GILBERT A. CHURCHILL, JR.*

A critical element in the evolution of a fundamental body of knowledge in marketing, as well as for improved marketing practice, is the development of better measures of the variables with which marketers work. In this article an approach is outlined by which this goal can be achieved and portions of the approach are illustrated in terms of a job satisfaction measure.

A Paradigm for Developing Better Measures of Marketing Constructs

In an article in the April 1978 issue of the *Journal of Marketing,* Jacoby placed much of the blame for the poor quality of some of the marketing literature on the measures marketers use to assess their variables of interest (p. 91):

> More stupefying than the sheer number of our measures is the ease with which they are proposed and the uncritical manner in which they are accepted. In point of fact, most of our measures are only measures because someone *says* that they are, not because they have been shown to satisfy standard measurement criteria (validity, reliability, and sensitivity). Stated somewhat differently, most of our measures are no more sophisticated than first asserting that the number of pebbles a person can count in a ten-minute period is a measure of that person's intelligence; next, conducting a study and finding that people who can count many pebbles in ten minutes also tend to eat more; and, finally, concluding from this: people with high intelligence tend to eat more.

Burleigh Gardner, President of Social Research, Inc., makes a similar point with respect to attitude measurement in a recent issue of the *Marketing News* (May 5, 1978, p. 1):

> Today the social scientists are enamored of numbers and counting . . .. Rarely do they stop and ask, "What lies behind the numbers?"
>
> When we talk about attitudes we are talking about constructs of the mind as they are expressed in response to our questions.
>
> But usually all we really know are the questions we ask and the answers we get.

Marketers, indeed, seem to be choking on their measures, as other articles in this issue attest. They seem to spend much effort and time operating by the routine which computer technicians refer to as GIGO—garbage in, garbage out. As Jacoby so succinctly puts it, "What does it mean if a finding is significant or that the ultimate in statistical analytical techniques have been applied, if the data collection instrument generated invalid data at the outset?" (1978, p. 90).

What accounts for this gap between the obvious need for better measures and the lack of such measures? The basic thesis of this article is that although the desire may be there, the know-how is not. The situation in marketing seems to parallel the dilemma which psychologists faced more than 20 years ago, when Tryon (1957, p. 229) wrote:

> If an investigator should invent a new psychological test and then turn to any recent scholarly work for

*Gilbert A. Churchill is Professor of Marketing, University of Wisconsin-Madison. The significant contributions of Michael Houston, Shelby Hunt, John Nevin, and Michael Rothschild through their comments on a draft of this article are gratefully acknowledged, as are the many helpful comments of the anonymous reviewers.

The AMA publications policy states: "No article(s) will be published in the *Journal of Marketing Research* written by the Editor or the Vice President of Publications." The inclusion of this article was approved by the Board of Directors because: (1) the article was submitted before the author took over as Editor, (2) the author played no part in its review, and (3) Michael Ray, who supervised the reviewing process for the special issue, formally requested he be allowed to publish it.

Journal of Marketing Research
Vol. XVI (February 1979), 64–73

> guidance on how to determine its reliability, he would confront such an array of different formulations that he would be unsure about how to proceed. After fifty years of psychological testing, the problem of discovering the degree to which an objective measure of behavior reliably differentiates individuals is still confused.

Psychology has made progress since that time. Attention has moved beyond simple questions of reliability and now includes more "direct" assessments of validity. Unfortunately, the marketing literature has been slow to reflect that progress. One of the main reasons is that the psychological literature is scattered. The notions are available in many bits and pieces in a variety of sources. There is no overriding framework which the marketer can embrace to help organize the many definitions and measures of reliability and validity into an integrated whole so that the decision as to which to use and when is obvious.

This article is an attempt to provide such a framework. A procedure is suggested by which measures of constructs of interest to marketers can be developed. The emphasis is on developing measures which have desirable reliability and validity properties. Part of the article is devoted to clarifying these notions, particularly those related to validity; reliability notions are well covered by Peter's article in this issue. Finally, the article contains suggestions about approaches on which marketers historically have relied in assessing the quality of measures, but which they would do well to consider abandoning in favor of some newer alternatives. The rationale as to why the newer alternatives are preferred is presented.

THE PROBLEM AND APPROACH

Technically, the process of measurement or operationalization involves "rules for assigning numbers to objects to represent quantities of attributes" (Nunnally, 1967, p. 2). The definition involves two key notions. First, it is the attributes of objects that are measured and not the objects themselves. Second, the definition does not specify the rules by which the numbers are assigned. However, the rigor with which the rules are specified and the skill with which they are applied determine whether the construct has been captured by the measure.

Consider some arbitrary construct, C, such as customer satisfaction. One can conceive at any given point in time that every customer has a "true" level of satisfaction; call this level X_T. Hopefully, each measurement one makes will produce an observed score, X_O, equal to the object's true score, X_T. Further, if there are differences between objects with respect to their X_o scores, these differences would be completely attributable to true differences in the characteristic one is attempting to measure, i.e., true differences in X_T. Rarely is the researcher so fortunate. Much more typical is the measurement where the X_o score differences also reflect (Selltiz et al., 1976, p. 164–8):

1. True differences in other relatively stable characteristics which affect the score, e.g., a person's willingness to express his or her true feelings.
2. Differences due to transient personal factors, e.g., a person's mood, state of fatigue.
3. Differences due to situational factors, e.g., whether the interview is conducted in the home or at a central facility.
4. Differences due to variations in administration, e.g., interviewers who probe differently.
5. Differences due to sampling of items, e.g., the specific items used on the questionnaire; if the items or the wording of those items were changed, the X_o scores would also change.
6. Differences due to lack of clarity of measuring instruments, e.g., vague or ambiguous questions which are interpreted differently by those responding.
7. Differences due to mechanical factors, e.g., a check mark in the wrong box or a response which is coded incorrectly.

Not all of these factors will be present in every measurement, nor are they limited to information collected by questionnaire in personal or telephone interviews. They arise also in studies in which self-administered questionnaires or observational techniques are used. Although the impact of each factor on the X_o score varies with the approach, their impact is predictable. They distort the observed scores away from the true scores. Functionally, the relationship can be expressed as:

$$X_o = X_T + X_S + X_R$$

where:

X_S = systematic sources of error such as stable characteristics of the object which affect its score, and

X_R = random sources of error such as transient personal factors which affect the object's score.

A measure is *valid* when the differences in observed scores reflect true differences on the characteristic one is attempting to measure and nothing else, that is, $X_o = X_T$. A measure is *reliable* to the extent that independent but comparable measures of the same trait or construct of a given object agree. Reliability depends on how much of the variation in scores is attributable to random or chance errors. If a measure is perfectly reliable, $X_R = 0$. Note that if a measure is valid, it is reliable, but that the converse is not necessarily true because the observed score when $X_R = 0$ could still equal $X_T + X_S$. Thus it is often said that reliability is a necessary but not a sufficient condition for validity. Reliability only provides negative evidence of the validity of the measure. However, the ease with which it can be computed helps explain

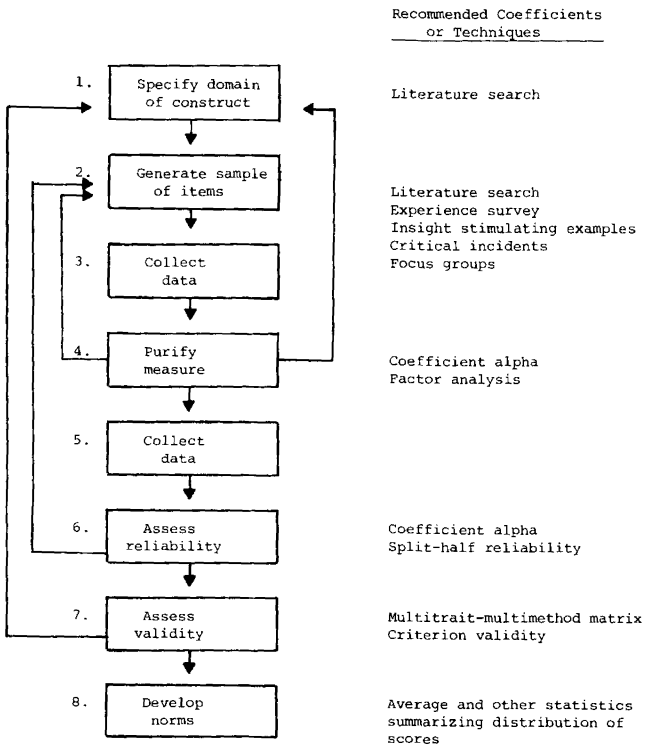
its popularity. Reliability is much more routinely reported than is evidence, which is much more difficult to secure but which relates more directly to the validity of the measure.

The fundamental objective in measurement is to produce X_o scores which approximate X_T scores as closely as possible. Unfortunately, the researcher never knows for sure what the X_T scores are. Rather, the measures are always inferences. The quality of these inferences depends directly on the procedures that are used to develop the measures and the evidence supporting their "goodness." This evidence typically takes the form of some reliability or validity index, of which there are a great many, perhaps too many.

The analyst working to develop a measure must contend with such notions as split-half, test-retest, and alternate forms reliability as well as with face, content, predictive, concurrent, pragmatic, construct, convergent, and discriminant validity. Because some of these terms are used interchangeably and others are often used loosely, the analyst wishing to develop a measure of some variable of interest in marketing faces difficult decisions about how to proceed and what reliability and validity indices to calculate.

Figure 1 is a diagram of the sequence of steps that can be followed and a list of some calculations that should be performed in developing measures of marketing constructs. The suggested sequence has worked well in several instances in producing measures with desirable psychometric properties (see Churchill et al., 1974, for one example). Some readers will undoubtedly disagree with the suggested process or with the omission of their favorite reliability or validity coefficient. The following discussion, which details both the steps and their rationale, shows that some of these measures should indeed be set aside because there are better alternatives or, if they are used, that they should at least be interpreted with the proper awareness of their shortcomings.

Figure 1
SUGGESTED PROCEDURE FOR DEVELOPING BETTER MEASURES



The process suggested is only applicable to multi-item measures. This deficiency is not as serious as it might appear. Multi-item measures have much to recommend them. First, individual items usually have considerable uniqueness or specificity in that each item tends to have only a low correlation with the attribute being measured and tends to relate to other attributes as well. Second, single items tend to categorize people into a relatively small number of groups. For example, a seven-step rating scale can at most distinguish between seven levels of an attribute. Third, individual items typically have considerable measurement error; they produce unreliable responses in the sense that the same scale position is unlikely to be checked in successive administrations of an instrument.

All three of these measurement difficulties can be diminished with multi-item measures: (1) the specificity of items can be averaged out when they are combined, (2) by combining items, one can make relatively fine distinctions among people, and (3) the reliability tends to increase and measurement error decreases as the number of items in a combination increases.

The folly of using single-item measures is illustrated by a question posed by Jacoby (1978, p. 93):

> How comfortable would we feel having our intelligence assessed on the basis of our response to a *single* question?" Yet that's exactly what we do in consumer research. . . . The literature reveals hundreds of instances in which responses to a single question suffice to establish the person's level on the variable of interest and then serves as the basis for extensive analysis and entire articles.
>
> . . . Given the complexity of our subject matter, what makes us think we can use responses to single items (or even to two or three items) as measures of these concepts, then relate these scores to a host of other variables, arrive at conclusions based on such an investigation, and get away calling what we have done "quality research?"

In sum, marketers are much better served with multi-item than single-item measures of their constructs, and they should take the time to develop them. This conclusion is particularly true for those investigating behavioral relationships from a fundamental

as well as applied perspective, although it applies also to marketing practitioners.

SPECIFY DOMAIN OF THE CONSTRUCT

The first step in the suggested procedure for developing better measures involves specifying the domain of the construct. The researcher must be exacting in delineating what is included in the definition and what is excluded. Consider, for example, the construct customer satisfaction, which lies at the heart of the marketing concept. Though it is a central notion in modern marketing thought, it is also a construct which marketers have not measured in exacting fashion.

Howard and Sheth (1969, p. 145), for example, define customer satisfaction as

> . . .the buyer's cognitive state of being adequately or inadequately rewarded in a buying situation for the sacrifice he has undergone. The adequacy is a consequence of matching actual past purchase and consumption experience with the reward that was expected from the brand in terms of its anticipated potential to satisfy the motives served by the particular product class. It includes not only reward from consumption of the brand but any other reward received in the purchasing and consuming process.

Thus, satisfaction by their definition seems to be attitude. Further, in order to measure satisfaction, it seems necessary to measure both expectations at the time of purchase and reactions at some time after purchase. If actual consequences equal or exceed expected consequences, the consumer is satisfied, but if actual consequences fall short of expected consequences, the consumer is dissatisfied.

But what expectations and consequences should the marketer attempt to assess? Certainly one would want to be reasonably exhaustive in the list of product features to be included, incorporating such facets as cost, durability, quality, operating performance, and aesthetic features (Czepeil et al., 1974). But what about purchasers' reactions to the sales assistance they received or subsequent service by independent dealers, as would be needed, for example, after the purchase of many small appliances? What about customer reaction to subsequent advertising or the expansion of the channels of distribution in which the product is available? What about the subsequent availability of competitors' alternatives which serve the same needs or the publishing of information about the environmental effects of using the product? To detail which of these factors would be included or how customer satisfaction *should be* operationalized is beyond the scope of this article; rather, the example emphasizes that the researcher must be exacting in the conceptual specification of the construct and what is and what is not included in the domain.

It is imperative, though, that researchers consult the literature when conceptualizing constructs and specifying domains. Perhaps if only a few more had done so, one of the main problems cited by Kollat, Engel, and Blackwell as impairing progress in consumer research—namely, the use of widely varying definitions—could have been at least diminished (Kollat et al., 1970, p. 328–9).

Certainly definitions of constructs are means rather than ends in themselves. Yet the use of different definitions makes it difficult to compare and accumulate findings and thereby develop syntheses of what is known. Researchers should have good reasons for proposing additional *new* measures given the many available for most marketing constructs of interest, and those publishing should be required to supply their rationale. Perhaps the older measures are inadequate. The researcher should make sure this is the case by conducting a thorough review of literature in which the variable is used and should present a detailed statement of the reasons and evidence as to why the new measure is better.

GENERATE SAMPLE OF ITEMS

The second step in the procedure for developing better measures is to generate items which capture the domain as specified. Those techniques that are typically productive in exploratory research, including literature searches, experience surveys, and insight-stimulating examples, are generally productive here (Selltiz et al., 1976). The literature should indicate how the variable has been defined previously and how many dimensions or components it has. The search for ways to measure customer satisfaction would include product brochures, articles in trade magazines and newspapers, or results of product tests such as those published by *Consumer Reports*. The experience survey is not a probability sample but a judgment sample of persons who can offer some ideas and insights into the phenomenon. In measuring consumer satisfaction, it could include discussions with (1) appropriate people in the product group responsible for the product, (2) sales representatives, (3) dealers, (4) consumers, and (5) persons in marketing research or advertising, as well as (6) outsiders who have a special expertise such as university or government personnel. The insight-stimulating examples could involve a comparison of competitors' products or a detailed examination of some particularly vehement complaints in unsolicited letters about performance of the product. Examples which indicate sharp contrasts or have striking features would be most productive.

Critical incidents and focus groups also can be used to advantage at the item-generation stage. To use the critical incidents technique a large number of scenarios describing specific situations could be made up and a sample of experienced consumers would be asked what specific behaviors (e.g., product changes, warranty handling) would create customer satisfaction or dissatisfaction (Flanagan, 1954; Kerlinger, 1973, p.

536). The scenarios might be presented to the respondents individually or 8 to 10 of them might be brought together in a focus group where the scenarios could be used to trigger open discussion among participants, although other devices might also be employed to promote discourse (Calder, 1977).

The emphasis at the early stages of item generation would be to develop a set of items which tap each of the dimensions of the construct at issue. Further, the researcher probably would want to include items with slightly different shades of meaning because the original list will be refined to produce the final measure. Experienced researchers can attest that seemingly identical statements produce widely different answers. By incorporating slightly different nuances of meaning in statements in the item pool, the researcher provides a better foundation for the eventual measure.

Near the end of the statement development stage the focus would shift to item editing. Each statement would be reviewed so that its wording would be as precise as possible. Double-barreled statements would be split into two single-idea statements, and if that proved impossible the statement would be eliminated altogether. Some of the statements would be recast to be positively stated and others to be negatively stated to reduce "yea-" or "nay-" saying tendencies. The analyst's attention would also be directed at refining those questions which contain an obvious "socially acceptable" response.

After the item pool is carefully edited, further refinement would await actual data. The type of data collected would depend on the type of scale used to measure the construct.

PURIFY THE MEASURE

The calculations one performs in purifying a measure depend somewhat on the measurement model one embraces. The most logically defensible model is the domain sampling model which holds that the purpose of any particular measurement is to estimate the score that would be obtained if *all* the items in the domain were used (Nunnally, 1967, p. 175–81). The score that any subject would obtain over the whole sample domain is the person's true score, X_T.

In practice, though, one does not use all of the items that could be used, but only a sample of them. To the extent that the sample of items correlates with true scores, it is good. According to the domain sampling model, then, a primary source of measurement error is the inadequate sampling of the domain of relevant items.

Basic to the domain sampling model is the concept of an infinitely large correlation matrix showing all correlations among the items in the domain. No single item is likely to provide a perfect representation of the concept, just as no single word can be used to test for differences in subjects' spelling abilities and no single question can measure a person's intelligence. Rather, each item can be expected to have a certain amount of distinctiveness or specificity even though it relates to the concept.

The average correlation in this infinitely large matrix, r, indicates the extent to which some common core is present in the items. The dispersion of correlations about the average indicates the extent to which items vary in sharing the common core. The key assumption in the domain sampling model is that all items, *if they belong to the domain of the concept*, have an equal amount of common core. This statement implies that the average correlation in each column of the hypothetical matrix is the same and in turn equals the average correlation in the whole matrix (Ley, 1972, p. 111; Nunnally, 1967, p. 175–6). That is, if all the items in a measure are drawn from the domain of a single construct, responses to those items should be highly intercorrelated. Low interitem correlations, in contrast, indicate that some items are not drawn from the appropriate domain and are producing error and unreliability.

Coefficient Alpha

The recommended measure of the internal consistency of a set of items is provided by coefficient alpha which results directly from the assumptions of the domain sampling model. See Peter's article in this issue for the calculation of coefficient alpha.

Coefficient alpha *absolutely* should be the first measure one calculates to assess the quality of the instrument. It is pregnant with meaning because the *square root* of coefficient alpha is the *estimated correlation of the k-item test with errorless true scores* (Nunnally, 1967, p. 191–6). Thus, a low coefficient alpha indicates the sample of items performs poorly in capturing the construct which motivated the measure. Conversely, a large alpha indicates that the k-item test correlates well with true scores.

If alpha is low, what should the analyst do?[1] If the item pool is sufficiently large, this outcome suggests that some items do not share equally in the common core and should be eliminated. The easiest way to find them is to calculate the correlation of each item with the total score and to plot these correlations by decreasing order of magnitude. Items with correlations near zero would be eliminated. Further, items which produce a substantial or sudden drop in the item-to-total correlations would also be deleted.

[1] What is "low" for alpha depends on the purpose of the research. For early stages of basic research, Nunnally (1967) suggests reliabilities of .50 to .60 suffice and that increasing reliabilities beyond .80 is probably wasteful. In many applied settings, however, where important decisions are made with respect to specific test scores, "a reliability of .90 is the minimum that should be tolerated, and a reliability of .95 should be considered the desirable standard" (p. 226).

If the construct had, say, five identifiable dimensions or components, coefficient alpha would be calculated for each dimension. The item-to-total correlations used to delete items would also be based on the items in the component and the total score for that dimension. The total score for the construct would be secured by summing the total scores for the separate components. The reliability of the total construct would not be measured through coefficient alpha, but rather through the formula for the reliability of linear combinations (Nunnally, 1967, p. 226–35).

Some analysts mistakenly calculate split-half reliability to assess the internal homogeneity of the measure. That is, they divide the measure into two halves. The first half may be composed of all the even-numbered items, for example, and the second half all the odd-numbered items. The analyst then calculates a total score for each half and correlates these total scores across subjects. The problem with this approach is that the size of this correlation depends on the way the items are split to form the two halves. With, say, 10 items (a very small number for most measurements), there are 126 possible splits.[2] Because each of these possible divisions will likely produce a different coefficient, what is the split-half reliability? Further, as the average of all of these coefficients equals coefficient alpha, why not calculate coefficient alpha in the first place? It is almost as easy, is not arbitrary, and has an important practical connotation.

Factor Analysis

Some analysts like to perform a factor analysis on the data before doing anything else in the hope of determining the number of dimensions underlying the construct. Factor analysis can indeed be used to suggest dimensions, and the marketing literature is replete with articles reporting such use. Much less prevalent is its use to confirm or refute components isolated by other means. For example, in discussing a test composed of items tapping two common factors, verbal fluency and number facility, Campbell (1976, p. 194) comments:

> Recognizing multidimensionality when we see it is not always an easy task. For example, rules for when to stop extracting factors are always arbitrary in some sense. Perhaps the wisest course is to always make the comparison between the split half and internal consistency estimates after first splitting the components into two halves on a priori grounds. That is, every effort should be made to balance the factor content of each half [part] *before* looking at component intercorrelations.

When factor analysis is done before the purification steps suggested heretofore, there seems to be a tendency to produce many more dimensions than can be conceptually identified. This effect is partly due to the "garbage items" which do not have the common core but which do produce additional dimensions in the factor analysis. Though this application may be satisfactory during the early stages of research on a construct, the use of factor analysis in a confirmatory fashion would seem better at later stages. Further, theoretical arguments support the iterative process of the calculation of coefficient alpha, the elimination of items, and the subsequent calculation of alpha until a satisfactory coefficient is achieved. Factor analysis then can be used to confirm whether the number of dimensions conceptualized can be verified empirically.

Iteration

The foregoing procedure can produce several outcomes. The most desirable outcome occurs when the measure produces a satisfactory coefficient alpha (or alphas if there are multiple dimensions) and the dimensions agree with those conceptualized. The measure is then ready for some additional testing for which a *new sample of data* should be collected. Second, factor analysis sometimes suggests that dimensions which were conceptualized as independent clearly overlap. In this case, the items which have pure loadings on the new factor can be retained and a new alpha calculated. If this outcome is satisfactory, additional testing with new data is indicated.

The third and least desirable outcome occurs when the alpha coefficient(s) is too low and restructuring of the items forming each dimension is unproductive. In this case, the appropriate strategy is to loop back to steps 1 and 2 and repeat the process to ascertain what might have gone wrong. Perhaps the construct was not appropriately delineated. Perhaps the item pool did not sample all aspects of the domain. Perhaps the emphases within the measure were somehow distorted in editing. Perhaps the sample of subjects was biased, or the construct so ambiguous as to defy measurement. The last conclusion would suggest a fundamental change in strategy, starting with a rethinking of the basic relationships that motivated the investigation in the first place.

ASSESS RELIABILITY WITH NEW DATA

The major source of error within a test or measure is the sampling of items. If the sample is appropriate and the items "look right," the measure is said to have *face* or *content* validity. Adherence to the steps suggested will tend to produce content valid measures. But that is not the whole story! What about transient personal factors, or ambiguous questions which pro-

[2] The number of possible splits with $2n$ items is given by the formula $\frac{(2n)!}{2(n!)(n!)}$ (Bohrnstedt, 1970). For the example cited, $n = 5$ and the formula reduces to $\frac{10!}{2(5!)(5!)}$.

duce guessing, or any of the other extraneous influences, other than the sampling of items, which tend to produce error in the measure?

Interestingly, all of the errors that occur within a test can be easily encompassed by the domain sampling model. All the sources of error occurring within a measurement will tend to lower the average correlation among the items within the test, but the average correlation is all that is needed to estimate the reliability. Suppose, for example, that one of the items is vague and respondents have to guess its meaning. This guessing will tend to lower coefficient alpha, suggesting there is error in the measurement. Subsequent calculation of item-to-total correlations will then suggest this item for elimination.

Coefficient alpha is the basic statistic for determining the reliability of a measure based on internal consistency. Coefficient alpha does not adequately estimate, though, errors caused by factors external to the instrument, such as differences in testing situations and respondents over time. If the researcher wants a reliability coefficient which assesses the between-test error, additional data must be collected. It is also advisable to collect additional data to rule out the possibility that the previous findings are due to chance. If the construct is more than a measurement artifact, it should be reproduced when the purified sample of items is submitted to a new sample of subjects.

Because Peter's article treats the assessment of reliability, it is not examined here except to suggest that test-retest reliability should *not* be used. The basic problem with straight test-retest reliability is respondents' memories. They will tend to reply to an item the same way in a second administration as they did in the first. Thus, even if an analyst were to put together an instrument in which the items correlate poorly, suggesting there is no common core and thus no construct, it is possible and even probable that the responses to each item would correlate well across the two measurements. The high correlation of the total scores on the two tests would suggest the measure had small measurement error when in fact very little is demonstrated about validity by straight test-retest correlations.

ASSESS CONSTRUCT VALIDITY

Specifying the domain of the construct, generating items that exhaust the domain, and subsequently purifying the resulting scale should produce a measure which is content or face valid and reliable. It may or may not produce a measure which has construct validity. Construct validity, which lies at the very heart of the scientific process, is most directly related to the question of what the instrument is in fact measuring—what construct, trait, or concept underlies a person's performance or score on a measure.

The preceding steps should produce an *internally consistent* or *internally homogeneous* set of items. Consistency is necessary but not sufficient for construct validity (Nunnally, 1967, p. 92).

Rather, to establish the construct validity of a measure, the analyst also must determine (1) the extent to which the measure correlates with other measures designed to measure the same thing and (2) whether the measure behaves as expected.

Correlations With Other Measures

A fundamental principle in science is that any particular construct or trait should be measurable by at least two, and preferably more, different methods. Otherwise the researcher has no way of knowing whether the trait is anything but an artifact of the measurement procedure. All the measurements of the trait may not be equally good, but science continually emphasizes improvement of the measures of the variables with which it works. Evidence of the *convergent validity* of the measure is provided by the extent to which it correlates highly with other methods designed to measure the same construct.

The measures should have not only convergent validity, but also *discriminant validity*. Discriminant validity is the extent to which the measure is indeed novel and not simply a reflection of some other variable. As Campbell and Fiske (1959) persuasively argue, "Tests can be invalidated by too high correlations with other tests from which they were intended to differ" (p. 81). Quite simply, scales that correlate *too highly* may be measuring the *same* rather than *different* constructs. Discriminant validity is indicated by "predictably low correlations between the measure of interest and other measures that are supposedly not measuring the same variable or concept" (Heeler and Ray, p. 362).

A useful way of assessing the convergent and discriminant validity of a measure is through the multitrait-multimethod matrix, which is a matrix of zero order correlations between different traits when each of the traits is measured by different methods (Campbell and Fiske, 1959). Table 1, for example, is the matrix for a Likert type of measure designed to assess salesperson job satisfaction (Churchill et al., 1974). The four essential elements of a multitrait-multimethod matrix are identified by the numbers in the upper left corner of each partitioned segment.

Only the reliability diagonal (1) corresponding to the Likert measure is shown; data were not collected for the thermometer scale because it was not of interest itself. The entries reflect the reliability of alternate forms administered two weeks apart. If these are unavailable, coefficient alpha can be used.

Evidence about the convergent validity of a measure is provided in the validity diagonal (3) by the extent to which the correlations are significantly different from zero and sufficiently large to encourage further examination of validity. The validity coefficients in

Table 1
MULTITRAIT-MULTIMETHOD MATRIX

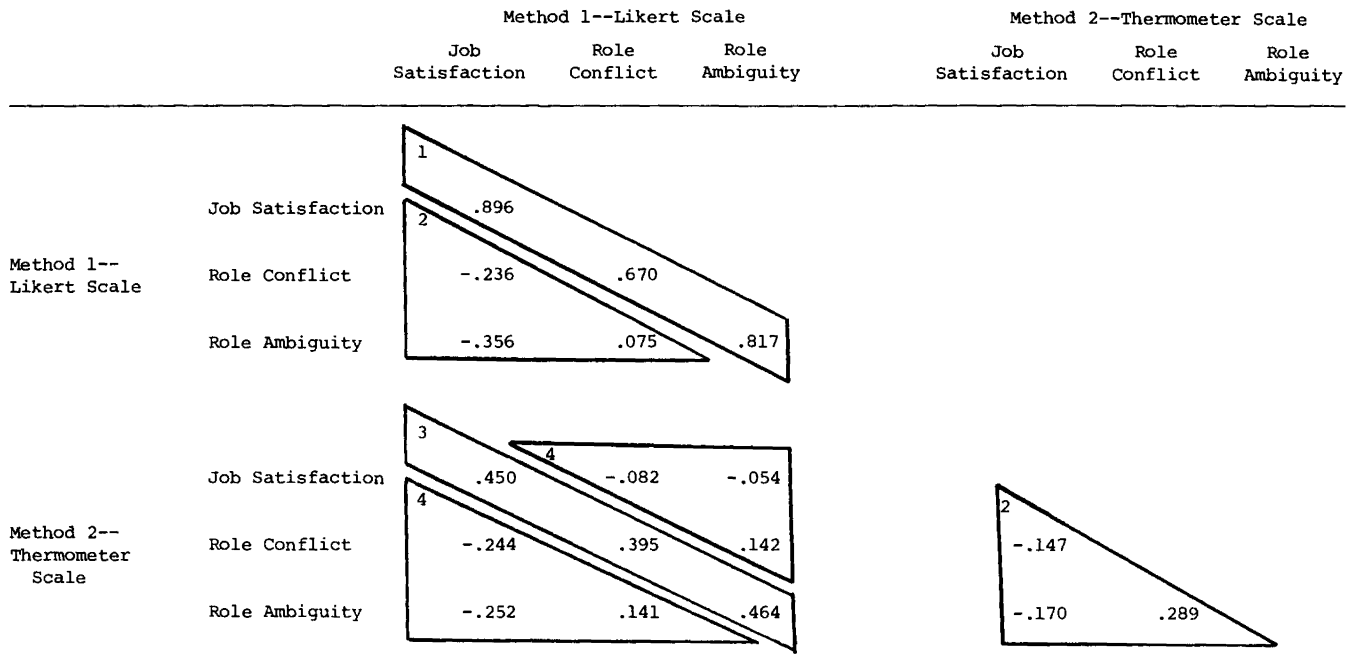


Table 1 of .450, .395 and .464 are all significant at the .01 level.

Discriminant validity, however, suggests three comparisons, namely that:

1. Entries in the validity diagonal (3) should be higher than the correlations that occupy the same row and column in the heteromethod block (4). This is a minimum requirement as it simply means that the correlation between two different measures of the same variable should be higher than the correlations "between that variable and any other variable which has neither trait nor method in common" (Campbell and Fiske, 1959, p. 82). The entries in Table 1 satisfy this condition.
2. The validity coefficients (3) should be higher than the correlations in the heterotrait-monomethod triangles (2) which suggests that the correlation within a trait measured by different methods must be higher than the correlations between traits which have method in common. It is a more stringent requirement than that involved in the heteromethod comparisons of step 1 as the off-diagonal elements in the monomethod blocks may be high because of method variance. The evidence in Table 1 is consistent with this requirement.
3. The pattern of correlations should be the same in all of the heterotrait triangles, e.g., both (2) and (4). This requirement is a check on the significance of the traits when compared to the methods and can be achieved by rank ordering the correlation coefficients in each heterotrait triangle; though a visual inspection often suffices, a rank order correlation coefficient such as the coefficient of concordance can be computed if there are a great many comparisons.

The last requirement is generally, though not completely, satisfied by the data in Table 1. Within each heterotrait triangle, the pairwise correlations are consistent in sign. Further, when the correlations within each heterotrait triangle are ranked from largest positive to largest negative, the same order emerges except for the lower left triangle in the heteromethod block. Here the correlation between job satisfaction and role ambiguity is higher, i.e., less negative, than that between job satisfaction and role conflict whereas the opposite was true in the other three heterotrait triangles (see Ford et al., 1975, p. 107, as to why this single violation of the desired pattern may not represent a serious distortion in the measure).

Ideally, the methods and traits generating the multitrait-multimethod matrix should be as independent as possible (Campbell and Fiske, 1959, p. 103). Sometimes the nature of the trait rules out the opportunity for measuring it by different methods, thus introducing the possibility of method variance. When this situation arises, the researcher's efforts should be directed to obtaining as much diversity as possible in terms of data sources and scoring procedures. If the traits are not independent, the monomethod correlations will be large and the heteromethod correlations between traits will also be substantial, and the evidence about

the discriminant validity of the measure will not be as easily established as when they are independent. Thus, Campbell and Fiske (1959, p. 103) suggest that it is preferable to include at least two sets of independent traits in the matrix.

Does the Measure Behave as Expected?

Internal consistency is a necessary but insufficient condition for construct validity. The observables may all relate to the same construct, but that does not prove that they relate to the specific construct that motivated the research in the first place. A suggested final step is to show that the measure behaves as expected in relation to other constructs. Thus one often tries to assess whether the scale score can differentiate the positions of "known groups" or whether the scale correctly predicts some criterion measure (criterion validity). Does a salesperson's job satisfaction, as measured by the scale, for example, relate to the individual's likelihood of quitting? It should, according to what is known about dissatisfied employees; if it does not, then one might question the quality of the measure of salesperson job satisfaction. Note, though, there is circular logic in the foregoing argument. The argument rests on four separate propositions (Nunnally, 1967, p. 93):

1. The constructs job satisfaction (A) and likelihood of quitting (B) are related.
2. The scale X provides a measure of A.
3. Y provides a measure of B.
4. X and Y correlate positively.

Only the fourth proposition is directly examined with empirical data. To establish that X truly measures A, one *must assume* that propositions 1 and 3 are correct. One must have a good measure for B, and the theory relating A and B must be true. Thus, the analyst tries to establish the construct validity of a measure by relating it to a number of other constructs and not simply one. Further, one also tries to use those theories and hypotheses which have been sufficiently well scrutinized to inspire confidence in their probable truth. Thus, job satisfaction would not be related to job performance because there is much disagreement about the relationship between these constructs (Schwab and Cummings, 1970).

DEVELOPING NORMS

Typically, a raw score on a measuring instrument used in a marketing investigation is not particularly informative about the position of a given object on the characteristic being measured because the units in which the scale is expressed are unfamiliar. For example, what does a score of 350 on a 100-item Likert scale with 1–5 scoring imply about a salesperson's job satisfaction? One would probably be tempted to conclude that because the neutral position is 3, a 350 score with 100 statements implies slightly positive attitude or satisfaction. The analyst should be cautious in making such an interpretation, though. Suppose the 350 score represents the highest score ever achieved on this instrument. Suppose it represents the lowest score. Clearly there is a difference.

A better way of assessing the position of the individual on the characteristic is to compare the person's score with the score achieved by other people. The technical name for this process is "developing norms," although it is something everyone does implicitly every day. Thus, by saying a person "sure is tall," one is saying the individual is much taller than others encountered previously. Each person has a mental standard of what is average, and classifies people as tall or short on the basis of how they compare with this mental standard.

In psychological measurement, such processes are formalized by making the implicit standards explicit. More particularly, meaning is imputed to a specific score in unfamiliar units by comparing it with the total distribution of scores, and this distribution is summarized by calculating a mean and standard deviation as well as other statistics such as centile rank of any particular score (see Ghiselli, 1964, p. 37–102, for a particularly lucid and compelling argument about the need and the procedures for norm development).

Norm quality is a function of both the number of cases on which the average is based and their representativeness. The larger the number of cases, the more stable will be the norms and the more definitive will be the conclusions that can be drawn, if the sample is representative of the total group the norms are to represent. Often it proves necessary to develop distinct norms for separate groups, e.g., by sex or by occupation. The need for such norms is particularly common in basic research, although it sometimes arises in applied marketing research as well.

Note that norms need not be developed if one wants only to compare salespersons i and j to determine who is more satisfied, or to determine how a particular individual's satisfaction has changed over time. For these comparisons, all one needs to do is compare the raw scores.

SUMMARY AND CONCLUSIONS

The purpose of this article is to outline a procedure which can be followed to develop better measures of marketing variables. The framework represents an attempt to unify and bring together in one place the scattered bits of information on how one goes about developing improved measures and how one assesses the quality of the measures that have been advanced.

Marketers certainly need to pay more attention to measure development. Many measures with which marketers now work are woefully inadequate, as the many literature reviews suggest. Despite the time and dollar costs associated with following the process suggested here, the payoffs with respect to the genera-

tion of a core body of knowledge are substantial. As Torgerson (1958) suggests in discussing the ordering of the various sciences along a theoretical-correlational continuum (p. 2):

> It is more than a mere coincidence that the sciences would order themselves in largely the same way if they were classified on the basis to which satisfactory measurement of their important variables has been achieved. The development of a theoretical science . . . would seem to be virtually impossible unless its variables can be measured adequately.

Progress in the development of marketing as a science certainly will depend on the measures marketers develop to estimate the variables of interest to them (Bartels, 1951; Buzzell, 1963; Converse, 1945; Hunt, 1976).

Persons doing research of a fundamental nature are well advised to execute the whole process suggested here. As scientists, marketers should be willing to make this committment to "quality research." Those doing applied research perhaps cannot "afford" the execution of each and every stage, although many of their conclusions are then likely to be nonsense, one-time relationships. Though the point could be argued at length, researchers doing applied work and practitioners could at least be expected to complete the process through step 4. The execution of steps 1–4 can be accomplished with one-time, cross-sectional data and will at least indicate whether one or more isolatable traits are being captured by the measures as well as the quality with which these traits are being assessed. At a minimum the execution of steps 1–4 should reduce the prevalent tendency to apply extremely sophisticated analysis to faulty data and thereby execute still another GIGO routine. And once steps 1–4 are done, data collected with each application of the measuring instrument can provide more and more evidence related to the other steps. As Ray points out in the introduction to this issue, marketing researchers are already collecting data relevant to steps 5–8. They just need to plan data collection and analysis more carefully to contribute to improved marketing measures.

REFERENCES

Bartels, Robert. "Can Marketing Be a Science?," *Journal of Marketing,* 15 (January 1951), 319–28.

Bohrnstedt, George W. "Reliability and Validity Assessment in Attitude Measurement," *in* Gene F. Summers, ed., *Attitude Measurement.* Chicago: Rand McNally and Company, 1970, 80–99.

Buzzell, Robert D. "Is Marketing a Science," *Harvard Business Review,* 41 (January–February 1963), 32–48.

Calder, Bobby J. "Focus Groups and the Nature of Qualitative Marketing Research," *Journal of Marketing Research,* 14 (August 1977), 353–64.

Campbell, Donald R. and Donald W. Fiske. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix," *Psychological Bulletin,* 56 (1959), 81–105.

Campbell, John P. "Psychometric Theory," *in* Marvin D. Dunette, ed., *Handbook of Industrial and Organizational Psychology.* Chicago: Rand McNally, Inc., 1976, 185–222.

Churchill, Gilbert A., Jr., Neil M. Ford, and Orville C. Walker, Jr. "Measuring the Satisfaction of Industrial Salesmen," *Journal of Marketing Research,* 11 (August 1974), 254–60.

Converse, Paul D. "The Development of a Science in Marketing," *Journal of Marketing,* 10 (July 1945), 14–23.

Cronbach, L. J. "Coefficient Alpha and the Internal Structure of Tests," *Psychometrika,* 16 (1951), 297–334.

Czepeil, John A., Larry J. Rosenberg, and Adebayo Akerale. "Perspectives on Consumer Satisfaction," *in* Ronald C. Curhan, ed., *1974 Combined Proceedings.* Chicago: American Marketing Association, 1974, 119–23.

Flanagan, J. "The Critical Incident Technique," *Psychological Bulletin,* 51 (1954), 327–58.

Ford, Neil M., Orville C. Walker, Jr. and Gilbert A. Churchill, Jr. "Expectation-Specific Measures of the Intersender Conflict and Role Ambiguity Experienced by Industrial Salesmen," *Journal of Business Research,* 3 (April 1975), 95–112.

Gardner, Burleigh B. "Attitude Research Lacks System to Help It Make Sense," *Marketing News,* 11 (May 5, 1978), 1+.

Ghiselli, Edwin E. *Theory of Psychological Measurement.* New York: McGraw-Hill Book Company, 1964.

Heeler, Roger M. and Michael L. Ray. "Measure Validation in Marketing," *Journal of Marketing Research,* 9 (November 1972), 361–70.

Howard, John A. and Jagdish N. Sheth. *The Theory of Buyer Behavior.* New York: John Wiley & Sons, Inc., 1969.

Hunt, Shelby D. "The Nature and Scope of Marketing," *Journal of Marketing,* 40 (July 1976), 17–28.

Jacoby, Jacob. "Consumer Research: A State of the Art Review," *Journal of Marketing,* 42 (April 1978), 87–96.

Kerlinger, Fred N. *Foundations of Behavioral Research,* 2nd ed. New York: Holt, Rinehart, Winston, Inc., 1973.

Kollat, David T., James F. Engel, and Roger D. Blackwell. "Current Problems in Consumer Behavior Research," *Journal of Marketing Research* 7 (August 1970), 327–32.

Ley, Philip. *Quantitative Aspects of Psychological Assessment.* London: Gerald Duckworth and Company, Ltd., 1972.

Nunnally, Jum C. *Psychometric Theory.* New York: McGraw-Hill Book Company, 1967.

Schwab, D. P. and L. L. Cummings, "Theories of Performance and Satisfaction: A Review," *Industrial Relations,* 9 (1970), 408–30.

Selltiz, Claire, Lawrence S. Wrightsman, and Stuart W. Cook. *Research Methods in Social Relations,* 3rd ed. New York: Holt, Rinehart, and Winston, 1976.

Torgerson, Warren S. *Theory and Methods of Scaling.* New York: John Wiley & Sons, Inc., 1958.

Tryon, Robert C. "Reliability and Behavior Domain Validity: Reformulation and Historical Critique," *Psychological Bulletin,* 54 (May 1957), 229–49.