1980

A Psychometric Study of an Evaluation Instrument Designed to Measure Adults' Health Knowledge and Attitudes

David William Rivers
Loyola University Chicago

A Psychometric Study of an Evaluation Instrument Designed to Measure Adults' Health Knowledge and Attitudes

by

David William Rivers

A Thesis Submitted to the Faculty of the Graduate School of Loyola University of Chicago in Partial Fulfillment of the Requirements for the Degree of Master of Arts

December

1980

ACKNOWLEDGMENTS

I acknowledge, first of all, my constant advisors Drs. Leonard Bickman and John Edwards. I am particularly grateful for their expeditious reviews of preliminary versions of this thesis. I owe a special thanks to my parents, William and Elizabeth Rivers, for their unending support and encouragement. Finally, I wish to express my appreciation and thanks to my wife, Carol, for her invaluable patience and understanding.

VITA

The author, David William Rivers, is the son of William Clarence Rivers and Elizabeth (Fuchs) Rivers. He was born November 24, 1955, in Washington, D.C.

His elementary education was obtained at Saint Mary's School in Rockville, Maryland, and Good Counsel High School, Wheaton, Maryland, where he graduated in 1973.

In September, 1973, he entered The Catholic University of America, Washington, D.C., and in May, 1977, graduated Magna Cum Laude with a Bachelor of Arts degree in psychology. While attending Catholic University, he became a member of the Phi Eta Sigma and Psi Chi honor societies.

In September, 1977, he was granted an assistantship in applied social psychology at Loyola University of Chicago.

TABLE OF CONTENTS

LIST OF TABLES

CONTENTS OF APPENDICES

INTRODUCTION

This study was conducted to assess the psychometric properties of an instrument designed to measure adults' knowledge, attitudes and behavior regarding a variety of health-related issues. The instrument itself was primarily developed for the purpose of determining the impact of the Chicago Heart Association's (CHA) heart health curriculum program on the teachers implementing the curriculum.

The ultimate goal of the program is to influence actual lifestyle behaviors in the areas of nutrition, exercise and cigarette smoking. The means of achieving such change is through the acquistion of health knowledge and the development of positive attitudes toward healthful living. The rationale for this approach is an extension of Ajzen and Fishbein's (1980) theory of reasoned action. It is assumed that education leads to knowledge acquisition which alters attitudes and eventually initiates behavior change (Swanson, 1972).

The quality of the evaluation of such a program depends on a number of things, particulary the research design and the measures of dependent variables. Though great care and expense is often taken to utilize superior research designs and analytic procedures, it is apparent that such attention is not payed as often to the development of valid, reliable measures of the variables involved. The lack of adequate measuring instruments has been one of the more significant deterrents to the advancement of social psychology as an explanatory and predictive science (Bohrnstedt, 1970). Over the years, social psychologists have for the most part ignored psychometric issues (Bickman, 1980). While attention has been focused on producing precise independent variables, the reliability and validity of dependent measures has been taken for granted. This has serious implications for the advancement of social psychological theory. It is often the case that the practical utility of a conceptual model is reduced because there is an absence of valid measures of the constructs included in the model (Cummings, Jette & Rosenstock, 1978). Statistically significant relationships among "psychological" variables may be found, but our understanding of human behavior is not advanced unless the instruments used to measure those variables are accurate reflections of the underlying psychological dimensions. And when the underlying variables are unobservable, abstract concepts such as one's attitude toward a nutritional diet, the measurement process is especially arduous. It becomes the objective of the researcher to assess the true (unobservable) relationships among variables, with measuring instruments that are only estimates of the concepts involved. The chief means of ensuring that an accurate measuring instrument has been developed is to establish its two basic psychometric properties: reliability and validity.

Reliability and validity concern the degree to which an instrument is free of measurement error. The two basic kinds of error that affect empirical measurements are random error and nonrandom error. A nonrandom error is one introduced into measurement by some factor that systematically affects the characteristic being measured or the process of measurement. Estimates of validity are concerned chiefly with nonrandom errors though random errors also have a diminishing effect on validity. (The effect of random error on validity will be discussed later in the context of the relationship between reliability and validity.)

Broadly defined, validity is the extent to which scores on a measuring instrument reflect the differences among individuals on the characteristic, or attribute, that one is seeking to measure. So any factor that systematically biases a measuring instrument (i.e., a nonrandom error) will result in an invalid estimate of the underlying attribute. That is, the measure will no longer be an accurate indicator of the characteristic. As Althauser and Heberlein note, "matters of validity arise when other factors -- more than one underlying construct or methods factors or other unmeasured variables -- are seen to affect the measures in addition to one underlying concept and random error" (1970:152).

Again, the measurement process is designed to assess the degree to which a person possesses a particular attribute. So, if we are measuring attitudes toward physical exercise, then people who receive high scores on the scale should actually have a more positive attitude toward exercise than people who receive low scores. But there is no way of knowing what an individual's actual attitude is. That is, the individual's "true" score cannot be known. Because there is no way of directly determining whether a score on an instrument reflects an individual's true position on the variable being measured, it is necessary to gather a variety of evidence to assess the validity of the instrument. This fact provides the reason why an instrument cannot be described as valid or invalid. It can only be said that the available evidence indicates that the instrument has some degree of validity. There are a variety of methods through which such evidence may be assembled and these will be described in a later section.

Whereas validity concerns both random and nonrandom error, reliability is particularly concerned with random errors only. The most widely used model for assessing random measurement error is referred to as classical test theory (Carmines and Zeller, 1979). Classical test theory involves a basic formulation consisting of an observed score, a true score, and random measurement error. Theoretically, any observed score is composed of two parts: a true score and an error component. Expressed as a formula:

$$X = t + e$$

where $\underline{X}$ is the observed score, $\underline{t}$ is the true score, and $\underline{e}$ is the random error. The <u>observed score</u> is the actual score obtained by a person on a measuring instrument. The <u>true score</u> is a hypothetical quantity that cannot be directly measured. Conceptually, it is an indicator of the person's true ability, attitude, or whatever the scale is designed to measure. Quantitatively, a person's true score is the average score that would be obtained if the person were remeasured an infinite number of times on that attribute. Again this is not a real, observable quantity but a hypothetical value. The random error component is due to chance factors of several types: factors operating within the individual (such as fatigue or temporary inattention); factors associated with the test itself (such as limited sampling of behaviors belonging to a universe or domain of content); and factors associated with administration or scoring procedures (such as limited time alotted for administration or subjectivity of scoring)(Sellitz, Wrightsman, and Cook, 1976). In some cases the error component may raise the observed score, in others it may lower it. The "positive" errors are just as likely as the "negative" errors, and their magnitudes are similar as well. That is, such errors are assumed to be random and independent, averaging over the long run to zero. But we do not measure people over the long run. In most instances we only measure them once. As a result, the observed score is only an estimate of the true score to the degree that random error is absent from that observation. But random error is always present to some degree. So then, the intent becomes to minimize error so as to obtain observed scores as close to true scores as possible.

Just as any one observed score is the sum of a person's true score plus random error, the total variance for a set of observed scores (S_o^2) is composed of two parts, the true score variance (S_t^2) and variance due to measurement errors (S_e^2):

$$S_o^2 = S_t^2 + S_e^s$$

Ordinarily, the variance of a sum is not the simple sum of the individual variances; also included would be the covariance between the true score and error. But in this case, it is assumed that the correlation between true scores and errors is zero, so that term (actually two times the covariance) drops out of the equation.

The above equation indicates that the greater the influence of true score variance on the total observed score variance, the more precise the scores are as estimates of true scores. If, on the other hand, most of the total observed score variance is variance due to measurement errors, then the observed scores are heavily influenced by chance and therefore lack reliability. Theoretically, the reliability of a measure (r_{xx}) is expressed as a ratio of true score to obtained score variance:

$$r_{xx} = \frac{S_t^2}{S_o^2}$$

Moreover, the reliability coefficient can be expressed in terms of error variance as follows:

$$r_{xx} = 1 - \frac{S_e^2}{S_o^2}$$

Thus the more error variance there is proportional to the observed variance the closer the reliability is to zero. Conversely, the degree to which the observed variance is uncontaminated, that is, contains no random measurement error, the closer the reliability is to one. A measure with a high reliability coefficient, therefore, is accurate in the sense that it produces observed scores that reflect the underlying, unobservable true scores. By contrast, the closer the reliability is to zero, the more the observed scores represent only error and chance factors that are unrelated to the characteristic being measured.

The above formulas are particularly useful for illustrating the conceptual notion of reliability. However they are not typically utilized when actually assessing the reliability of a given measure. Rather, as will be described in the next section, product-moment correlation coefficients are used in the operational definitions of reliability. That is, reliability is estimated through the use of correlation coefficients.

THE ASSESSMENT OF RELIABILITY

Generally, measures of reliability can be classified into three groups; coefficients of stability, equivalence, and internal consistency. The three vary in the procedures used to collect relevant data and in the meaning or interpretation of the resulting coefficient. Differences in meaning basically derive from differences in how consistency is defined.

COEFFICIENT OF STABILITY

A measure of stability attempts to address the issue of whether or not the results obtained during one administration of an instrument are replicable. That is, are scores on a measure due to true variation or are they the result of situation-specific factors? The more influence that extraneous factors have on scores, the more the instrument will reflect random error variation instead of true score variaiton and will provide unstable scores. The primary means of determining the stability of an instrument is through the test-retest method.

The test-retest method involves administering the same measure to the same sample of respondents at two different points in time. It is assumed that the responses to the measure will correlate over time because they reflect the same true underlying variable. The more influence that situation-specific factors have, the lower the correlation will be. It is assumed that these factors represent random error and this error is uncorrelated across parallel measurements (ie, test-retest administrations of the same instrument) (Carmines and Zeller, 1979). So the reliability coefficient is equal to the correlation between the scores on the same measure obtained at two points in time. And again, this correlation reflects the degree to which the instrument contains true score variation versus random error variation.

Though the test-retest method is a simple procedure that corresponds closely to the conceptual notion of reliability, the information that it provides is ambiguous. There are several alternative explanations that could account for a low correlation between the results of the two administrations other than an unreliable or unstable instrument.

When utilizing the test-retest method problems may occur with respect to the time interval between measurements. If the interval is short then respondents may actually remember their earlier responses and attempt to duplicate them on the second administration. As a result, during an interval of two to four weeks it is likely that memory will be such a factor so as to substantially overestimate the reliability of a measure (Nunnally, 1964). On the other hand, the longer the time interval is, the more likely it is that the concept being measured will itself change (Bohrnstedt, 1970). A low test-retest correlation may be an underestimate of the reliability of a scale if true change in the characteristic being measured has occured. It is also possible that true change will not result in a lowering of the correlation. If each respondent's score increases for example, the correlation may not be lowered at all. On the other hand, if all of the respondents scores increase but a ceiling effect occurs, the correlation will be lowered. It is necessary therefore, to look at change in the distribution of scores over time, not change in the mean score.

Another problem that leads to deflated estimates of reliability when utilizing the test-retest method is reactivity. Reactivity refers to the fact that a respondent's sensitivity or responsiveness to the variable under study may be heightened by the measurement of that variable (Campbell and Stanley, 1966). For example, measuring a person's attitude toward a regular program of physical exercise may enhance the person's interest in the matter and cause him or her to examine the benefits of such a program. This might result in a true change in attitude across time which would not have occured if the person had not been surveyed. The test-retest correlation will be lower, due not to an unreliable instrument, but due rather to reactivity.

COEFFICIENT OF EQUIVALENCE

The correlation between scores from two forms of an instrument given at the same time is a coefficient of equivalence. Known as the alternative-form method, the procedure may also involve administering the measures at different points in time, thus incorporating aspects of the test-retest method. It is intended that the two forms provide measures of the same underlying concept. The chief means of ensuring such equivalence is by randomly selecting items to be included on each form. That is, twice as many items as are needed for one instrument are created and these are randomly divided to provide two forms.

The alternative-form method is superior to the simple test-retest method because it reduces the influence of the respondent's memory on the administration of the second instrument. Yet again, if the forms are administered at two points in time it will be difficult to distinguish true change from unrelibility unless the characteristic being measured is a relatively enduring one (Nunnally and Durham, 1975).

The primary drawback of the alternative-forms method is the practical difficulty of constructing two forms of an instrument that are parallel. Given the properties of parallel measurements, that task can be rather arduous, if not impossible.

COEFFICIENTS OF INTERNAL CONSISTENCY

Determining reliability from a single administration of one form of an instrument yields reliability estimates known as coefficients of internal consistency. These coefficients convey the degree of consistency of the content within the single instrument form.

One of the earliest devised methods of determining internal consistency is the split-half technique. This method involves dividing the total set of items into two halves and calculating the correlation between the scores on each half. But this correlation would be the reliability for each half of the scale rather than the total scale, so it is necessary to statistically correct the correlation in order to obtain an estimate of the reliability of the entire scale. This correction is known as the Spearman-Brown prophecy formula (Spearman, 1910; Brown, 1910).

There is a difficulty with the split-half technique in that the correlation between halves will vary, depending on how the items are divided. There are various ways to divide a scale, e.g., odd versus even-numbered items. But each split is likely to result in a different correlation between the halves. So, the split-half technique does not allow one to arrive at a single, determinate estimate of reliability.

Probably the best estimate of internal consistency is coefficient alpha (Cronbach, 1951). Though thought of as only a measure of internal consistency, coefficient alpha is based on both the average correlation among items (the internal consistency) and the number of items, as can be seen in the following expression:

$$\alpha = \frac{Nr}{1+r(N-1)}$$

where <u>N</u> is equal to the number of items and <u>r</u> is equal to the mean interitem correlation. Alpha is a generalization of KR20 (Kuder and Richardson, 1937), which is used to estimate the reliability of scales composed of dichotomously scored items.

Coefficient alpha addresses the issue of homogeneity. That is, alpha can be interpreted as assessing the degree to which all of the items in a scale are measuring the same characteristic. This question is particularly relevant with respect to the assessment of the CHA survey. As stated earlier, the knowledge and attitude instruments in question are designed to assess the impact of a program. This assessment becomes a question of individual differences in which the object is to classify respondents along a specified dimension (e.g., attitude toward smoking). A summary score is calculated that indicates the position of each respondent on a continuum of the attribute being measured. The logic of this process of adding items together is based on the condition that the items are positively correlated (i.e., internally consistent or homogeneous). So, adequate internal consistency is necessary in order to justify the calculation of a total score over a number of items (Nunnally,1978). Otherwise it would make little sense to speak of the total score as representing the measure of a single characteristic.

As noted above, coefficient alpha is not only a function of the average interitem correlation, but it also varies depending on the number of items in the scale. Thus, increasing the number of items on a scale can increase the scale's reliability, provided there is not a reduction in the average interitem correlation (Carmines and Zeller, 1979). However there is a point of diminishing returns; the adding of items indefinitely makes progressively less impact on the reliability. Moreover, it takes a greater amount of time and resources to construct a longer instrument and a longer instrument results in increased respondent burden.

It is important to note the alpha has the same logical status as coefficients derived from the other methods of assessing reliability. It is assumed that the items of a scale are parallel measurements and that they only differ from one another because of strictly random error. And given that parallel measurements have equal intercorrelations, the average interitem correlation accurately estimates all of the correlations in the item matrix. Thus, "logically, using the average correlation in the calculation of alpha amounts to exactly the same thing as calculating a simple correlation between parallel measurements" (Carmines and Zeller, 1979:47).

From the above discussion it is evident that coefficient alpha is perhaps the best estimate of reliability available. Each of the other methods described -- test-retest, alternative-forms, and split-half -- have either methodological or practical limitations which can be compensated for in the use of coefficient alpha. Where appropriate these other methods may be employed, but only as an adjunct to alpha, not as a substitute (Nunnally and Durpham, 1975).

THE ASSESSMENT OF VALIDITY

As defined earlier, validity is concerned with the degree to which scores on a measuring instrument reflect the differences among individuals on the characteristic that one is seeking to measure. However it is not the instrument itself that is validated, but rather one validates the data that arise from the measure and the specific procedure employed (Cronbach, 1971). As a result, validity is not a static, enduring property. A scale may be considered valid for one purpose but not for another. Or it may have some degree of validity today but that may not be the case at some point in the future (Crano and Brewer, 1973).

In addition to viewing the validity of a measure as being unstable over time, it is also true that even at one point in time it is not possible to say that a measure is valid or invalid. Rather validity is a relative, descriptive term, not an all-or-none property (Nunnally, 1978). An investigator can only establish the extent or the degree to which a measure may be valid. While the concept of validity is relatively straightforward, there are in fact different types of validity. Each of these takes a somewhat different approach as a means of assessing the extent to which an instrument measures what it is intended to measure. The use of the different approaches allows one to gather a variety of evidence regarding the validity of a measure. This assures the researcher a more complete assessment than would be achieved through the use of only one approach. The following sections describe in detail the meaning, methods, and interpretation of the different types of validity.

CONTENT VALIDITY

An assessment of content validity gives evidence as to how well the set of items on an instrument represents the domain of the concept about which generalizations are to be made. In other words, content validity depends on the adequacy with which a specified domain of content is sampled (Nunnally and Durham, 1975). Content validity is important in that if one is attempting to measure a complex concept (e.g., an attitude), it is necessary to include items that are representative of all of the important aspects and dimensions of the concept. The extent to which a measure taps into each of the attributes of a given characteristic, the more likely it is that scores on the instrument will truly represent the quality of the characteristic.

An important issue to be raised here regards the dimensionality of a construct. Earlier it was stated that the reliability of a measure depends on its internal consistency, i.e., the degree to which the items in scale are measuring the same thing. On the other hand, content validity is concerned with how well the items are measuring different things. Though it is essential that a measure tap into a variety of dimensions, each dimension should be closely related to the overall concept. In empirical terms -- the items on a scale should be intercorrelated enough to provide homogeneity (i.e., reliability), but they should not be so highly correlated that they are only measuring a single aspect of the concept (i.e., a lack of content validity). In other words, at one level the attribute being measured should be unidimensional; at another level, within the attribute, there may be several dimensions each of which needs to be tapped into in order to provide a more valid indicator of the attribute.

The assessment of the content validity of measure is not a straightforward task. In fact, as Nunnally (1978) recommends, one should not test content validity, but rather it should be ensured by the plan and procedures of construction. Determing what is a relevant or important dimension may be done on a theoretical or empirical basis. Theoretical considerations may indicate that a given number of dimensions are believed to be important to the concept and should be assessed. Empirical approaches such as factor analysis are used to examine the relationships among a number of already existing items to determine how sets of items (factors) are related to one another. The statistical importance of these factors is used as an indication of content validity . Of course, in utilizing the empirical approach, it is necessary to first construct a number of items that measure the concept.

Two basic steps can be employed that provide the researcher with a structured, systematic approach to the development of a content valid measuring instrument (Bohrnstedt, 1970). First, the researcher should specify the full domain of content that is relevant to the trait or attribute in question. The domain should be stratified such that each dimension and aspect is represented. For example, in designing a scale to measure attitudes toward smoking, one might identify a variety of dimensions -- smokers' rights, value of good health, personal appearance, etc. The more dimensions that are identified the easier it will be to construct items and the more valid the scale will be.

The second step then is to write several items that capture the shades of meaning associated with each dimension. It is not possible to specify the exact number of items that ought to be constructed for each dimension. It can only be said that it is better to have too many items than too few. This is important because subsequent item analyses will likely eliminate many of the items due to poor internal consistency. If only an item or two are written for a particular dimension and they are eliminated, that aspect of the construct will not be represented and the content validity of the scale will be weakened. On the other hand, if several items had been written, it is likely that the final instrument would contain at least one or two items for that dimension.

The general approach to be used to ensure content validity is evident. It has been utilized a great deal by educators in the development of achievement and proficiency tests (Carmines and Zeller, 1979). Social psychologists however, have typically not been able to approach content validity to as great a degree. The main reason for this is that the theoretical concepts in social psychology (e.g., attitudes) have not been described with the required exactness. Specifying the domain of content for an arithmetic test is a considerably easier task than it is for a scale designed to measure attitudes toward a regular program of physical exercise. Social psychology deals with concepts that are most often abstract; and few if any of these have an agreed upon domain of content.

It is not to be suggested that content validity is an ideal that except in certain circumstances cannot be achieved. Rather, methods of scale consturction should always be aimed at ensuring that the domain of content is specified as best it can and items are written so as to sample this content adequately. The more carefully and systematically that the researcher undertakes the process of scale construction the more likely it will be that the final product will indeed have a high degree of validity. But in no way is content validation a fully sufficient means of assessing the overall validity of measures in social psychology (Carmines and Zeller, 1979). That task is left to validation procedures that are more empirically grounded. These are described below.

CRITERION-RELATED VALIDITY

Criterion-related validity probably has the closest correspondance to the conceptual notion of validity. It is a pragmatic approach; validity is judged in terms of the accuracy with which some criterion can be predicted based on the results of the measuring instrument at hand. If the criterion exists in the present, the assessment is of *concurrent validity*. If the criterion is to be obtained in the future, *predictive validity* is assessed. Whichever method is used the degree of criterion-related validity depends only on the degree of correspondance between the measuring instrument and the criterion. That is, the correlation between the scale and the criterion is taken to be the validity coefficient (Bohrnstedt, 1970).

Criterion-related validity is so empirically grounded that it may in some cases be totally atheoretical (Carmines and Zeller, 1979). The only concern is that the criterion can be predicted accuately. For example, if it were found that attendance at college football games correlated highly with success in graduate school, then such attendance would be a valid measure for predicting success in graduate school. This example points out an important weakness of criterion-related validity. The simple correlation between two variables does not allow one to say why such a relationship exists. Thus in terms of theory building, criterion-related validity may provide little if any information regarding the interrelationships of variables (Crano and Brewer, 1973). This is especially true if the concern is only to predict the criterion with little interest in why the measuring instrument is related to the criterion.

However it is usually the case that theoretical considerations do enter the process of criterion-related validation. When trying to validate a measure, there must be some basis for choosing a criterion measure. The selection process is often guided by theory to some degree. As a result, observed correlations may have more meaning than simply indicating how well the criterion can be predicted. That is, if a relationship between the measure and a criterion is expected on theoretical grounds, then a high correlation does provide some confirmation for the theory.

Although criterion-related validity is chiefly concerned with accurate predictions of criterion, it can also be a source of information regarding the nature of theoretically based relationships among variables. In the latter case criterion-related validation also gives evidence of construct validity. That is, the degree to which variables behave in theoretically predicted manner establishes the construct validity of the instruments used to measure those variables. The following section will describe those methods that are designed to more directly assess the construct validity of measuring instruments.

CONSTRUCT VALIDITY

The most theoretically oriented type of validity assessment is that of construct validity. In essence a construct is a hypothetical variable; it is a trait, an attribute, or a process which is inferred to have real existence and to give rise to measurable phenomena (Chaplin, 1975; Selltiz, Wrightsman & Cook, 1976). Many instruments are designed to measure the degree to which an individual possesses some characteristic or trait (e.g., attitude toward smoking). But because these characteristics are abstract phenomena, it is impossible to determine directly whether or not the instrument is indeed measuring the characteristic in question. Rather, it is necessary to infer indirectly as to how well scores on a given instrument truly represent varying levels of the attribute the instrument is supposed to be measuring. The process of construct validation attempts to establish how the attribute relates to other theoretically meaningful variables. That is, does the measure of an attribute correlate highly with a measure or indicator of another variable that is known to be related to the attribute or is thought to be, based on a well specified theory? And moreover, is the measure uncorrelated with variables known or thought to be unrelated to the attribute?

The importance of construct validation cannot be overstated. The science of psychology is chiefly concerned with finding functional relations among important variables. Hypotheses about human behavior are tested by studying the effect of one variable upon another. The extent to which any conclusions can be drawn regarding behavior (or cognitions and emotions for that matter) depends on the construct validity of the instruments that are purportedly measuring the relevant variables (Nunnally and Durham, 1975).

The process of construct validation is essentially a three step process (Carmines and Zeller, 1979). First, the theoretical relationship between a set of two or more concepts must be specified. Second, the empirical relationship between the concepts needs to be examined. Lastly, the construct validity of the measure(s) must be interpreted in light of the empirical evidence.

It is clear that some sort of theoretical framework is necessary in order to establish the construct validity of a particular measure. However, it is not necessary to have an extensive or fully developed theory (Cronbach and Meehl, 1955). The main requirement is only that the theory be detailed enough in order to state a set of hypotheses that involve the particular concepts. The construct validity of the measures used as indicators of the theoretical concepts can then be empirically tested based on those hypotheses.

A common hypothesis regarding many theoretical concepts is that two particular groups or samples of individuals will differ significantly with respect to the concept. For example, if it is believed that Democrats have more liberal political attitudes than do Republicans, then an empirical test of construct validity would be to administer an instrument designed to measure political attitudes to a group of Democrats and to a group of Republicans. If the group of Democrats receive scores on the measure that indicate that they have significantly more liberal political attitudes than Republicans, then there is evidence that the scale possesses some degree of construct validity. This straightforward approach to construct validation is referred to as the known-groups method (Crano and Brewer, 1973).

By its name the known-groups method implies a significant drawback of the technique. The groups that are employed in the empirical test are supposed to be known to be different with regard to the construct in question. In many situations, this requirement is very difficult to meet. Often times it is only assumed that the groups are different. The example above is a case in point. Research on the nature of political attitudes may not be so conclusive that it allows one to draw the conclusion that indeed Democrats have more liberal political attitudes than do Republicans. Perhaps it can only be assumed that that is the case. If such an assumption is going to be made in employing the known-groups method, then the researcher must gather additional evidence that supports the assumption that the groups are different. Additional self-report measures might be used or perhaps reports from others; these could be used to establish the status of the individual on the criterion variable.

In any case, even if the groups are not absolutely known to be different, the technique can still be utilized -- provided there are data to support any assumptions that are made. The known-groups method has been employed in a number of studies where the authors did not state that the groups were actually "known" to be different. Rather they presented data to support that assumption (for example, Crewe, 1967; Fischer, 1970; Parcel, 1975). It is evident that the known-groups technique as well as other assessments of construct validity, are closely related to theory testing. In the example above, if the measure of political attitudes was known to be valid and reliable, administering the measure to the two different groups becomes a test of the theory of political attitudes. However, in construct validation the assumption is necessarily made that the theory is correct and it is the measures that are being tested.

Like the known-groups approach to construct validation, the multitrait-multimethod matrix technique (Campbell and Fiske, 1959) examines the conceptual relationships among variables. Through the use of correlation matrices the technique simultaneously examines two more aspects of validity -- <u>convergent validity</u> and <u>discriminant validity</u>. Convergent validity refers to the notion that differents methods of measuring the same trait, or abstract concept, should yield similar results. That is, the methods should converge if they are validily measuring the same concept. Discriminant validity refers to the notion that similar (and different) methods of measuring different traits should yield different results. That is, two (or more) instruments should be able to discriminate between different traits if they are valid measures of those traits. Correlations between the measures of same traits and different traits are examined to determine the amount of variation that is shared among the measures. For example, if the measures are to have construct validity then the correlation between similar measures of the same trait should be higher than the correlation between two different traits utilizing those same measurement methods. This should be due to the sharing of common variation between the two similar traits over and above that common variation due to the similarity of the measurement technique. This is just one example of the many comparisons of correlation coefficients that are made within the multitrait-multimethod matrix as a means of assessing the construct validity of the various measures. Essentially it is a matter of attributing common variation to the methods being used to measure the traits (indicating a lack of construct validity), or attributing common variation to the traits themselves (indicating evidence of construct validity) (Crano and Brewer, 1973).

The basic requirement that should be met in order to utilize the multitrait-multimethod matrix is that there are at least two different traits measured in at least two different ways. The methods as well as the traits should be as maximally dissimilar as possible (Sullivan and Feldman, 1979). Table 1 exemplifies a typical configuration of multitrait-multimethod intercorrelations; in this case, three traits --

TABLE 1

EXAMPLE MULTITRAIT-MULTIMETHOD MATRIX

		Method 1			Method 2		
		Trait A (1)	Trait B (2)	Trait C (3)	Trait A (4)	Trait B (5)	Trait C (6)
	S (1)	r_{11}					
1	E (2)	r_{12}	r_{22}				
	D (3)	r_{13}	r_{23}	r_{33}			
	S (4)	r_{14}	r_{24}	r_{34}	r_{44}		
2	E (5)	r_{15}	r_{25}	r_{35}	r_{45}	r_{55}	
	D (6)	r_{16}	r_{26}	r_{36}	r_{46}	r_{56}	r_{66}

A, B, and C -- are measured with two methods -- 1 and 2. The entries in the table represent correlation coefficients. Each coefficient is one of four different "kinds" of correlations (Sullivan and Feldman, 1979). The first kind of correlation is actually a reliability coefficient, i.e., r_{11}, r_{22}, etc. The effect of reliability on the validity of measures will be disussed in a later chapter.

The second kind of correlation in Table 1 is that between the same trait measured with different methods, i.e., r_{14}, r_{25}, r_{36}. These are known as validity coefficients and are the focus of the convergent validity assessment.

The third type of correlation in Table 1 is that between different traits measured with the same method. These different-trait, same-method correlations are represented in Table 1 by the following entries: r_{12}, r_{13}, r_{23}, r_{45}, r_{46}, r_{56}. Lastly, the fourth kind of correlation is between different traits measured with different methods, i.e., r_{15}, r_{16}, r_{26}, r_{24}, r_{34}, r_{35}. These are referred to as different-trait, different-method correlations. These last two types of correlation coefficients are the focus of discriminant validity, that is, the degree to which the measures can discriminate between different traits.

Given these four different kinds of correlation coefficients, there are four criteria that are used in examining these correlations for evidence of construct validity (Campbell and Fiske, 1959). These are outlined in Table 2.

TABLE 2

CRITERIA FOR EXAMINING CORRELATION COEFFICIENTS
IN A MULTITRAIT-MULTIMETHOD MATRIX

1. The validity coefficients should be significantly different from zero and large enough to encourage further study.

2. Each validity coefficient should be larger than all different-trait, different-method correlations that are in the same row or column as the validity coefficient.

3. Each validity coefficient should be larger than the different-trait, same-method correlations which involve the same trait used for the validity coefficient.

4. The pattern of correlations should be the same within each triangle of coefficients representing the different-trait, same-method correlations and different-trait, different-method correlations.

It is very likely that in most cases, all of the criteria set forth in Table 2 will not be met by the data. Even with valid measures this may be true. Differing levels of reliability and validity, and chance fluctuations due to sampling error will result in inconsistent patterns within the correlation matrix (Sullivan and Feldman, 1979). When inconsistencies do occur there is no clearly defined route that should be taken to make an assessment of construct validity. However, because validity is not an all-or-nothing quality it is possible in many cases to make some statement regarding the construct validity of the various measures. Clearly, different patterns of correlations will provide varying levels of evidence for validity. At the highest level, when the four criteria are met, the evidence might be viewed as conclusive. In other instances, it might only be said that there is an indication that the measures have some degree of validity, but this assessment ought not be considered conclusive.

As a means of concluding this discussion of the various methods of assessing the validity of measuring instruments, it should be emphasized that these approaches need not be viewed in isolation. Rather, each assessment can be used to help interpret the other. (For example, the construct validity of measure is likely to be dependent on the extent to which content validity was ensured during the scale construction process.) The process of instrument validation is one of accumulating a variety of evidence (positive or negative) from each of the different types of validity (Crano and Brewer, 1973).

Just as the different types of validity are not mutually exclusive, they are also closely related to reliability. The next chapter will discuss the impact that reliability has on the validity of a measuring instrument.

THE RELATIONSHIP BETWEEN RELIABILITY AND VALIDITY

Though reliability and validity are distinct properties that a useful measuring instrument must have, an instrument cannot be valid unless it is reliable. This fact is evident if one considers that if a measure is unstable or inconsistent, i.e., contains random errors, it is not possible that it could measure any construct validly. Again, validity concerns the degree to which an observed score represents the "true" score. And if there is variable error present in the observed score (i.e., it is unreliable), the correspondence between the observed score and the "true" score will be limited, and thus, the measure will be less valid. In fact, the square of the correlation between the observed score and the "true" score is equal to the reliability of the measure (Bohrnstedt, 1970), as illustrated below:

$$r_{xy}^2 = r_{xx}$$

Furthermore, validity as determined by the correlation of a measure with some outside criterion can never exceed the correlation of an observed score with its "true" score to the extent that the measure is unreliable:

$$r_{xy} \leq r_{xt}(r_{xx})^{1/2}$$

where r_{xy} is the correlation between the measure and the criterion (observed validity coefficient), r_{xt} is the hypothetical correlation between the measure and the "true" score (the "true" validity of the measure), and r_{xx} is the reliability of the measure. In other words, the square root of the reliability of a measure places an upper limit on the correlation of the measure with an outside criterion (or any other measure for that matter).

The above logic can be extended to estimate what the correlation between two variables might be if the measures employed were made to be perfectly reliable. This estimate is known as the <u>correction for attenuation</u> and is shown below:

$$r'_{xy} = \frac{r_{xy}}{(r_{xx} r_{xy})^{1/2}}$$

where r'_{xy} is the correlation corrected for attenuation (Nunnally, 1978). How the correction for attenuation ought to be used has been an issue of considerable debate. It has been proposed that two pieces of information be considered before deciding to utilize the correction for attenuation (Bohrnstedt, 1970; Nunnally, 1978). The first consideration pertains to the estimates of reliability. The correction should only be used if reliability has been appropriately assessed with a large enough number of cases to allow confidence in it as a good estimate of the population parameter.

The purpose for which the correction for attenuation will be used is the second point that must be considered. It is never appropriate in a research study to correct the correlation between two variables and then report the corrected correlation as an indicator of the relationship between the variables (Bohrnstedt, 1970). However, if one is examining causal relationships and wishes to estimate what the true causal relation between two variables is, then correcting for attenuation might be in order. The objective in this case is to estimate what the true relationship is between two variables by correcting the fact that our measures of those variables are unreliable. Again, it is necessary that good estimates of reliability are available. Overall, it can be said that correcting for attenuation can be used in the early stages of research when one is exploring the relationships between variables but it should not be consistently applied as a tool to compensate for inadequate measuring instruments.

In summary it can be said that the extent to which a measure is unreliable (i.e., contains random errors), its validity is necessarily lessened. This is not to say that high reliability means high validity; an instrument might consistently be measuring the wrong thing. Reliability is a necessary but not a sufficient condition for validity. Reliability shows that something is being dependably measured, but not necessarily the concept of interest. "A valid measure with low reliability is more useful than a reliable measure of something one does not care to measure" (Selltiz et al., 1976, p. 197).

The psychometric properties that a useful measuring instrument must possess have been detailed in the preceding chapters. The following chapters will describe how the CHA health survey was developed and subsequently assessed regarding its adequacy as a useful tool to measure adults' health knowledge and attitudes.

DEVELOPMENT OF THE SURVEY

The CHA survey under consideration in this study is designed to measure adults' knowledge, attitudes and behavior regarding a variety of health-related issues. Specifically the survey contains several measures as outlined in Table 3. (See Appendix A for a copy of these measures.) Construction of these instruments entailed a multiphase pilot test and item refinement procedure. The primary basis for instrument refinement consisted of an examination of the results of item analyses. Where appropriate, inspection of item-total correlations, item discrimination, response option distrubution, and internal consistency furnished the rationale for item selection and refinement. Consultation with CHA staff and advisors provided additional input into the scale construction process. Their expert advice was utilized to assure content validity, particularly with regard to the relevance and importance of items on each measure. The next sections will describe in more detail the development process for each of the measures outlined in Table 3.

TABLE 3

CHA HEALTH SURVEY:
CONTENT AREAS AND METHODS OF MEASUREMENT

Content Area	Measurement Method
Health Knowledge	Multiple choice (4-alternative)
Attitude toward smoking	Likert (7-point scale)
Attitude toward physical exercise	Likert (7-point scale)
Attitude toward a nutritional diet	Likert (7-point scale)
Attitude toward smoking	Behavior Consequences (7-point scale)
Attitude toward nonsmoking	Behavioral Consequences (7-point scale)
Attitude toward physical exercise	Behavioral Consequences (7-point scale)
Attitude toward a nutritional diet	Behavioral Consequences (7-point scale)
Smoking behavior	Current status (5-point scale)
Exercise behavior	Inventory (Hours per week)
Diet behavior	Inventory (Servings per week)

HEALTH KNOWLEDGE QUESTIONNAIRE

The first step in constructing the knowledge questionnaire was to identify the types and areas of knowledge teachers were expected to gain during the curriculum program. A thorough review of CHA materials, including film strip scripts, information packet, and activity sheets, indicated that over two dozen specific content areas were covered. These were categorized into five groups corresponding to the five modules of the program and 30 items were prepared for each section. The items were primarily constructed from CHA materials, although other items were taken and modified from existing instruments, such as the cognitive test used in a previous CHA program. The first phase adult knowledge test consisted of 160, four alternative, multiple choice items.

After several phases of administering the knowledge test and conducting item analyses, a 50-item instrument was assembled. At the early stages , items were analyzed by determining the percentage of respondents who gave each of the four responses to each question. Each item was also assessed in terms of how well it discriminated between those people who scored above and below the median. On the basis of these analyses, it was possible to identify the foils (wrong answers) that were over- or underutilized, and items that failed to discriminate between high and low scoring respondents. In general, items were retained on the early pilot instruments if between 20 percent and 80 percent of the respondents answered them correctly, if each foil was used at least five percent of the time, and if they were more likely to be answered correct by people having above average total scores. (The total score for a respondent is simply the total number of correct answers.)

During subsequent phases of pilot testing, inspection of item-total correlations and estimates of internal consistency (i.e., coefficient alpha) were undertaken in addition to the analyses described above. These analyses were aimed at evaluating the instrument as a whole. That is, the primary concern became one of developing a reliable, homogenous set of items whereas earlier analyses were focused more at the individual item level.

All of the procedures just described -- precise specification of the content domain, multiphase pilot testing, and thorough analyses of individual items and the test as a whole — were undertaken for the purpose of constructing a measure of health knowledge, that is content valid, sensitive, and internally consistent. The following sections describe how similar methods were utilized in the construction of two sets of attitude assessment instruments.

HEALTH ATTITUDE LIKERT SCALES

The first task in designing any attitude scale is to precisely identify the object of attitudes one wishes to measure. That is, people have attitudes toward or against an object, which can be a person, group, physical object, abstract idea, event, or behavior (Fishbein and Ajzen, 1975). Following examination of curriculum materials and consultation with CHA staff it was decided that the attitude scales (Likert and Behavioral Consequences Methods) should deal with attitudes toward three types of behavior: (1)smoking, (2)regular exercise, and (3)eating a properly nutritious diet.

The procedure for constructing the Likert summated ratings scales (Likert, 1930) began by writing a large number (75) of statements, 25 for each of the three attitude domains. The content of the items was based on the review of curriculum materials, existing attitude scales and relevant attitude research. Approximately equal numbers of moderately favorable and unfavorable items were constructed. The response format consisted of a 7-point scale ranging from strongly disagree to strongly agree. (See Appendix A.) The total attitude score was computed by recoding all reversed-direction items and summing the values over the total number of items for each domain.

The appropriate item analyses for a Likert scale follows from the way in which the scale is conceptualized. It is assumed that the probability of agreeing with a positive statement and the probability of disagreeing with a negative statement about the attitude object is a linear function of the favorability of the respondent's attitude toward the object (Fishbein and Ajzen, 1975). This operating characteristic or traceline of scale items is indicated by the size and direction of the correlation between an item and the total scale score. This total is actually a corrected total in that it is the sum of the ratings of all the items excluding the one in question. Thus, the chief item analysis procedure consists of examining these corrected item-total correlations which should be approximately .30 or higher. Though there is not an absolute cut-off point that should be utilized, reliable scales of this length (10-15 items) generally have item-total correlations of .30 or higher. As discussed earlier, these correlations should not be so high however, that the items are measuring only a single dimension of the concept. Adequate item-total correlations indicate that the scale has high internal consistency. Thus by definition, a properly constructed Likert scale is also a reliable scale. That is, it provides a homogeneous measure of the construct of interest.

Like the knowledge questionnaire, the Likert scales were subjected to a multiphase pilot testing process. At each stage, item analyses were conducted to select those items with the strongest correlations with the total score. The final instruments contain ten items for each of the three attitude domains -- smoking, exercise, and diet. The next section will detail the development of the other set of attitude assessment measures, the Behavioral Concequences scales.

HEALTH ATTITUDE BEHAVIORAL CONSEQUENCES SCALES

The Behavioral Consequences scales were formulated under the guidance of expectancy-value theory, which suggests that the probability of a person's behavior, with respect to some object, is a function of the expected value (positive or negative) of the consequences of the behavior and the probability that the behavior will lead to those consequences (Fishbein and Ajzen, 1975; Peak, 1955; Rosenberg, 1956). Fishbein's theory in particular states a person's intention to behave in a certain way is partly determined by attitudes toward the behvior which, in turn, is a function of these two variables. Thus, these scales were developed to measure people's health related goals and the relationship of their behavior to those goals.

The first step in developing this instrument was to construct a list of goals, values, or consequences of behavior that are generally related to health issues and particularly relevant to smoking, exercise, and diet. This list was prepared, by referring to existing items from previous research that used the expectancy-value approach, past research on social values (e.g., Edwards, 1967), suggestions from CHA personnel, CHA materials and other sources.

The expectancy-value approach requires that the list of consequences be rated according to two quite different sets of instructions. First, the consequences are rated in terms of the extent to which they are seen as instrumentally related to some specified behavior, i.e., the extent to which some behavior leads to or prevents goal attainment. (For example, "Engaging in a regular program of physical exercise prevents/leads to my getting heart disease.") Second, the consequences are rated in terms of their affective value, such as their desirability or importance to the respondent. (For example, "My living longer than average is bad, makes me dissatisfied, unhappy/is good, makes me satisfied, happy.")

The Behavioral Consequences scale allows the calculation of an attitude score based on each of the ratings described above. The attitude index is computed by multiplying the instrumentality rating of a behavior (i.e., prevents or leads to) by the affective rating (important versus unimportant) for each of the consequences and summing these products over the consequences. For the initial pilot test, a list of 26 consequences was developed, as described above, and phrased in personalized terms. Again using multiphase pilot testing and item analyses a final list of consequences was constructed, containing 16 items. (This process will described in more detail below.) Each consequence was rated in terms of its importance to the respondent, but the instrumentality rating with respect to a given behavior was only made for a subset of those 16. This is because some consequences are only relevant to a particular behavior; for another behavior the consequence may be seen as being unrelated to the behavior by the vast majority of people, and thus provides an insensitive measure of one's attitude toward the behavior. So the final instrument contained 11 instrumentality ratings for each behavior, some of which were unique to that behavior, others of which were rated for all of the behaviors. The ratings were made on 1-7 point scale, with appropriate labels for each task. (See Appendix A.)

Thus an attitude score for each behavior was computed by multiplying the instrumentality rating by the affective rating for each of 11 consequences and summing these products over the 11 consequences. In order for this weighted sum attitude score to make sense psychologically, it was necessary to convert the responses from the unipolar 1-7 point scales to bipolar (-3 to +3) scales for both the instrumentality and importance ratings. Thus, if a person says that a behavior strongly prevents (-3) an unimportant consequence (-3), or that a behavior strongly leads to (+3) an important consequence (+3), both would indicate a positive attitude toward the behavior (i.e., a product of +9). Alternatively, if a person says that a behavior strongly prevents (-3) an important consequence (+3), or strongly leads to (+3) an unimportant consequence (-3), both would indicate a negative attitude toward the behavior (i.e., a product of -9). This type of calculation yields a highly sensitive attitude measure due to the very wide potential range of scores. In this case, the range of scores could be from -99 (extremely negative attitude) to +99 (extremely positive attitude).

As stated above, a multiphase pilot testing procedure was employed for the development of the Behavioral Consequences scales. For the first phase, the initial list of 26 consequences was rated in terms of two behaviors only: smoking and nonsmoking. The reason for dealing just with the smoking issue was the centrality of that topic to the program, its importance as a health problem in society generally, and its clarity as a form of behavior in comparison with the other two behaviors (exercise and nutrition). Since the population can be roughly divided into two groups, those who currently smoke to some degree and those who do not, four different kinds of instrumentality judgments could be made: Smokers can rate the consequences of (1)their continuing to smoke and (2)their stopping smoking; and nonsmokers can rate the consequences of (1)their continuing to not smoke and (2)their starting to smoke. A screening question was used to divide the respondents into five smoking categories. The categories were: (1)someone who smokes just about every day (regular); (2)someone who smokes once in awhile but not everyday (occasional); (3)someone who used to smoke regularly but has quit (exregular); (4)someone who used to smoke once in awhile but quit (ex-occasional); and (5)someone who has never smoked. People in the first two categories were defined as smokers, and people in the latter three categores were defined as nonsmokers.

After completing the screening question, the smokers were asked to rate each consequence three times: first, the extent to which they believed their continuing to smoke would either prevent or lead to each consequence; second, the extent to which they believed that their stopping smoking would either lead to or prevent each consequence; and third, the affective value of each consequence in terms of its personal importance and degree of satisfaction it would produce. Nonsmokers also rated each consequence three times: first, in terms of how their continuing to not smoke was related to each consequence; second, how their starting to smoke would relate to each consequence; and third, how personally important each consequence was. Thus the Behavioral Consequences scales permit the calculation of two smoking attitude scores: attitudes toward continuing one's present behavior (either smoking or not smoking) and attitudes toward the alternative behavior (either stopping or starting to smoke, respectively).

Several analyses were conducted to determine the sensitivity and relevance of the items individually and the scales as a whole. For example, the ratings of individual consequences were compared to determine which were the most important in distinguishing among different attitudes. As a result, consequences were eliminated from the scale if they were redundant (i.e., were very highly correlated with another consequence) or were nondiscriminating (i.e., did not distinguish between individuals having very positive versus very negative attitudes). A shorter version of the Behavioral Consequences scales was then constructed for the second pilot phase. This instrument included a few additional items specifically related to the other two attitudinal topic areas: exercise and nutrition. Additional pilot testing and item analyses were conducted for these two scales as well as the smoking scales. After this series of analyses the final instrument was constructed having 11 consequence ratings for each behavior. The consequences were rated as follows: (1) the extent to which engaging in a regular program of physical exercise leads to or prevents each, (2) the extent to which sticking to well-balanced, low cholesterol diet leads to or prevents each, (3) and (4) the extent to which smoking or nonsmoking leads to or prevents each, and (5) the desirability or goodness/badness of each consequence.

HEALTH BEHAVIOR SCALES

Individual behavior with regard to smoking, exercise, and diet was assessed through self-report questionnaire responses. Current smoking behavior was assessed via the screening question utilized in the Behavioral Consequences scales. For the exercise behavior measure, an inventory of activities was developed. The inventory represents a broad spectrum of athletic, sport, physical activities. (See Appendix A.) The response format consisted of having the respondent indicate the amount of time each week spent on each of the activities. In order to calculate a total score that validly assesses the degree of physical activity, a table of caloric expenditure for various activities (Morehouse and Miller, 1976) was consulted. The amount of time respondents reported they spent engaging in the various activities was multiplied by a weight from 2 to 10, based on the table entry for that activity. These products were then summed to arrive at the total score representing the degree and amount of physical activity engaged in per week.

An inventory was similarly developed as a measure of diet behavior. The foods in the inventory represent a number of selections from each of the five main food groups. In order to calculate a total score that reflects the nutritional value of one's diet, a nutrition food guide (Chicago Heart Association, 1977) was consulted. Foods in the inventory were weighted -2, -1, or +2, to reflect the degree to which each food was recommended according to the food guide. Respondents indicated the number of servings per week they typically have of each food, and these responses were multiplied by the appropriate weight. These products were then summed to arrive at a total score that represents the nutritional value of the respondent's diet. One drawback of this measure is that a high score should indicate good nutritional habits, but it may also be the result of someone eating a great amount of a particular good food. This would not be an example of good nutritional behavior.

In summary, it can be stated that a great deal of time and expense went into the development of the CHA survey. The measures of health knowledge and attitudes went through a detailed progression of pilot testing and instrument refinement. Though the behavior measures were not as thoroughly pilot tested, a careful, systematic approach to their development was taken to ensure that they provide relevant, representative assessments of the three behaviors. The objective of this study then, was to evaluate the degree to which this comprehensive scale construction process was successful. That is, the measures of health knowledge, and attitudes toward smoking, exercise, and diet were assessed in terms of their psychometric properties. The next chapter will delineate those procedures and methods that were employed to examine the reliability and validity of each of the measures.

METHOD

SUBJECTS

The CHA survey was administered to two groups of respondents. Group I consisted of 181 elementary school teachers who were participating in a preliminary workshop for the CHA curriculum program. Of this sample there were 46 males and 135 females. Group II consisted of 20 students in the University of Illinois graduate program of Health, Physical Education, and Recreation. These respondents were recruited from a variety of classes and participated on a volunteer basis. There were 9 males and 11 females in this group.

PROCEDURE

All respondents completed the CHA survey, consisting of the health knowledge questionnaire, the health attitude Likert scales, the health attitude Behavioral Consequences scales, and the health behavior inventories. Again, the survey was administered to Group I during a preliminary workshop session. A brief set of instructions was read to them indicating that the survey should be self-explanatory and that all results would be strictly confidential. The respondents in Group II completed the survey on an individual basis. A cover sheet was included, again indicating that the survey was self-explanatory and that the confidentiality of their responses would be preserved. They were also instructed to return the survey to a specified location when they had completed it.

ANALYSES

In order to provide a perspective with which to view the results of the psychometric assessments, a brief review of the primary analyses that were conducted is necessary. With the exception of the known-groups assessment, all analyses were conducted utilizing data from Group I only.

The reliability of the measures was estimated by assessing their internal consistency. Specifically, this involved calculating coefficient alpha for each measure (Cronbach, 1951).

The concurrent validity of each of the seven attitude scales was assessed by examining the relationship between scores on the attitude scale and scores on the relevant behavior measure. The two sets of scores were correlated to determine if there was a linear relationship between, for example, one's attitude toward physical exercise (as measured by the Likert method) and self-reports of one's behavior with regard to physical exercise.

Two approaches were taken to assess the construct validity of the measures. The known-groups technique was utilized to compare the responses of Group I versus those of Group II. It was believed that these groups were different in terms of health knowledge and attitudes because the school teachers were considered to have moderate knowledge of and average attitudes toward health; whereas the graduate students in health education were known to have above average knowledge of and more favorable attitudes toward a variety of health issues. For this reason, the groups were expected to receive significantly different scores on the knowledge measure and each of the attitude measures. To test for significance of mean differences between the groups on each of the measures, the "t" statistic was used.

The construct validity of the attitude measures was also assessed through the multitrait-multimethod matrix technique. As described earlier, this technique can be employed to examine the convergence between independent measures of the same attribute and discrimination between measures of different attributes. In this study, a matrix of intercorrelations of three theoretically unrelated attitudes (toward cigarette smoking, physical exercise, and a nutritional diet) as measured by two independent methods (Likert and Behavioral Consequences) was studied.

The results of each of the above analyses will be presented and discussed independently of one another. After the results of all of the analyses have been described, a synthesis of all of the assessments will be presented. This approach should allow for a systematic and thorough interpretation of the psychometric properties of each measure.

RESULTS AND DISCUSSION

RELIABILITY

The reliability of the knowledge and attitude measures was assessed by calculating coefficient alpha for each measure. Again, coefficient alpha is an index of the internal consistency, or homogeneity, of a measure. These reliability coefficients and a number of descriptive statistics are presented in Table 4.

Interpreting reliability coefficients should be done in terms of the purposes for which a particular measure will be used. In certain instances when important decisions are made with respect to specific test scores (e.g., academic admissions testing), reliabilities of .90 or higher are necessary (Nunnally, 1978). However, when measures are being utilized for research purposes, a reliability coefficient of approximately .60 is adequate. As indicated in Table 4, the coefficient alphas for the knowledge and attitude measures exceed that level. Indeed, with the exception of the Likert diet scale, the alphas are a good deal higher than the level necessary.

This evidence of measurement reliability is particularly encouraging in view of the limited number of items on the attitude scales. In fact, the levels of reliability for these brief measures equal or exceed those obtained with much longer instruments (see, for example, Solleder, 1979).

TABLE 4

RELIABILITY COEFFICIENTS AND DESCRIPTIVE STATISTICS FOR KNOWLEDGE AND ATTITUDE MEASURES

Measure	Number of Items	Possible Range	Obtained Range	Mean	S.D.	Coeff. Alpha
Health Knowledge	50	0 to 50	11 to 44	26.76	6.67	.76
Attitudes (Likert)						
Smoking	10	10 to 70	10 to 62	34.19	10.57	.71
Exercise	10	10 to 70	26 to 68	49.74	8.96	.71
Diet	10	10 to 70	26 to 69	49.86	8.26	.65
Attitudes (Behavioral Consequences)						
Smoking	11	-99 to +99	-99 to +51	-30.18	27.25	.78
Nonsmoking	11	-99 to +99	-99 to +99	35.77	30.48	.85
Exercise	11	-99 to +99	-99 to +99	41.92	28.41	.86
Diet	11	-99 to +99	-99 to +99	33.43	25.90	.83

The distribution of scores is also an important factor in determining the adequacy of a measuring instrument (Nunnally, 1978). Reference to the scale means, standard deviations, and ranges in Table 4 reveals that ceiling and floor effects were not characteristics of these scales. That is, there is sufficient room for scores to occur above and below the mean of each measure. In addition, a high level of sensitivity is indicated for each measure by the broad range of obtained scores. That a measure is sensitive means that it is capable of distinguishing among levels of a characteristic to a very specific or exact degree (Selltiz et al., 1976). For example, a very insensitive measure of attitudes toward smoking might distinguish only two postions: pro-smoking and anti-smoking. On the other hand, the Behavioral Consequences smoking scale provides a highly sensitive measure of attitudes toward smoking; respondents received scores ranging from extremely anti-smoking (-99) to very pro-smoking (+51). This is a range of 150 different positions with respect to one's attitude toward smoking obtained by this sample of respondents. The potential range of positions on the Behavioral Consequences scales is 199. Having such sensitive measures is essential when the objective is to monitor small or gradual shifts in a particular characteristic.

In summary it can be said that the knowledge and attitude measures possess a more than adequate degree of internal consistency. Moreover, the distributional characteristics of the knowledge questionnaire indicates that it is highly sensitive and not subject to ceiling or floor effects. It is evident then that these measures are relatively uncontaminated by random measurement error and thus produce observed scores that are good estimates of the underlying, unobservable "true" scores. Whether the observed scores are representative of the variables of interest is a matter of validity.

CONCURRENT VALIDITY

The foremost consideration in attempting to establish the concurrent validity of a measure is the selection of an appropriate criterion measure (Vincent, 1970). The choice of criterion measures in this study was based on a conceptual relationship between attitudes and behavior. Fishbein and Ajzen (1975) have noted that an appropriately constructed measure of behavior can serve as a valid indicant of attitude toward the behavior. Specifically, the behavior measure should be a multiple-act, repeated observation assessment. A measure of a single behavior at one point in time is not likely to be very related to a measure of attitudes toward the behavior (Wicker, 1969). On the other hand, a global measure of behavior, one based on multiple acts at different times, is more apt to be correlated with attitudes toward the behavior. This argument is based on the notion that though a person possesses favorable attitudes toward a given set of behaviors (e.g., exercising), he or she may not perform a single act of that behavior in a particular situation (e.g., calisthenics on Monday). However, that person is likely to perform one or more other behaviors (e.g., swimming, racketball) over a period of time. Therefore, a criterion measure that is based on observations of different behaviors at different points in time represents a general measure of attitude toward the behavior in question and can be used to assess the concurrent validity of the attitude measure (Fishbein and Ajzen, 1975).

Though the criterion measures utilized in this study were obtained at only one point in time, they are actually measures of behavior over time. Furthermore, they are indicators of multiple acts with respect to each behavioral domain. For example, with respect to exercise behavior, respondents were requested to indicate the amount of time they typically spent participating in 26 different physical activities. The 26 different activities provides the multiple-act criterion, and repeated observation is achieved by having the respondents recall their behavior over a period of time. (See Appendix A.) Similarly, the diet behavior inventory includes multiple acts (20 different types of food) and repeated observations. With regard to cigarette smoking the situation is slightly different. Though the respondents did not indicate multiple acts directly, the assessment of current status is a measure of global behavior. That is, the question is not phrased: "Did you smoke a king-size, filter-tip cigarette today?". Rather, respondents indicated whether they "ususlly smoke cigarettes" or "smoke cigarettes once in awhile." Thus, multiple acts are assessed indirectly and are based on a broad period of time. Tables 5 and 6 provide descriptive data regarding each of the behavior measures.

In order to examine the concurrent validity of the attitude measures, the correlation between scores on those measures with scores on the relevant behavior measures were calculated. So a correlation between smoking attitudes and smoking behavior could be calculated, the

TABLE 5

DESCRIPTIVE STATISTICS FOR EXERCISE AND DIET HEALTH BEHAVIOR MEASURES

Measure	Range	Mean	S.D.
Exercise Health Inventory	0 to 184	42.34	33.00
Diet Health Inventory	-99 to +77	-2.56	29.18

TABLE 6

RESPONDENT'S CURRENT SMOKING STATUS

Current Status

	1	2	3	4	5
	Usually smoke every day	Smoke once in awhile, but not every day	Used to smoke every day, but not now	Have smoked a few times, but not now	Never smoked
N	40	13	40	28	60
(%)	(22)	(7)	(22)	(16)	(33)

nominal level smoking status variable was dichotomized; categories one and two being combined to indicate current smokers, categories three, four, and five being combined to indicate current nonsmokers. Thus, the correlation between attitudes with regard to smoking and current smoking is a point-biserial correlation (Guilford and Fruchter, 1973). The concurrent validity correlations are presented in Table 7.

There is evidence for the concurrent validity of some of the attitude measures in Table 7. Each of the correlations between current smoking status and smoking attitudes are significant at the .001 level. The correlation between smoking status and the smoking Likert scale (.41) is somewhat higher than either of the other two measures of smoking attitudes -- Behavioral Consequences smoking scale (.33) and Behavioral Consequences nonsmoking scale (-.29). The evidence for the validity of the exercise Likert scale ($r=.19$) and the Likert diet scale ($r=.20$) is weaker though the correlations are statistically significant ($p<.01$). Neither the exercise or diet Behavioral Consequences scales correlate significantly with the relevant behavior measures.

Thus, this assessment of validity indicates that each of the smoking attitude measures possess a moderate degree of validity, the exercise and diet Likert scales possess a weaker amount of validity, and virtually no evidence is given for the validity of the exercise and diet Behavioral Consequences scales. This evidence is not concluseve however. As pointed out earlier, individual assessments of validity do not determine that a given indicator should be absolutely accepted or rejected as valid but only increase (or decrease) the likelihood of val-

TABLE 7

CONCURRENT VALIDITY
CORRELATION COEFFICIENTS

Smoking (N=172)

	Likert Smoking Scale	Smoking Consequences Behavioral Scale	Nonsmoking Consequences Behavioral Scale
Current Smoking Status	.41**	.33**	-.29**

Exercise (N=161)

	Exercise Likert Scale	Exercise Behavioral Consequences Scale
Exercise Behavior Score	.19*	.08

Diet (N=150)

	Diet Likert Scale	Diet Behavioral Consequences Scale
Diet Behavior Score	.20*	-.02

* $p<.01$
** $p<.001$

idity (Curtis and Jackson, 1962). A more thorough evaluation depends on additional assessments of validity, the results of which will be discussed below.

Though the results of the concurrent validity assessment are less than favorable for some of the measures, they are not without alternative interpretations. Low correlations between the attitude measures and the behavior measures may be due to the inadequacy of the behavior measures, not the attitude measures. That is, the behavior or "criterion" measures may themselves be unreliable or invalid. They were not thoroughly assessed and thus firm conclusions cannot be drawn regarding other measures which employ these as criteria. In addition, they are not ideal multiple-act criterion measures as suggested by Fishbein and Ajzen (1975). They are only approximations. The only evidence regarding the validity of the behavior measures pertains to the manner in which they were developed. It can be said that the development process did ensure a good amount of content validity for these measures. That is, they are representative samples of the relevant content domains.

So there is only a small degree of evidence that the criterion measures used in the concurrent validity assessment are valid indicators of the concepts of interest. This is not an unusual circumstance (Selltiz et al., 1976). The best solution, as indicated before, is to keep in mind the limitations of the assessment and supplement it with additional information. The following section will describe the results of two additional assessments of validity.

CONSTRUCT VALIDITY

In order to examine the construct validity of the knowledge and attitude measures, two approaches were utilized -- the known-groups technique and the multitrait-multimethod matrix technique. The known-groups assessment involved comparing the mean responses of Group I (a sample of elementary school teachers) to Group II (a sample of graduate students specializing in health, physical education, and recreation) on the health knowledge measure, the three Likert scales, and the four Behavioral Consequences scales. Again, it was believed that Group II would receive significantly higher scores on the knowledge test, have significantly more favorable attitudes toward nonsmoking, exercise, and a nutritional diet, and have significantly less favorable attitudes toward smoking. To test if there was a significant difference between the groups' mean scores on each measure, the $\underline{t}$ statistic was used. Separate variance estimates were used because homogeneity of variance tests confirmed the belief that the scores of Group II would be more homogeneous than those for Group I (Winer, 1971). The results of the known groups validity assessment are presented in Table 8.

With the exception of the Likert diet scale, comparisons of the groups' means reveal that the measures differentiated the groups as predicted. More specifically, differences between the groups' means on all of the measures except Likert diet and exercise were statistically significant at the .05 level or better.

Thus, there is confirmation regarding the construct validity of the knowledge test, the Likert smoking scale, and each of the Behavioral

TABLE 8

KNOWN GROUPS VALIDITY ASSESSMENT

Measure	Group I Mean	S.D.	N	Group II Mean	S.D.	N	t
Health Knowledge	26.76	6.67	181	34.40	5.59	20	5.68***
Attitudes (Likert)							
Smoking	34.19	10.57	177	28.65	8.02	20	-2.82**
Exercise	49.74	8.96	179	51.55	5.90	20	1.22
Diet	49.86	8.26	178	49.30	4.69	20	-.46
Attitudes (Behavioral Consequences)							
Smoking	-30.18	27.25	177	-53.05	25.44	20	-3.78***
Nonsmoking	35.77	30.48	177	59.45	22.68	20	4.26***
Exercise	41.92	28.41	177	52.70	15.55	20	2.64*
Diet	33.43	25.90	175	43.30	19.49	20	2.07*

* $p<.05$

** $p<.01$

*** $p<.001$

Consequences scales. It is apparent however that the Likert exercise scale and the Likert diet scale possess a lesser degree of validity. Though the mean difference on the Likert exercise scale was not statistically significant, it was in the predicted direction. The small number of respondents in Group II suggests that repeating the study with a larger group might well substantiate this trend. The results for the Likert diet scale are much less favorable. The finding that members of Group II tended to express less favorable nutrition attitudes may be interpreted as an indication that this measure may not be a valid indicator. Alternatively, the assumption that the comparison group (Group II) would be more likely than the teachers to adhere to a healthful diet may have been unfounded. A variety of information testifies to this interpretation. First, Group II in fact received lower scores on the diet inventory scale than did Group I. Thus, this measure of behavior indicates that the assumption of known groups with regard to diet might have been inaccurate. Given that the members of Group II were participating in an educational program that emphasizes exercise and not diet, this interpretation becomes even more tenable. In addition, Group II's scores on the Behavioral Consequences measures are consistent with the belief that they have less favorable attitudes toward maintaining a nutritional diet than they do toward exercise; their mean scores on these measures were 43.30 and 52.70 respectively. So a more plausible interpretation regarding the construct validity of the Likert diet scale is not that it is invalid, but rather the known group chosen for this study provided an inappropriate criterion against which the validity of the diet measures should be judged. Administration of these measures to a different known group, for example, nutritionists and dieticians, would furnish a more effective evaluation of their validity.

The known-groups technique applied in this study has provided strong evidence regarding the validity of a number of the measures on the CHA survey. The construct validity of the Likert diet and exercise scales was not confirmed. Further study of these two scales employing different known groups and a larger number of respondents should be conducted to furnish more conclusive evidence regarding their validity. The positive evidence for the knowledge test, Likert smoking scale, and Behavioral Consequences scales also needs to be verified through additional study. One such method that can be used to supply additional evidence regarding validity is the multitrait-multimethod matrix technique.

The multitrait-multimethod technique was employed in this study to assess the convergent and discriminant validity of the attitude measures. Specifically the interrelationships of three traits (attitudes toward smoking, exercise, and a nutritional diet) as measured by two methods (Likert and Behavioral Consequences) were examined. The multitrait-multimethod correlations are presented in Table 9.

The correlations in the main diagonal of the matrix are actually the reliability estimates of each measure, i.e., coefficient alpha. Recalling that the square root of the reliability of a measure places an upper limit on the possible correlation of the measure with any other

TABLE 9

MULTITRAIT-MULTIMETHOD
CORRELATION MATRIX

		Likert Method			Behavioral Consequences Method		
		Smoking (1)	Exercise (2)	Diet (3)	Smoking (4)	Exercise (5)	Diet (6)
L	S (1)	.71					
	E (2)	-.34	.71				
	D (3)	-.35	.38	.65			
BC	S (4)	.39	-.28	-.24	.78		
	E (5)	-.19	.39	.23	-.56	.86	
	D (6)	-.25	.33	.26	-.58	.76	.83

A,B = Different-trait, different-method triangles
C,D = Different-trait, same-method triangles

measure, it can be seen that the unreliability of the measures is not a major problem in this matrix. The lowest estimate is for the Likert diet scale (.65) and the square root of this coefficient is approximately .81. The Likert diet scale therefore, cannot be expected to correlate any higher than .81 with another measure. The other reliability estimates are somewhat better, with square roots ranging from .84 for the Likert smoking and exercise scales to .93 for the Behavioral Consequences exercise scale. So, the restrictions due to unreliability are variable across the different measures. Though this does not appear to be a significant problem, the information can be used nonethleless to aide in the interpretation of the other coefficients in the matrix.

The validity coefficients in Table 9 are the correlations between the same-trait, different-method variables, i.e., r_{14}, r_{25}, r_{36}. The validity coefficients for smoking and exercise as measured by the two different methods are both .39, while the coefficient for diet as measured by each method is .26. According to Campbell and Fiske's first criterion, these validity coefficients should be large enough. (See Table 2.) All are significantly different from zero at the .001 level.

The second criterion is that each validity coefficient should be larger (in absolute value) than all of the different-trait, different-method correlations that are in the same row or column as the validity coefficient. This criterion is met for the validity coefficient for smoking as measured by the Likert and Behavioral Consequences methods (.39 > -.19, -.25, -.28, -.24) and for the validity coefficient for exercise as measured by the two methods (.39 > -.19, .33, -.28, .23). However, the validity coefficient for diet does not fully meet this criterion (.26 > -.25, .23, -.24; but, .26 < .33). So with the exception of the measures of attitudes toward a nutritional diet, the Likert and Behavioral Consequences scales meet the criterion that the validity coefficients be larger than the corresponding different-trait, different-method correlations.

The third criterion is that each validity coefficient should be larger than the different-trait, same-method correlations which involve the same trait as the validity coefficient. This criterion is only met for the smoking and exercise validity coefficients, and then only with regard to the different-trait, Likert-method correlations. The diet validity coefficient does not meet the criterion at all, and none of the three validity coefficients are greater than any of the three different-trait, Behavioral Consequences-method correlations. Again the validity of the diet measures is suspect, and the validity of the other measures is brought into doubt as well.

The fourth criterion is that the same pattern of correlations should be evidenced within each of the triangles. This criterion is met in three of the four triangles. In triangles A, C, and D, the correlation is highest between attitudes toward exercise and diet, then between smoking and diet, and lastly between smoking and exercise. In triangle B, the pattern of correlations is just the opposite. Because this different pattern occurs in a different-method triangle it provides no evidence regarding the superiority of one method over the other. That is, the different-trait, same-method pattern of correlations is the same for each method. So apparently, the different methods are comparable to the extent that they result in similar patterns of correlations among different traits. However, the inconsistent pattern in triangle B points to a lack of validity in each method. The fact that two of the correlations in that triangle involve measures of attitudes toward a nutritional diet, indicates that the problem may again be with regard to measures of this trait as opposed to being a problem with the methods themselves.

A definitive evaluation of any multitrait-multimethod matrix is difficult (Cummings et al., 1978). The differences among correlations in this matrix were in many cases not very substantial. Sampling errors could easily alter many of the patterns of correlations that were pointed out. Also, there is a good deal of inconsistency with respect to the four criteria used to assess the matrix. Such inconsistency might be taken as an indicator of the invalidity of all the measures employed (Campbell and Fiske, 1959). On the other hand, a more critical look at the matrix in Table 9 may reveal a more favorable outcome.

The analysis of the validity coefficients with respect to the first criterion suggested that the two measures of attitudes toward diet were not as valid as the other measures, though the coefficient was statistically significant from zero. This was substantiated by the fact that the validity coefficent for diet was not larger than the corresponding different-trait, different-method correlations (Criterion 2). In addition, this coefficient was not larger than the different-trait, Likert-method correlations, while the other two validity coefficients did meet this portion of the third criterion. And again, two of the correlations in the triangle with the inconsistent pattern were coefficients involving Likert measures of diet attitudes (Criterion 4). All of these results point to the invalidity of the measures of attitudes toward a nutritional diet, particualrly the Likert scale.

The inadequacy of either the Likert or Behavioral Consequences measures of attitudes toward a nutritional diet may be due more to the nature of the concept than to the methods of measurement (Nunnnally, 1967). That is, people's attitudes toward maintaining a nutritional diet may not be as well-formulated as are their attitudes toward smoking cigarettes, for example. This hypothesis is supported by the fact that the reliability coefficient for the Likert diet scale is lower than any of the other coefficients. This low reliability estimate indicates a lack of consistency in the responses made to the items. This could certainly be due to a poor measuring instrument, but it may also be that people do not have a well-defined attitude toward maintaining a nutritional diet. The conclusion that the multitrait-multimethod matrix provides little evidence for the construct validity of the diet measures must remain. However, the reason for invalidity may lie more with the nature of the construct, than with the methods.

The evidence for the validity of the measures of attitudes toward smoking and exercise was better than for the measures of diet attitudes, but it was not entirely supportive. In particular, the different-trait, Behavioral Consequences-method correlations were substantially higher than the validity coefficients for smoking and exercise. This would indicate that the Behavioral Consequences method lacks discriminant validity. This finding is not surprising given the manner in which the scale is constructed and total scores computed. At least eight of the consequences utilized on each scale are shared by one of the other two scales. Since these consequences are rated only once in terms of their importance, this part of the product (instrumentality rating X importance rating) will be the same in each scale that contains that particular consequence. Of course the instrumentality rating of the consequences may differ across the different behaviors but at least eight importance ratings will be shared by one scale or the other. It is likely then, that the Behavioral Consequences scales as they are designed on the CHA survey are going to be intercorrelated to an extent greater than what generally occurs due to an overlap of trait variation only. Thus their lack of discriminant validity should not be taken as convincing evidence that either the Behavioral Consequences or Likert measures of attitudes toward smoking and exercise are invalid. It merely suggests that a firm conclusion cannot be drawn in total favor of either validity or invalidity.

To sum up, it has been shown that the multitrait-multimethod matrix as employed in this study indicates that the Likert and Behavioral Consequences measures of attitudes toward smoking and exercise possess a moderate degree of convergent validity. The evidence regarding the convergent validity of both methods of measuring of attitudes toward a nutritional diet was less favorable. As is often the case, the discriminant validity of the measures was lacking (Campbell and Fiske, 1959). So while the different methods achieved similar results while measuring the same traits in two of three cases, they were not so successful at achieving dissimilar results while measuring different traits. One viable explanation for the lack of discriminant validity, is that the traits utilized may not be maximally different. In fact, attitudes toward smoking, exercise and diet may all be part of an overall "health attitude." In this case, substantial discriminant validity would not be expected.

This study has employed a number of different approaches and techniques to assess the psychometric properties of an instrument designed to measure adults' knowledge and attitudes regarding a variety of health issues. Analyses of the data have resulted in the gathering of a considerable amount of evidence, especially with regard to the validity of these measures. Much of the evidence is favorable, some of it is unfavorable. It has been mentioned a number of times that the best way to evaluate the adequacy of measuring instruments is to gather such a variety of evidence. The final section of this paper summarizes the findings of this study and attempt to integrate the evidence that has been furnished by each of the analyses. Where appropriate, conclusions are drawn regarding the psychometric quality of each measure. Finally, recommendations are made regarding future studies that might be conducted to substantiate the findings of the current study.

CONCLUSIONS

The purpose of this study was to assess the reliability and validity of an evaluation instrument designed to measure adults' health knowledge and attitudes. The outcomes of a number of analyses are presented in Table 10. This summary of findings can be used as an aide in interpreting the myriad of evidence that was supplied by the various assessments. Each measure will be discussed in turn with respect to the outcome of each assessment.

The health knowledge instrument was designed to be a measure of general health knowledge and knowledge regarding a number of specific issues relevant to the CHA curriculum program. The reliability of the knowledge test was found to be good and it exhibited excellent distributional characteristics. The known groups analysis showed that the test could significantly discriminate between a group of school teachers and a group of health education graduate students. Though only assessed with one method it is apparent that the knowledge test possesses strong construct validity and can be used as an indicator of general health knowledge as it relates to cardiovascular fitness.

The Likert smoking scale showed an adequate degree of reliability, especially given the limited number of items. In terms of concurrent validity the scale was found to correlate relatively well with a self-report measure of current smoking status. The construct validity

TABLE 10

SUMMARY OF FINDINGS

Measure	Assessment			
		Validity		
	Reliability	Concurrent	Known Groups	Multitrait-Multimethod
Health Knowledge	++		++	
Attitudes (Likert)				
Smoking	++	+	++	+
Exercise	++	+	-	+
Diet	+	+	?	-
Attitudes (Behavioral Consequences)				
Smoking	++	+	++	+
Nonsmoking	++	+	++	
Exercise	++	-	+	+
Diet	++	-	?	-

++ Strong evidence in favor.
+ Weak evidence in favor.
- Evidence not in favor.
? Inconclusive finding.
Blank indicates measure was not assessed with that technique.

of this scale was strongly demonstrated through the known-groups method and moderately demonstrated in the multitrait-multimethod matrix. In the latter case, the Likert smoking scale was found to a have a good amount of convergent validity, but only a small amount of discriminant validity. Thus, overall, there is a good deal of evidence that the Likert smoking scale is a reliable and valid indicator of adults' attitudes toward smoking cigarettes.

As with the Likert smoking scale, the Likert exercise scale exhibited an adequate degree of reliability. The evidence for concurrent validity was rather weak. Though statistically significant, the Likert exercise scale was only somewhat correlated with a composite self-report measure of exercise behavior. In terms of construct validity the scale did not discriminate between the known groups but it did exhibit a moderate amount of convergent validity in the multitrait-multimethod matrix. Overall, there is an indication that the Likert exercise scale has some validity but it ought to be used with caution until this tendency can be substantiated. As pointed out earlier, administering this scale, as well as the others, to a larger known group might provide more conclusive information.

The Likert diet scale exhibited only a fair degree of reliability. Furthermore the only evidence for validity was a moderate correlation with a composite self-report measure of nutritional behavior. The low degree of validity may be due in part to the low reliability. And the low reliability may be a function of the nature of attitudes toward maintaining a nutritional diet. As discussed earlier, it may be that people do not have well-defined attitudes regarding nutrition. One possible remedy that could be used to improve this scale would be to add more items. Provided that they do not decrease the average interitem correlation, it is likely that the scale's releability will increase and accordingly the scale's validity may improve. However until such evidence is presented it is recommended that the Likert diet scale not be used to assess adults' attitudes toward maintaining a nutritional diet.

The reliability of the Behavioral Consequences smoking scale was found to be quite adequate. The concurrent validity assessment showed that this scale correlated well with a measure of current smoking status. The known-groups assessment provided strong evidence regarding the construct validity of this scale. And the multitrait-multimentod matrix demonstrated the convergent validity of the Behavioral Consequences smoking scale, while evidence for its discriminant validity was lacking. So overall it is apparent that this scale designed to measure attitudes toward smoking possesses adequate reliability and validity.

The results of the analyses of the Behavioral Consequences nonsmoking scale were comparable to those for the Behavioral Consequences smoking scale. The reliability of the scale was somewhat higher and the results of the concurrent validity and known groups assessments suggest that this scale is a useful instrument for measuring adults' attitudes toward nonsmoking.

The Behavioral Consequences exercise scale was found to have the highest degree of reliability of all of the measures. The assessment of concurrent validity was not so favorable as the scale was uncorrelated with the exercise behavior measure. A fair amount of construct validity was evidenced with the known-groups assessment and in the multitrait-multimethod matrix. As with a number of the other measures, convergent validity was exhibited, but not discriminant validity. In sum it can be said that more evidence should be gathered regarding the validity of this scale, and in its present form it should be used with caution.

The only favorable evidence for the Behavioral Consequences diet scale is that it was fairly reliable. It did not correlate with a measure of nutritional behavior, nor did it exhibit convergent validity in the multitrait-multimethod matrix. Though this scale did distinguish between the group of teachers and the known group of health education graduate students, this comparison group of respondents may have been an inappropriate criterion in that the respondents reported maintaining a less nutritional diet than the teachers. Thus, the evidence regarding this scale is ambiguous at best. Further study of the Behavioral Consequences diet scale might employ a more suitable known group and a more definitive conclusion might be drawn regarding its usefulness as a measure of attitudes toward a nutritional diet. Until further evidence is provided, the use of this scale is not recommended.

In summary, the results of a number of psychometric analyses have indicated that some of the instruments on the CHA health survey can be used as reliable and valid measures, others should be used with caution, and still others should not be used in their present form. All of the measures had an adequate degree of reliability, though the estimate for the Likert diet scale was somewhat lower than those for the other scales. However, the degree of validity varied across the measures. Specifically, the health knowledge questionnaire, the Likert smoking scale, the Behavioral Consequences smoking scale, and the Behavioral Consequences nonsmoking scale exhibited consistently good levels of validity for each assessment employed. These measures appear to be psychometrically sound. The Likert exercise scale and the Behavioral Consequences exercise scale tended toward good levels of validity, but this was not consistent across the different assessments. These two measures should be used with caution. To use with caution means that any results obtained with the use of these measures ought to be interpreted in the light of the inconclusive evidence of validity. Additional measures of these attitudes might be considered as a means of substantiating obtained results, at least until it has been concluded that the measures possess an adequate degree of validity. The Likert diet scale and the Behavioral Consequences diet scale showed little evidence of validity on any of the assessments. These measures need further refinement, which may include the addition of new items and rewording of existing items. This of course would necessitate additional pilot testing, item analyses, and further assessments of reliability and validity.

The above conclusions point out an interesting finding. The relative degree of validity across the three concepts was the same for each of the measurement techniques. That is, both the Likert and Behavioral Consequences measures were most valid for smoking, then for exercise, then for diet. It is possible that this pattern of results is due to the nature of the constructs. It was noted earlier that people may not have a well-defined attitude toward a nutritional diet. On the other hand, it is likely that people have fairly well-formulated attitudes toward cigarette smoking. Attitudes toward physical exercise may lie somewhere in between. This does not imply that that people's attitudes toward the three issues are necessarily different in direction or degree. It only means that the more defined and formulated a person's attitude is toward a particular object, the more likely it is that a measure of that attitude will be valid and reliable. A measure of an ill-defined attribute will by necessity be unstable and therefore less valid.

Finally, it must be emphasized that the validation process does not stop here. Even those measures that have been indicated as being adequately reliable and valid need to be periodically reinvestigated. Scale validity may change from time to time and from sample to sample. Thus it becomes necessary to ensure that what is now a reliable and valid indicator of a theoretical construct continues to be so in the future.

REFERENCES

Ajzen, I., & Fishbein, M. _Understanding attitudes and predicting social behavior_. Englewood Cliffs, NJ: Prentice-Hall, 1980.

Althauser, R., & Heberlein, T. Validity and the multitrait-multimethod matrix. In E. Borgatta & G. Bohrnstedt (Eds.), _Sociological methodology_. San Francisco: Jossey-Bass, 1970.

Bickman, L. Personal communication. November, 1980.

Bohrnstedt, G. Reliability and validity assessment in attitude measurement. In G. Summers (Ed.), _Attitude measurement_. Chicago: Rand McNallly, 1970.

Brown, W. Some experimental results in the correlation of mental abilities. _British Jounal of Psychology_, 1910, 3, 296-322.

Campbell, D., & Fiske, D. Convergent and discriminant validation by the multitrait-multimethod matrix. _Psychological Bulletin_, 1959, 56, 81-105.

Campbell D., & Stanley, J. _Experimental and quasi-experimental designs for research_. Chicago: Rand McNally, 1966.

Carmines, E., & Zeller, R. _Reliability and validity assessment_. Beverly Hills: Sage, 1979.

Chaplin, J. _Dictionary of psychology_. New York: Dell, 1975.

Chicago Heart Association. _The heart saver eating style_. Chicago: Author, 1977.

Crano, W., & Brewer, M. _Principles of research in social psychology_. New York: McGraw-Hill, 1973.

Crewe, N. Comparison of factor analytic and empirical scales. _Proceedings of the 75th Annual Convention of the American Psychological Association_, 1967, 367-368.

Cronbach, L. Coefficient alpha and the internal structure of tests. _Psychometrika_, 1951, 16, 296-334.

Cronbach, L. Test validation. In R. Thorndike (Ed.), _Educational measurement_. Washington: American Council on Education, 1971.

Cronbach, L., & Meehl, P. Construct validity in psychological tests. *Psychological Bulletin*, 1955, 52, 281-302.

Cummings, K., Jette, A., & Rosenstock, I. Construct validation of the Health Belief Model. *Health Education Monographs*, 1978, 6, 394-405.

Curtis, R., & Jackson, E. Multiple indicators in survey research. *The American Journal of Sociology*, 1962, 195-204.

Edwards, J. An ordered list of values. *Experimental Social Psychology Bulletin*, Ohio State University, Mimeo, 1967.

Fischer, E. Attitudes toward limiting family size: Convergence of factorial and known-groups validity approaches. *Proceedings of the 79th Annual Convention of the American Psychological Association*, 1971, 263-264.

Fishbein, M. Attitude and the prediction of behavior. In M. Fishbein (Ed.), *Readings in attitude theory and measurement*. New York: Wiley, 1967.

Fishbein, M., & Ajzen, I. *Belief, attitude, intention and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley, 1975.

Guilford, J., & Fruchter, B. *Fundamental statistics in psychology and education*. New York: McGraw, 1973.

Kuder, G., & Richardson, M. The theory of the estimation of test reliability. *Psychometrika*, 1937, 2, 151-160.

Likert, R. A technique for the measurement of attitudes. *Archives of Psychology*, 1932, No. 140.

Morehouse, L., & Miller, A. *The physiology of exercise*. St. Louis: Mosby, 1976.

Nunnally, J. *Educational measurement and evaluation*. New York: McGraw-Hill, 1964.

Nunnally, J. *Psychometirc theory*. New York:McGraw-Hill, 1978.

Nunnally, J., & Durham, R. Validity, reliability, and special problems of measurement in evaluation research. In E. Struening & M. Guttentag (Eds.), *Handbook of evaluation research*, Beverly Hills: Sage, 1975.

Parcel, G. Development of an instrument to measure attitudes toward the personal use of premarital contraception. *The Journal of School Health*, 1975, 45, 157-160.

Peak, H. Attitude and motivation. In M. Jones (Ed.), Nebraska Symposium on Motivation. Vol. 3, Lincoln, Nebraska: University of Nebraska Press, 149-188, 1955.

Rosenberg, M. Cognitive structure and attitudinal affect. Journal of Abnormal and Social Psychology, 1956,50, 295-302.

Selltiz, C., Wrightsman, L., & Cook, S. Research methods in social relations. New York: Holt, Rinehart and Winston, 1976.

Solledor, M. Evaluation instruments in health education. Washington, DC: Associates for the Advancement of Health Education, 1979.

Spearman, C. Correlation calculated from faulty data. British Journal of Psychology, 1910, 3, 271-295.

Sullivan, J., & Feldman, S. Multiple indicators: An introduction. Beverly Hills: Sage, 1979.

Swanson, J. Second thoughts on knowledge and attitude effects upon behavior. Journal of School Health, 1972, 42, 363-365.

Vincent, R. A scale to measure attitude toward smoking marihuana. The Journal of School Health, 1970, 454-456.

Wicker, A. Attitudes vs. actions: The relationship of verbal and overt behavioral responses to attitude objects. The Journal of Social Issues, 1969, 25, 41-78.

Winer, B. Statistical principles in experimental design. New York: McGraw-Hill, 1971.

Appendix A

HEALTH KNOWLEDGE QUESTIONNAIRE

The following are fifty questions concerning various health issues and facts. Please mark one answer for each question by circling your choice on the questionnaire.

1. Blood pressure is measured by the pressure

   a. in the valves
   b. on the walls of the veins
   c. in the vena cava
   d. on the walls of the arteries

2. What kind of fat is most likely to raise the level of blood cholesterol?

   a. saturated fat
   b. polyunsaturated fat
   c. monounsaturated fat
   d. vegetable fat

3. Which of the following is not true about cigarette smokers, compared with nonsmokers?

   a. smokers have a higher occupational level
   b. smokers are more likely to have accidents
   c. smokers are more likely to die in fires
   d. smokers have slower reflexes

4. Which set of risk factors is most likely to lead to coronary artery disease?

   a. high blood pressure, leukemia, obesity
   b. rheumatism, diabetes, high blood pressure
   c. increased cholesterol, smoking, a family history of coronary diseases
   d. increased cholesterol, family history of coronary disease, emphysema

5. If you strained your quadriceps you would have strained a _____ muscle.

   a. shoulder
   b. arm
   c. leg
   d. back

6. A resting heart rate of 150 beats per minute is:

   a. average for adult females
   b. average for adult males
   c. average for adult males and females
   d. above average for adult males

7. What happens to respiration as the result of regular exercise?

   a. blood vessels in the lungs contract
   b. vlood vessels in the lungs dilate
   c. one can take in more air with each breath
   d. lungs become more porous

8. A broken bone that breaks through the skin is called a

   a. compound fracture
   b. greenstick fracture
   c. stress fracture
   d. double fracture

9. Which of the following statements about the size of muscle fibers is true?

   a. they vary in size throughout the body
   b. they are about the size of a hair
   c. they change in size with age
   d. they vary greatly in size from one person to another

10. The mandible is a

   a. bone
   b. muscle
   c. nerve
   d. specific rib of the rib cage

11. When the blood leaves the heart to carry oxygen to the rest of the body, what blood vessel does it travel through?

   a. pulmonary artery
   b. aorta
   c. pulmonary vein
   d. vena cava

12. Coronary artery disease means that _____ is developing in the arteries of the heart.

    a. a blood clot
    b. protein deposits
    c. sugar deposits
    d. atherosclerosis

13. Of the four main types of blood, which is the most rare type of blood, the one that the fewest people have?

    a. O
    b. AB
    c. A
    d. B

14. After eating a meal, the food passes through the small intestine in about

    a. 12 hours
    b. 7 hours
    c. 5 hours
    d. 2 hours

15. A heart attack occurs when the blood supply is cut off in the

    a. vena cava
    b. pulmonary artery
    c. aorta
    d. coronary arteries

16. How many bones does the average person have?

    a. 98
    b. 206
    c. 451
    d. 1021

17. Blood moves from the right ventricle to the

    a. body
    b. left ventricle
    c. right clavicle
    d. lungs

18. Which of the following is a good source of protein?

    a. wheat bread
    b. cheese
    c. artichokes
    d. celery

19. Which type of blood vessel allows oxygen and nutrients to pass to the body and wastes to enter the blood?

    a. capillaries
    b. veins
    c. arteries
    d. arterioles

20. Most of the important nutrients are removed from food while the food is in the

    a. small intestine
    b. large intestine
    c. stomach
    d. duodenum

21. Which of the following chemicals in cigarette smoke is probably most responsible for causing lung cancer?

    a. nicotine
    b. tar
    c. carbon monoxide
    d. lead

22. A person with tension headaches is asked to relax his or her muscles while a machine shows whether or not those muscles are being relaxed. This procedure is called

    a. psychocybernetics
    b. stress resistance
    c. autonomic control
    d. biofeedback

23. Which of the following diseases is <u>least</u> likely to be caused by cigarette smoking?

    a. tuberculosis
    b. stroke
    c. heart damage
    d. hypertension

24. Which parts of the blood work to form blood clots and scabs if you get a cut?

    a. red blood cells
    b. white blood cells
    c. platelets
    d. plasma

25. Which of the following is the most basic building material for your body?

    a. vitamins
    b. carbohydrates
    c. protein
    d. plasma

26. Which of the following does <u>not</u> happen when people exercise?

    a. they have less energy
    b. reaction time improves
    c. they sleep better
    d. decrease in resting heart rate

27. When the blood leaves the heart to go to the lungs, what blood vessel does it travel through?

    a. aorta
    b. pulmonary vein
    c. pulmonary artery
    d. vena cava

28. Which of the following foods contains smoething from each of the five basic food groups (meat, fruits and vegetables, milk products, cereals, fats)?

    a. tomatoes stuffed with tuna
    b. sausage pizza
    c. turkey and dressing
    d. peanut butter and jelly sandwich

29. Which of the following statements if false?

    a. smoking and blood pressure are closely related
    b. cigarette smoking narrows the blood vessels in your skin
    c. cigarette smoking makes your heart beat faster
    d. if you are going to smoke, the best way to avoid cancer is to smoke a pipe

30. What causes fatigue during exercise?

    a. muscle tissue becomes porous
    b. waste products build up in the muscles
    c. muscle fibers contract at different rates
    d. muscle fibers change in size

31. The main air passage between the mouth and lungs is the

    a. esophagus
    b. bronchial tube
    c. trachea
    d. larynx

32. Which of the following meals would be the highest in fat content?

    a. chicken and cheddar cheese casserole
    b. tuna fish and cheese
    c. steak and eggs
    d. turkey and stuffing

33. Which part of the blood is yellowish in color as it carries food through the body?

    a. plasma
    b. hemoglobin
    c. white blood cells
    d. platelets

34. When sunshine falls on our skin, it helps our body to make

    a. vitamin A
    b. vitamin B
    c. vitamin C
    d. vitamin D

35. Which of the following statements is true?

    a. smoking affects males and females differently
    b. cigarette smoking enlarges the blood vessels
    c. cigarettes are equally harmful for adults and teenagers
    d. among middle aged men, the rate of heart attack is about the same for smokers and nonsmokers

36. Which of the following bones is found in your arm?

    a. radius
    b. tibia
    c. femus
    d. scapula

37. Which of the following medical instruments is used for looking inside the ears and nose?

    a. oscilloscope
    b. stethoscope
    c. sphygmomanometer
    d. otoscope

38. Enzymes break down proteins, carbohydrates, and fats into tiny particles called

    a. atoms
    b. molecules
    c. stomach acid
    d. bile

39. Which of the following is most true about heart beats?

    a. the smaller an animal is, the slower their heart beats
    b. the bigger an animal is, the slower their heart beats
    c. heart beat is about the same for animals of all sizes
    d. the heart of an adult beats faster than the heart of a baby

40. The main source of energy for your body is

    a. fats
    b. vitamins
    c. proteins
    d. carbohydrates

41. What is the cause of atherosclerosis?

    a. undetermined
    b. a virus
    c. a bacterium
    d. an enzyme deficiency

42. The liver aids digestion by making a substance that breaks down fat. That substance is called

    a. acid
    b. enzymes
    c. bile
    d. carbohydrates

43. How many different kinds of muscle tissue are there in your body?

a. 2
b. 3
c. 5
d. 6

44. Which of the following statements about blood pressure is <u>false</u>?

a. Your blood pressure tends to go down as you get older
b. blood pressure fluctuates continually
c. the tendency toward high blood pressure is often inherited
d. emotions and stress can temporarily raise blood pressure

45. When the blood is coming back from the body to the heart, which part of the heart does it go into first?

a. left atrium
b. left ventricle
c. right ventricle
d. right atrium

46. Which of the following foods is a <u>low</u> <u>fat</u> meat?

a. lamb
b. beef
c. pork
d. veal

47. Which of the following helps equalize air pressure in the nasal cavity?

a. sinuses
b. tympanic membrane
c. lungs
d. septum

48. Unlike the tars in cigarettes, nicotine has the greatest effect on:

a. respiration
b. energy level
c. circulation
d. relaxation

49. Which of the following acts directly as a cleaning and filtering system for the lungs?

    a. bronchi
    b. alveoli
    c. trachea
    d. cilia

50. Which of the following best describes the effect of nicotine?

    a. it dilates the blood vessels
    b. it constricts the blood vessels
    c. it dilates the veins, but not the arteries
    d. it constricts the veins, but not the arteries

LIKERT ATTITUDE SCALES

The following are 30 statements about various health related issues. Please indicate your degree of agreement or disagreement with each statement according to the following scale.

Strongly disagree			Neither agree nor disagree			Strongly agree
1	2	3	4	5	6	7

Mark your choice in the space to the left of each item.

A. SMOKING

_____1 Smoking can be stimulating and keep you going.

_____2 Smoking should be banned in all public places.

_____3 Under no circumstances should characters in TV programs and movies be shown smoking.

_____4 The manufacture and sale of cigarettes should be outlawed.

_____5 If people wouldn't smoke they could concentrate better on their work.

_____6 Smoking a few cigarettes a day really isn't bad for you.

_____7 The supposed dangers of smoking are not as great as the media tend to portray them.

_____8 The world would be a more pleasant place if people didn't smoke.

_____9 Smoking is an issue of freedom of choice.

_____10 All advertising for cigarettes should not be banned.

B. EXERCISE

_____11 People often get carried away with exercise programs and harm their bodies.

_____12 Thinking more clearly is a direct result of a regular program of exercise.

Strongly disagree			Neither agree nor disagree			Strongly agree
1	2	3	4	5	6	7

_____13 It is not necessary to exercise regularly to have an attractive body.

_____14 A regular program of exercise takes too much time.

_____15 Many forms of illness are the result of a lack of exercise.

_____16 Exercising is the best way of overcoming tension.

_____17 People would feel more energetic if they exercised regularly.

_____18 Good parents force their children to exercise.

_____19 A lot of people do not exercise regularly and it doesn't hurt them.

_____20 The benefits of exercise have been exaggerated lately.

C. DIET

_____21 Eating nutritious foods is the best way to make you feel healthy.

_____22 Lots of people eat non-nutritious foods and it doesn't bother them.

_____23 Empty calorie food (i.e., "junk food") is actually more nutritious than most people realize.

_____24 Not eating properly is the main reason people do not perform effectively at work.

_____25 Good parents do not allow their children to eat empty calorie foods (i.e., so-called "junk food").

_____26 The enjoyment of eating fattening foods makes up for any harm they might do.

_____27 Nutritious foods taste much better than non-nutritious "junk" foods.

_____28 Eating empty calorie food (so-called "junk food") is good for people's morale.

Strongly disagree			Neither agree nor disagree			Strongly agree
1	2	3	4	5	6	7

_____29 You can raise your level of intellectual functioning by sticking to a strict well-balanced diet.

_____30 People who eat just the food that is "good for them" are not much fun to be with.

BEHAVIORAL CONSEQUENCES ATTITUDE SCALES

The following pages contain lists of various outcomes or consequences that may be associated with certain behaviors. On each page, you will be asked to indicate the degree to which you feel that the behavior in question is related to each consequence. Your ratings should reflect how strongly you feel the behavior either prevents or leads to the consequences. The ratings should be made on a 1 to 7 scale, where:

1 = the behavior very strongly prevents the consequence
2 = the behavior strongly prevents the consequence
3 = the behavior somewhat prevents the consequence
4 = the behavior is unrelated to the consequence
5 = the behavior somewhat leads to the consequence
6 = the behavior strongly leads to the consequence
7 = the behavior very strongly leads to the consequence

Make your rating by writing a number from 1 to 7 on the line at the left of each listed consequence.

EXERCISE INSTRUMENTALITY RATINGS

Below is a list of experiences. For each item listed, indicate the extent to which you feel that your engaging in a regular program of physical exercise either leads to or prevents your experiencing each consequence. Make your rating by writing a number from 1 to 7 on the line to the left of each item, where 1 = exercise very strongly prevents and 7 = exercise very strongly leads to the consequence.

Engaging in a regular program of exercise

Very strongly prevents	Strongly prevents	Somewhat prevents	Unrelated to	Somewhat leads to	Strongly leads to	Very strongly leads to
1	2	3	4	5	6	7

Rating	Consequence
______1	my feeling weak
______2	my participating in sports
______3	my enjoying life
______4	my getting heart disease
______5	my being irritable
______6	my being in good health
______7	my living longer than average
______8	my feeling self-disciplined
______9	my being overweight
______10	my feeling mentally dull
______11	my feeling relaxed

DIET INSTRUMENTALITY RATINGS

Below is a list of experiences. For each item listed, indicate the extent to which you feel that your sticking to a well balanced, low cholesterol diet either leads to or prevents your experiencing each consequence. Make your rating by writing a number from 1 to 7 on the line to the left of each item, where 1 = proper diet very strongly prevents and 7 = proper diet very strongly leads to the consequence.

Sticking to a well balanced, low cholesterol diet

Very strongly prevents	Strongly prevents	Somewhat prevents	Unrelated to	Somewhat leads to	Strongly leads to	Very strongly leads to
1	2	3	4	5	6	7

Rating	Consequence
______12	my feeling clean
______13	my participating in sports
______14	my enjoying life
______15	my being irritable
______16	my being in good heatlh
______17	my living longer than average
______18	my feeling self-disciplined
______19	my being overweight
______20	my getting cancer
______21	my feeling relaxed
______22	my getting heart disease

CURRENT SMOKING STATUS

23. Please indicate your current status with regard to smoking cigarettes by carefully circling one of the numbers from 1 to 5 on the following scale.

    1. I usually smoke cigarettes just about every day.
    2. I now smoke cigarettes once in awhile, but not every day.
    3. I used to smoke cigarettes just about every day, but I don't smoke them now.
    4. I have smoked cigarettes a few times, but I don't smoke them now.
    5. I have never smoked cigarettes.

If you circled either 1 or 2 above, please follow the instructions for smokers on the following two pages.

If you circled 3, 4, or 5 above, please follow the instructions for nonsmokers on the following two pages.

SMOKING INSTRUMENTALITY RATINGS

Instructions for smokers (If you circled 1 or 2 on the previous page): For each item listed, indicate the extent to which you feel that your continuing to smoke cigarettes either leads to or prevents each outcome. Make your rating by writing a number from 1 to 7 on the line to the left of each item, where 1 = smoking very strongly prevents and 7 = very strongly leads to the consequence.

Instructions for nonsmokers (If you circled 3, 4, or 5 on the previous page): For each item listed, indicate the extent to which you feel that your starting to smoke cigarettes would either lead to or prevent each outcome. Make your rating by writing a number from 1 to 7 on the line to the left of each item, where 1 = smoking would very strongly prevent and 7 = smoking would very strongly lead to the consequence.

Smokers: Your continuing to smoke
Nonsmokers: Your starting to smoke

Very Strongly prevents	Strongly prevents	Somewhat prevents	Unrelated to	Somewhat leads to	Strongly leads to	Very strongly leads to
1	2	3	4	5	6	7

Rating	Consequence
______24	my feeling clean
______25	my participating in sports
______26	my enjoying life
______27	my having extra money
______28	my getting heart disease
______29	my being irritable
______30	my having a poor appetite
______31	my living longer than average
______32	my feeling self-disciplined

Very strongly prevents	Strongly prevents	Somewhat prevents	Unrelated to	Somewhat leads to	Strongly leads to	Very strongly leads to
1	2	3	4	5	6	7

Rating	Consequence
______33	my being unattractive to other people
______34	my getting cancer

NONSMOKING INSTRUMENTALITY RATINGS

Instructions for smokers: For each item listed, indicate the extent to which you feel your quitting smoking would either lead to or prevent each outcome. Make your rating by writing a number from 1 to 7 on the line to the left of each item, where 1 = quitting smoking would very strongly prevent and 7 = quitting smoking would very strongly lead to the consequence.

Instructions for nonsmokers: For each item listed, indicate the extent to which you feel your continuing not to smoke cigarettes either leads to or prevents each outcome. Make your rating by writing a number from 1 to 7 on the line to the left of each item, where 1 = your continuing not to smoke very strongly prevents and 7 = your continuing not to smoke very strongly leads to the consequence.

Smokers: Your quitting smoking
Nonsmokers: Your continuing not to smoke

Very strongly prevents	Strongly prevents	Somewhat prevents	Unrelated to	Somewhat leads to	Strongly leads to	Very strongly leads to
1	2	3	4	5	6	7

Rating	Consequence
______35	my feeling clean
______36	my participating in sports
______37	my enjoying life
______38	my having extra money
______39	my getting heart disease
______40	my being irritable
______41	my being in good health
______42	my living longer than average
______43	my feeling self-disciplined

Very strongly prevents	Strongly prevents	Somewhat prevents	Unrelated to	Somewhat leads to	Strongly leads to	Very strongly leads to
1	2	3	4	5	6	7

Rating	Consequence
______44	my being unattractive to other people
______45	my getting cancer

VALUE IMPORTANCE RATINGS

Finally, please evaluate each of the listed consequences in terms of how good, satisfied and happy; or bad, dissatisfied and unhappy you would feel if you experienced them. Make your rating by writing a number from 1 to 7 on the line to the left of each item, where 1 = extremely bad and 7 = extremely good.

Rating Scale

1	2	3	4	5	6	7
Extremely bad, dissatisfied, unhappy						Extremely good, satisfied, happy

Rating	Consequence
______46	my feeling clean
______47	my participating in sports
______48	my enjoying life
______49	my having extra money
______50	my getting heart disease
______51	my being irritable
______52	my being in good health
______53	my having a poor appetite
______54	my living longer than average
______55	my feeling self-disciplined
______56	my being overweight
______57	my getting cancer
______58	my feeling relaxed
______59	my being unattractive to other people

Rating Scale

1	2	3	4	5	6	7
Extremely bad, dis-satisfied, unhappy						Extremely good, satisfied, happy

Rating	Consequence
______60	my feeling weak
______61	my feeling mentally dull

EXERCISE HEALTH INVENTORY

For each type of the following activities, indicate the amount of time each week you typically spend on them by writing the number of hours and/or minutes on the line in front of each activity. If you never or nearly never engage in the activity, place a zero on the line. For seasonal activities (e.g., snow skiing), indicate the average number of hours per week during the season.

____hrs. ____mins. 1. Walking for pleasure and/or to work

____hrs. ____mins. 2. Hiking/backpacking

____hrs. ____mins. 3. Bicycling to work and/or for pleasure

____hrs. ____mins. 4. Using stairs when elevator is available

____hrs. ____mins. 5. Dancing

____hrs. ____mins. 6. Calisthenics (home exercise)

____hrs. ____mins. 7. Health club exercise

____hrs. ____mins. 8. Jogging/walking combination

____hrs. ____mins. 9. Jogging/running combination

____hrs. ____mins. 10. Weight lifting

____hrs. ____mins. 11. Swimming

____hrs. ____mins. 12. Snow skiing

____hrs. ____mins. 13. Ice or roller skating

____hrs. ____mins. 14. Baseball

____hrs. ____mins. 15. Basketball

____hrs. ____mins. 16. Racketball, handball

____hrs. ____mins. 17. Softball

____hrs. ____mins. 18. Table tennis (ping pong)

____hrs. ____mins. 19. Tennis

____hrs. ____mins. 20. Soccer

____hrs. ____mins. 21. Badminton

____hrs. ____mins. 22. Volley ball

____hrs. ____mins. 23. Hunting

____hrs. ____mins. 24. Bowling

____hrs. ____mins. 25. Golf (walking, pulling clubs or cart, or carrying clubs)

____hrs. ____mins. 26. Lawn mowing

____hrs. ____mins. 27. Other (specify) ________________________

DIET HEALTH INVENTORY

For each of the following foods, indicate how many times per week you typically have a serving of them by writing the number of times on the line in front of each food. You may use fractions if you have an average of less than one serving per week.

Average number of servings per week		Average number of servings per week	
_____	1. bacon	_____	11. butter
_____	2. sausage	_____	12. beef (rib roasts)
_____	3. fish	_____	13. milk (whole or 2%)
_____	4. fruit	_____	14. noodles
_____	5. ice cream	_____	15. beans
_____	6. hamburger	_____	16. cake, pie and pastries
_____	7. french fries	_____	17. pork roast
_____	8. chicken, turkey	_____	18. cheese
_____	9. vegetables	_____	19. eggs
_____	10. hot dogs	_____	20. lunchmeat

APPROVAL SHEET

The thesis submitted by David William Rivers

has been read and approved by the following committee:

Dr. Leonard Bickman, Director

Professor, Psychology, Loyola

Dr. John Edwards

Associate Professor, Psychology, Loyola

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

12/9/80

Date

Leonard Bickman

Director's Signature