLUSION

Dietary studies use a scientific approach and conclusions are based on scientific evidence. Using confusing definitions of terms, makes it difficult to draw scientific conclusions. The outcomes of dietary studies are crucial and used by professionals and lay people as a source of information. Not having a clear definition of terms, method of conducting studies and method of analyzing data makes it difficult to critically evaluate studies. While scientists create definitions and draw conclusions independent of each other, there are some trends of comparison.

A convention of terminology must be established. Perhaps the use of the words reliability, reproducibility and validity are not the best or the most accurate way to judge the correctness of dietary studies. A convention must be established so that all researchers use the same terminology when dealing with the outcomes of dietary studies. The language used must be precise, accurate and recognized by all. Failure to pay attention to this creates a huge gap and the potential for presenting misleading information. A dietary study should also include information about assumptions made regarding person, place and time when a food list is selected for inclusion in a questionnaire.

REFERENCES

American Psychological Association, Committee to Develop Standards (1985). Standards for educational and psychological testing, Washington, D.C.: Author.

Angbratt, M. and Moller, M. (1999). Questionnaire about calcium intake: can we trust the answers? *Osteoporosis International*, **9**(3), 220–225.

Beaton, G.H (1994). Approaches to analysis of dietary data: Relationship between planned analysis and choice of methodology. *American Journal of Clinical Nutrition*, **59 (suppl.)**, 253S–261.

Field, A.E., Peterson, K.E., Gortmaker, S.L., Chung, L., Rockett, H., Fox, M. K. and Colditz, G.A. (1999). Reproducibility and validity of a food frequency questionnaire among fourth to seventh grader inner-city school children: implications of age and day-to-day variation in dietary intake. *Public Health Nutrition*, **2**(3), 293–300.

Food Market Commentary (1991). *Agriculture Canada*, **13**, 1.

Hankin, J.H. (1992). Dietary intake methodology. In *Research: successful approaches* (E.R. Monsen, ed.), pp. 173–194. Chicago: The American Dietetic Association.

Joachim, G. (1997). Supply and demand a framework for explaining variability in dietary intake and its impact in data. *Nutrition and Health*, **11**, 289–299.

Joachim, G. (1998). Sources of variability in the reproducibility of food frequency questionnaires. *Nutrition and Health*, **12**, 181–188.

Joachim, G. (1999). The relationship between habits of food consumption and reported reactions to food in people with inflammatory bowel disease- testing the limits. *Nutrition and Health*, **13**, 69–83.

Kohls, R.L. and Uhl, J.N. (1990). *Marketing Agricultural Products*, 7th ed. Macmillan Publishing: New York.

Liebman, B. (1992). The changing American diet. In *Nutrition 92/93* (C.C. Cook Fuller, ed.). 5th ed. Dushkin Publishing Group: Guilford, Cn.

Loether, H.J. and McTavish, D.G. (1993). *Descriptive and Inferential Statistics*. 4th ed. Allyn and Bacon: Toronto.

Metcalf, P., Swinburn, B., Scragg, R. and Dryson, E. (1997). Reproducibility and validity of a food frequency questionnaire in European and Polynesian New Zealanders. *Ethnicity and Health*, **2**(4), 297–308.

Subar, A.F., Frey, C.M., Harlan, L.C. and Kahle, L. (1994). Differences in reported food frequency by season of the questionnaire administration. The 1987 national health interview survey. *Epidemiology*, **2**, 226–233.

Swan, P.D., Spitler, D.L., Whiddon, S., Evans, B.W. and Wells, M.F. (1997). Assessment of lifetime patterns of dairy food intake and physical activity. *Nutrition and Health*, **11**(4), 255–270.

Takatsuka, N., Kurisu, Y., Nagata, C., Owaki, A., Kawakami, N. and Shimizu, H. (1997). Validation of a simplified diet history. *Journal of Epidemiology*, **7**(1), 33–41.

Talbot, L.A. (1995). *Principles and Practice of Nursing Research.* Mosby: St. Louis.
Tarasuk, V. and Beaton, G.H. (1991). The nature and individuality of within subject variation in energy intake. *American Journal of Clinical Nutrition*, **54**, 464–470.
Tarasuk, V. and Beaton, G.H. (1992). Day to day variation in energy and nutrient intake: evidence of individuality in eating behavior? *Appetite*, **18**, 43–54.
Willett, W.C. (1990). Nature of variation in diet. In *Nutritional Epidemiology Monographs in Epidemiology and Biostatistics* (B. MacMahon, ed.) pp. 34–51. Oxford University Press: London.
Willett, W.C. (1998). Reproducibility and validity of food-frequency questionnaires. In *Nutritional Epidemiology Monographs in Epidemiology and Biostatistics.* 2nd ed. (J. Kelsey, M. Marmot, P. Stolley and M. Vessey, eds.) pp 101–147. Oxford University Press: London.
Woods, N.F. and Catanzaro, M. (1988). *Nursing Research Theory and Practice.* Mosby: St. Louis.

(*Received 30 September 2000*)