***Reliability Studies***

J. PAUL PETER*

The basic theories and measurement procedures for reliability and the closely related concept of generalizability are reviewed, illustrated, and evaluated for use in marketing research. A critique is given of a subset of previous marketing research studies in which reliability estimates were used and recommendations are made for future research.

# Reliability: A Review of Psychometric Basics and Recent Marketing Practices

Valid measurement is the *sine qua non* of science. In a general sense, validity refers to the degree to which instruments truly measure the constructs which they are intended to measure. If the measures used in a discipline have not been demonstrated to have a high degree of validity, that discipline is not a science.

A necessary (but not sufficient) condition for validity of measures is that they are reliable. Reliability can be defined broadly as the degree to which measures are free from error and therefore yield consistent results. For example, ordinal level measures are reliable if they consistently rank order subjects in the same manner; interval level measures are reliable if they consistently rank order and maintain the distance between subjects (up to a linear transformation). Of course, behavioral measures are seldom if ever totally reliable and valid, but the degree of their validity and reliability must be assessed if research is to be truly scientific.

Marketing researchers seldom assess the reliability (much less the validity) of their measures (Heeler and Ray, 1972, p. 369). For example, in consumer behavior, which is traditionally viewed as a marketing area, Jacoby (1976, p. 6) reports that in the entire 300-item brand loyalty literature only one study has presented a measure of test-retest reliability for that construct. Rogers (1976, p. 299) states that there is a lack of evidence of the accuracy and stability over time of measures in the adoption-diffusion literature. Ryan and Bonfield (1975, p. 22) criticize the common use of single-item scales and lack of concern with reliability in the attitude-behavioral intention literature, and Kassarjian (1971, p. 415) points out that too often researchers are disinterested in reliability (and validity) criteria in the study of personality. The problem is not unique to these areas; the situation is similar with respect to measures of other core constructs (Jacoby, 1976, p. 6). In fact, of the more than 400 consumer behavior studies surveyed for this research, less than 5% assessed the reliability of the measures employed.

If the state of the art in marketing is to develop beyond its current condition, a useful starting point would be the regular assessment of reliability in marketing research studies and the development of highly reliable scales. Not only is reliability a necessary condition for validity, but unreliable measures attenuate (lessen) the correlation between measures. Thus, if reliability is not assessed and the correlation between measures of two constructs is low, marketing researchers have no way of knowing whether there is simply little relationship between the two constructs or whether the measures are unreliable.

The purpose of this article is to provide a resource for marketing researchers interested in understanding reliability theory and assessing the reliability of their measures. The first section is a discussion of traditional reliability theory and measurement. Though much of the psychometric literature has been concerned with analyzing appropriate assumption structures, formulas, and methods for assessing reliability, the focus here is on discussing basic concepts and evaluating

*J. Paul Peter is Associate Professor of Marketing, Washington University. The author gratefully acknowledges the many useful suggestions and contributions provided by the reviewers.



*Journal of Marketing Research*
Vol. XVI (February 1979), 6-17

reliability assessment procedures for use in marketing research.[1]

The second section is concerned with the reformulation of reliability as generalizability theory. Though generalizability theory has not replaced traditional reliability theory, it does provide a unified conceptual and operational approach for addressing reliability issues.

The final section is a review of applications of reliability assessment in one area of marketing, consumer behavior. Consumer behavior was selected for this review because it is the most heavily researched area in marketing and has borrowed most heavily from psychology. Thus, reliability was expected to be assessed more often in consumer behavior than in other areas of marketing research. However, this is not meant to imply that reliability assessment is appropriate only for consumer behavior constructs; reliability needs to be assessed regularly for constructs in all areas of marketing research.

## *TRADITIONAL APPROACHES TO RELIABILITY*

The voluminous literature on reliability began with the work of Spearman in 1904 and 1910. Though several other formulations and assumptions structures for deriving reliability theory have been advanced, Spearman's notion of true and error components remains the most influential model in psychological research (Campbell, 1976, p. 18). In this section the Spearman approach is discussed and basic methods for assessing the reliability of a measurement scale are evaluated. The term "scale" is used here to mean a multi-item scale and not simply a single item. One of the advantages of multi-item scales is that they allow measurement errors to cancel out against each other and thus the reliability of the scale is increased. In addition, multi-item scales may be necessary to approach valid measurement of factorially complex constructs.

### *Reliability Theory*

This basic approach starts with the notion that the mean and variance of any observed scale score can each be divided into two parts.[2] In terms of the mean, the two parts are the true score and the error score or

$$X_{\text{observed}} = X_{\text{true}} + X_{\text{error}}. \tag{1}$$

Conceptually, the true score is a perfect measure of the property being measured. However, in practice, the true score can never really be known and generally is assumed to be the mean score of a large number of administrations of the same scale to the same subject. The error score is an increase or decrease from the true score resulting from measurement error. Measurement error is the source of unreliability and its primary cause is that items in the scale are not measuring the same phenomenon.

The variance of an observed scale score also is assumed to have a true component and an error component or

$$V_{\text{observed}} = V_{\text{true}} + V_{\text{error}}. \tag{2}$$

The true variance component includes all systematic variance. In one sense, it is a misnomer because it includes both variance from the phenomenon under investigation and all other sources of systematic variance. (Determination of the difference between types of systematic variance is a validity question.) The error variance component includes all random or nonsystematic variance. In terms of the previous definition of reliability, systematic variance does not affect either the rank order or distance between subjects but random or error variance does and thus error variance lowers the reliability of measures. A reliability coefficient ($r_{tt}$), therefore, is nothing more than the ratio of true variance to observed variance or the percentage of total variance which is of the systematic type. Symbolically,

$$r_{tt} = \frac{V_{\text{true}}}{V_{\text{observed}}}. \tag{3}$$

Because $V_{\text{true}}$ cannot be estimated directly, equation 3 cannot be used to compute a reliability coefficient. However, because $V_{\text{true}} = 1 - V_{\text{error}}$ and $V_{\text{error}}$ can be estimated, equation 3 can be rewritten into a computational formula as

$$r_{tt} = 1 - \frac{V_{\text{error}}}{V_{\text{observed}}} \tag{4}$$

or by further multiplying through by $V_{\text{observed}}$,

$$r_{tt} = \frac{V_{\text{observed}} - V_{\text{error}}}{V_{\text{observed}}}. \tag{5}$$

These two equations are both theoretical and practical. As a theoretical matter, they exemplify the notion that measurement error (error or random variance) reduces the reliability of measures. As a practical matter, an analysis of variance approach has been suggested (e.g., Alexander, 1947; Burt, 1955; Hoyt, 1941) for estimating these sources of variance. Basically, the ANOVA model employs the mean square of the residual as an estimate of $V_{\text{error}}$, the mean square between individuals as an estimate of $V_{\text{observed}}$,

[1]Excellent reviews and comparisons of alternative formulations are provided by Bohrnstedt (1970), Campbell (1976), and Tryon (1957). The Campbell review is the most recent and includes some formulations not found in the earlier works. More comprehensive and technical treatments can be found in Lord and Novick (1968) and Cronbach et al. (1972).

[2]Parts of this section are based on Guilford (1954) and Kerlinger (1973). This is by no means the only rationale for deriving reliability theory and it has been subject to criticism (e.g., Tryon, 1957). However, in spite of its limitations, it provides a useful framework for discussing reliability concepts.

and substitutes each into equation 4 or 5.[3] However, this is a method of reliability assessment.

*Reliability Measurement*

There are three basic methods for assessing the reliability of a measurement scale: test-retest, internal consistency, and alternative forms.[4] All three methods attempt to determine the proportion of variance in a measurement scale that is systematic. Basically, these methods correlate scores obtained from a scale with scores from some form of replication of the scale. If the correlation is high, most of the variance is of the systematic type and, with some degree of consistency, the measures can be depended upon to yield the same results.

The basic difference among the three methods is in what the scale is to be correlated with to compute the reliability coefficient. In test-retest, the identical set of measures is applied to the same subjects at two different times. The two sets of obtained scores are then correlated. In internal consistency, a measurement scale is applied to subjects at one point in time; subsets of items within the scale are then correlated. In alternative forms, two similar sets of items are applied to the same subjects at two different times. Scale items on one form are designed to be similar (but not identical) to scale items on the other form. The resulting scores from the two administrations of the alternative forms are then correlated.

*Test-retest reliability.* In this method of reliability assessment the same scale is applied a second time to the same subjects under conditions as similar as the investigator can make them. The scores from the two administrations then are correlated and the resulting index is interpreted in terms of the stability of performance of the measures over time. A two-week interval is the generally recommended retest period.

The retest method involves at least three basic problems. First, different results may be obtained depending on the length of time between measurement and remeasurement. In general, the longer the time interval the lower the reliability (Bohrnstedt, 1970, p. 85). Second, if a change in the phenomenon occurs between the first and second administration, there is no way to distinguish between change and unreliability.[5] Third, the retest correlation is only partly dependent on the correlation between different items in the scale, because a portion of the correlation of sums includes the correlation of each item with itself. Such correlations would be expected to be much higher than those found between different items and could produce a substantial correlation between test and retest (Nunnally, 1967, p. 215).

Although the retest method provides useful information about the stability of measures, the problems suggest that it should not be used as the sole method of reliability assessment. Rather, if the retest method is employed, it should be supplemented with internal consistency estimates for each administration.

*Internal consistency reliability.* The basic form of this method is split-halves in which item scores obtained from the administration of a scale are split in half and the resulting half scores are correlated. The scale is usually split in terms of odd and even numbered items or on a random basis. Internal consistency measures assess the homogeneity of a set of items.

Though split-halves is the basic form of internal consistency estimate, there is one basic problem with using it: different results may be obtained depending on how the items are split in half. Thus, the researcher is faced with the bothersome question of what is the "real" reliability coefficient. One approach to overcoming this problem is to determine the mean reliability coefficient for all possible ways of splitting a set of items in half. A formula which accomplishes this step is Cronbach's (1951) coefficient alpha which is the most commonly accepted formula for assessing the reliability of a measurement scale with multi-point items.[6] Alpha is formulated as

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_{i=1}^{k} \sigma_i^2}{\sigma_t^2}\right) \tag{6}$$

where:

$k$ = number of parts (usually items) in the scale,
$\sigma_i^2$ = variance of item $i$, and
$\sigma_t^2$ = total variance of the scale.

[3]For computational examples, see Kerlinger (1973, p. 447-51). A more comprehensive ANOVA approach is presented in the generalizability section of this article.

[4]Strictly speaking, reliability is never really measured but only estimated. For convenience, however, the terms "measure," "estimate," and "assess" are used interchangeably in this article.

[5]See Heise (1969) and Wiley and Wiley (1971) for approaches to overcoming this problem.

[6]If items are scored dichotomously, e.g., yes-no, true-false, A vs. B, a special case of $\alpha$, Kuder-Richardson Formula 20 (Kuder and Richardson, 1937) is the appropriate formula. KR-20 is formulated as

$$\text{KR-20} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_{i=1}^{k} pq}{\sigma_t^2}\right)$$

where:

$k$ = number of items on the scale,
$p$ = proportion of responses of the first type,
$q$ = proportion of responses of the second type $(1 - p)$, and
$\sigma_t^2$ = total variance of the scale.

**Table 1**
LOWER HALF COVARIANCE MATRIX FOR SIX PROBABILITY OF LOSS ITEMS

| | | | | | |
|---|---|---|---|---|---|
| 3.49[a] | | | | | |
| 1.07 | 2.46 | | | | |
| 2.04 | .83 | 3.37 | | | |
| 1.45 | 1.62 | 1.97 | 3.62 | | |
| 1.10 | 1.00 | 1.80 | 1.61 | 3.62 | |
| 1.91 | .58 | 2.30 | 1.35 | 2.03 | 3.52 |

[a]Underlined values are the item variances.

Because of the facts that (1) alpha is one of the most important deductions from the theory of measurement error (Nunnally, 1967, p. 96) and (2) the majority of marketing research studies employ scales or items of the type alpha was designed to evaluate, a numerical example is provided hereafter. Though a computer program is available in the marketing literature for calculating alpha (Vigderhous, 1974), the calculations can be made easily from the covariance matrix of a set of items.

In a portion of a previously reported study (Peter and Ryan, 1976), the reliability of several six-item perceived risk scales was assessed by using coefficient $\alpha$. One scale was intended to measure the probability of six types of loss from the purchase of a Ford Pinto. The six types of loss examined were financial, social, performance, psychological, physical, and convenience. The items were scored from 1 (improbable) to 7 (probable) by 108 subjects. Table 1 is the covariance matrix for the resulting scores.

Because the total variance can be restructured as the sum of the item variances plus two times the sum of the item covariances, alpha can be restructured into a computational formula as[7]

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sum_{i=1}^{k}\sigma_i^2 + 2\sum_{i>}^{k}\sum_{j}^{k}\sigma_{ij}}\right). \tag{7}$$

The first step is to compute $\sum_{i=1}^{6}\sigma_i^2$ which is the sum of the item variances, the diagonal elements of the covariance matrix or

$$\sum_{i=1}^{6}\sigma_i^2 = (3.49 + 2.46 + 3.37 + 3.62 + 3.62 + 3.52) = 20.08.$$

The next step is to compute two times the sum of the covariance elements, the off-diagonal elements of the covariance matrix or

$$2\sum_{i=1}^{6}\sum_{j=1}^{6}\sigma_{ij} = 2(1.07 + 2.04 + 1.45 + 1.10 + 1.91 + .83 + 1.62 + 1.00 + .58 + 1.97 + 1.80 + 2.30 + 1.61 + 1.35 + 2.03) = 45.32.$$

Alpha then can be determined as

$$\alpha = \frac{6}{6-1}\left(1 - \frac{20.08}{20.08 + 45.32}\right) = .83.$$

Thus, the scale demonstrates a high degree of reliability, and correlations between the sum score of the scale items and other variables would be affected very little by attenuation.[8]

In addition to alpha, there are numerous other formulas for computing internal consistency estimates of reliability. However, alpha is a general formula and many other approaches have been shown to be different computational forms which will yield similar results (Tryon, 1957). Although some aspects of deriving alpha have been criticized (Bentler, 1972; Tryon, 1957), it is a most useful formula for assessing the reliability of measures in marketing research.

*Alternative form reliability.* In this method of assessing reliability, the same subjects are measured with two scales at two different times, usually two weeks apart. Each scale is designed to be similar in content to the other but different enough that the first measurement will not substantially affect remeasurement. The resulting scores from the two administrations of the alternative forms then are correlated to obtain a reliability coefficient. Alternative forms assess the equivalency of content of sets of items.

[7]Not only is equation 7 more convenient computationally, but also it illustrates why increasing the number of items on a scale almost always increases the scale's reliability—the number of covariance terms in the denominator increases geometrically with the number of items whereas the number of variance terms increases only arithmetically. An increase of $m$ items increases the variance by $m$ elements but the covariance by $m(m-1)$ elements. Thus, although $k/k-1$ in the formula decreases with an increase in the number of items and the additional variance would have a negative effect on the value of alpha, the geometric increase in the covariance elements more than offsets these effects. Thus, unless all of the covariance of added items with the original items is almost zero and the variance of the added items is not, an increase in the number of items on a scale will increase the reliability of the scale.

[8]Given a reliability estimate, the standard error of measurement can be computed by the formula

$$\sigma_{meas} = \sigma_t\sqrt{1 - r_{tt}}$$

where:

$\sigma_{meas}$ = the standard error of measurement,
$\sigma_t$ = the standard deviation of total scores, and
$r_{tt}$ = the reliability estimate.

In the example, $\sigma_{meas} = 8.08\sqrt{1 - .83} = 3.33$. Although confidence intervals around true scores can be computed by using this estimate (see Bohrnstedt, 1970, p. 83–4; Nunnally, 1967, p. 220) such analysis has little practical value for most types of marketing research.

The primary problem with use of alternative forms is in the development of substantially equivalent alternative measures. For example, strict definitions of alternative forms state that the mean, variance, and intercorrelation of items on each form must be equivalent (Gulliksen, 1950). Though this problem has been overcome to some extent in educational testing, it remains a serious consideration for the measurement of other behavioral constructs.

An even more perplexing problem with alternative forms is "proving" that the two measures are equivalent in content. For example, if the correlation between scores on the two forms is low, it is difficult to determine whether the measures have intrinsically low reliability or whether one of the forms is simply not equivalent in content to the other. Although this problem can be investigated (Nunnally, 1967, p. 211–13), the end result most likely will be the development of yet another alternative form.

The importance of assessing reliability with alternative forms depends on the phenomenon under investigation. If the phenomenon is expected to vary over relatively short periods of time, then alternative form measures may be necessary for examining changes. Thus, though the alternative form method may be necessary for the investigation of some marketing constructs, coefficient alpha usually will provide a close estimate of alternative forms reliability (Nunnally, 1967, p. 211).

## *REFORMULATING RELIABILITY AS GENERALIZABILITY THEORY*

One problem not explicitly addressed by traditional approaches to reliability is that measurement error can come from many sources and each definition of error changes the meaning of the "reliability coefficient" (Gleser et al., 1965). For example, the components of variance which are "true" and "error" in computing a test-retest correlation are different from those for an internal consistency estimate. One approach to simultaneously analyzing multiple sources of variance in a measurement procedure has been the reformulation of reliability as generalizability theory by Cronbach and his associates (1963, 1972; Gleser et al., 1965).

### *Generalizability Theory*

Generalizability theory is based in part on the concept of sampling. However, the primary focus is not on the sampling of people from populations of people, but rather on sampling "conditions of measurement" from universes of possible measurement conditions.

A condition of measurement (or simply, condition) is a specific aspect of a measurement procedure, e.g., the specific times measures are taken or the specific items used in the scale. The general term for an aspect of a measurement procedure is a "facet of measurement," i.e., time, instrument, and observers are several common facets.

The question in generalizability theory is whether scores obtained in the sampled conditions of measurement are representative of the universe scores for those conditions. The universe score is analogous to the true score in traditional reliability theory and is conceptualized as the mean score a subject would provide if measured over all conditions in a universe.

Measures cannot be taken over all conditions in a universe. However, it is still possible to determine the correlation between observed scores and universe scores, because it is assumed in generalizability theory that conditions of measurement are randomly sampled from the universe of conditions. Given this assumption, the correlation between any two sets of observed scores for a condition can be shown, on the average, to be equal to the correlation between observed scores and universe scores (Campbell, 1976, p. 144). Thus, a "coefficient of generalizability" can be determined from observed scores and is defined as the ratio of universe-score variance to expected observed-score variance (Cronbach et al., 1972, p. 17).

Though generalizability coefficients can be computed for a measurement procedure, this is not the main thrust of generalizability research. The main goal of generalizability research is to simultaneously assess the components of variance in a measurement procedure. Variance can come from many sources in a measurement procedure and, depending on the interest of the researcher, at least one of these sources is unwanted or "error." By simultaneously investigating multiple sources of variance, the researcher can develop more efficient measurement procedures.

The main benefit of generalizability theory is that it explicitly recognizes that there are many universes to which a researcher may wish to generalize. For example, the researcher may wish to generalize from a sample of items to the universe of items, from a sample of times of measurement to the universe of times of measurement, from a sample of places of measurement to the universe of places of measurement, from a sample of observers to the universe of observers, etc. Measurement procedures are designed in generalizability studies to investigate the universes of interest by sampling conditions of measurement from each of them. In other words, for each universe of interest, a facet is included in the generalizability study. Thus, traditional reliability methods can be viewed as single-facet generalizability studies, e.g., a test-retest correlation is concerned with whether scores obtained from a measurement instrument are generalizable to the universe scores across all times of possible measurement. Clearly, even if the test-retest correlation were high, no statement could be made about the generalizability of the measures to other universes. To generalize to other universes, other measurement procedures would have to be employed.

Table 2
ANALYSIS OF VARIANCE FOR HYPOTHETICAL BRAND LOYALTY *G* STUDY

| *Source of variance* | *Sum of squares* | *d.f.* | *Mean square* |
|---|---|---|---|
| Subjects ($p$) | 5999.400 | 99 | 60.600 |
| Items ($i$) | 1087.200 | 9 | 120.800 |
| Occasions ($j$) | 6540.000 | 2 | 3270.000 |
| $pi$ | 1335.609 | 891 | 1.499 |
| $pj$ | 712.800 | 198 | 3.600 |
| $ij$ | 67.500 | 18 | 3.750 |
| Residual | 2227.500 | 1782 | 1.250 |

*Generalizability Measurement*

As previously noted, analysis of variance could be used to assess the reliability of a measurement scale. The logic of the ANOVA model can be expanded to include multiple sources of variance and this is precisely the method used to assess the various components of variance in generalizability research.

For example, suppose a researcher is interested in investigating a brand loyalty measurement procedure. Eventually, the measurement procedure is to be used to help make a decision about whether or not a particular brand should be dropped from a product line.

The first step would be to perform a generalizability ($G$) study. A $G$ study is used to estimate variance components in the measurement procedure. If the $G$ study indicates that the components of variance which the researcher considers "error" are minimal, then the measurement procedure would be considered acceptable for use in a decision ($D$) study. The $D$ study would be the actual study of brand loyalty on which the decision to drop the brand from the product line would be based.

The design of the $G$ study depends on the facets of interest to the researcher. Suppose the researcher were interested in two facets, time and instrument, and therefore administered a 10-item brand loyalty scale to 100 subjects on three occasions. The first step in the analysis of the $G$ study would be to perform a three-way analysis of variance, subjects × items × occasions. Table 2 is an analysis of variance table for this hypothetical problem.

Because subjects, items, and occasions are assumed to be randomly sampled, the observed mean squares are unbiased estimates of the expected mean square (EMS). These expected mean squares also have been shown by Cornfield and Tukey (1956) to be the sum of certain variance components, and it is these variance components which are of primary interest in generalizability research. The formulas for obtaining the seven sources of variance of interest in the sample problem are:

$$(8)\quad \mathrm{EMS}_p = \sigma_e^2 + n_i\sigma_{pj}^2 + n_j\sigma_{pi}^2 + n_i n_j\sigma_p^2$$

$$(9)\quad \mathrm{EMS}_i = \sigma_e^2 + n_p\sigma_{ij}^2 + n_j\sigma_{pi}^2 + n_p n_j\sigma_i^2$$

$$(10)\quad \mathrm{EMS}_j = \sigma_e^2 + n_i\sigma_{pj}^2 + n_p\sigma_{ij}^2 + n_p n_i\sigma_j^2$$

$$(11)\quad \mathrm{EMS}_{pi} = \sigma_e^2 + n_j\sigma_{pi}^2$$

$$(12)\quad \mathrm{EMS}_{pj} = \sigma_e^2 + n_i\sigma_{pj}^2$$

$$(13)\quad \mathrm{EMS}_{ij} = \sigma_e^2 + n_p\sigma_{ij}^2$$

$$(14)\quad \mathrm{EMS}_{\mathrm{res}} = \sigma_e^2.$$

To determine the unknown variance components, the analysis starts with $\sigma_e^2$ which is 1.25.[9] With this estimate, and because every term in each of the equations is known except one of the variance components, the researcher can solve for the variance components. Because each of the observed mean squares is an unbiased estimate of its respective EMS, the EMS terms are taken from Table 2; the number of persons, $n_p = 100$; the number of items, $n_i = 10$; the number of occasions, $n_j = 3$. For example, to solve for the variance component of the $ij$ interaction term, the substitution would be:

$$\mathrm{EMS}_{ij} = \sigma_e^2 + n_p\sigma_{ij}^2$$
$$3.75 = 1.25 + 100\sigma_{ij}^2$$
$$\sigma_{ij}^2 = .025.$$

Working backward through each of the equations, one can estimate each of the components of variance. In addition, by summing the variance components and then dividing the sum into each component, one can determine the percentage of total variance which is attributable to each source. The results of these computations for the sample problem are reported in Table 3.

For the purpose of a $G$ study, analysis is made directly from these components of variance. The large effects are those from subjects, occasions, and the residual. The large subject variance component would suggest that the sample is rather heterogeneous and the measurement procedure is capable of discriminating between the various subjects.

The major source of variance is from occasions which would suggest that subjects are not responding in the same way at different times of measurement. However, another possibility might be that different coders were employed at different times of measurement and this effect is masked because "coders" was not included as a facet in the study. Subsequent analysis including coders as a separate facet could be performed to determine the extent of this effect. The high residual could stem from other facets affecting the scores which are not explicitly accounted for in the design.

[9] It would be more proper to label $\sigma_e^2$ as $\sigma_{pij,e}^2$ to account for the within-cell error term as there is only one observation per cell in this design. However, $\sigma_e^2$ is used here to simplify the notation.

Table 3
ESTIMATED COMPONENTS OF VARIANCE AND PERCENTAGES OF TOTAL VARIANCE FOR HYPOTHETICAL BRAND LOYALTY STUDY

| *Source of variance* | *Mean square* | *d.f.* | *Estimate of variance component* | *Percentage of total variance* |
|---|---|---|---|---|
| Subjects ($p$) | 60.60 | 99 | 1.892 | 26.51 |
| Items ($i$) | 120.80 | 9 | .389 | 5.45 |
| Occasions ($j$) | 3270.00 | 2 | 3.264 | 45.73 |
| $pi$ | 4.20 | 891 | .083 | 1.16 |
| $pj$ | 3.60 | 198 | .235 | 3.29 |
| $ij$ | 3.75 | 18 | .025 | .35 |
| Residual | 1.25 | 1782 | 1.250 | 17.51 |
| | | | $\sigma_r^2 = 7.138$ | 100.00% |

The coefficient of generalizability can be determined by dividing an estimate of universe-score variance by an estimate of expected observed-score variance. In this example, the estimate of the universe-score variance is the variance component for subjects, 1.892. The estimate of the expected observed-score variance is the sum of the variance components for $p$, $pi$, $pj$, and $e$ or 1.892 + .083 + .235 + 1.250 = 3.46. The coefficient of generalizability thus equals 1.892/3.46 or .547.

Though the illustrative study used here is a completely crossed, two-facet design, the possibilities for study designs for generalizability are unlimited. Not only can any number of facets be examined simultaneously, but nested designs or designs with any number of fixed and random facets can be accommodated.

By formulating reliability problems in terms of generalizability, researchers are forced to recognize that unwanted or error variances could come from many sources in a measurement procedure. Even if reliability is assessed by use of a traditional method, only one facet of measurement and type of error variance is being explicitly considered. Thus, a high reliability coefficient of one type cannot be interpreted to mean that the measurement procedure will yield consistent results across all potential facets.

Although the generalizability formulation provides a framework for extensive investigation of measurement procedures, two practical considerations may limit its use in marketing research. First, the design and interpretation of generalizability studies can become very complex. For example, interpreting a higher order three- or four-way interaction can be a most challenging task. Second, whether generalizability studies are profitable (reward-cost) depends on how important the various sources of error variance are expected to be in a measurement procedure. In most cases, the primary source of measurement error is from the items in a measurement scale. In other words, the scale items do not systematically measure the same phenomenon. Thus, even though coefficient alpha does not consider many sources of measurement error, Nunnally (1967, p. 210–11) suggests that it is surprising what little difference these other sources usually make. This is particularly true for situations in which instructions are easily understood and there is little subjectivity in scoring. Thus, the necessity of performing a multifaceted generalizability study depends on how important sources of variance other than the items are expected to be.

## *RELIABILITY ASSESSMENT PRACTICES IN MARKETING RESEARCH*

To investigate reliability assessment practices, a sample of 400 empirical research studies was surveyed in one area of marketing, consumer behavior. These studies were surveyed primarily from the *JMR* (1972–1976), *JCR* (1974–1976), *JM* (1974–1976), *Advances in Consumer Research* (1975–1977), and the *AMA Proceedings* (1974–1976).[10] Studies which were found to include some form of reliability assessment were reviewed closely to investigate problems in the area. There is no intent here to be critical of past research efforts and in fact these studies are laudable for at least attempting to address the reliability issue.

Tables 4 and 5 summarize the results of the survey. No studies were found which included either an alternative form or a multifaceted generalizability approach. This outcome could be expected because alternative forms are often difficult to develop and generalizability is a relatively new approach.

Nineteen studies were found which included reliability estimates. Two studies (Best et al., 1977; Lundstrom and Lamont, 1976) employed both test-retest

[10]In addition, a computer scan of the *Psychological Abstracts* was performed by the Bibliographical Services for Research system. In this scan, the keywords Consumer Behavior, Consumer Research, and Consumer Attitudes were cross-referenced with both test reliability and reliability. The scan produced a total of nine references for the years 1974–1976, of which only two were empirical research employing a reliability measure. Thus, this method was not effective in locating appropriate articles because reliability estimates are not always reported in the abstracts.

## Table 4
SUMMARY OF STUDIES INCLUDING TEST-RETEST RELIABILITY ESTIMATES

| *Authors* | *Study area* | *Nature of scales* | *Type of scale*[a] | *No. of items in scale* | *No. of points per item* | *Retest period* | *Reliability coefficient* | *Sample size* |
|---|---|---|---|---|---|---|---|---|
| Best, Hawkins, & Albaum (1977)[b] | Attitudes | Beliefs (about 5 department stores) | S | 10 per store | 6 | 10 days | range .41–.61 | 70 |
| Bettman, Capon, Lutz (1975) | Attitudes | Beliefs/evaluations (for four brands of toothpaste) | S | 36 | 11 | Immediate | .672[c] | 72 |
| Brooker (1975) | Consumer self-actualization | Consumer self-actualization | A vs. B | 20 | 2 | 4 weeks<br>5 weeks | .57<br>.67 | 13<br>24 |
| Green & Devita (1974) | Consumer utility for item collection | Personal preference for meal and dessert combinations[d] | Category ratings | 15 | 9 | 1 hour | range .8–.95 | 27 |
| Landon (1974) | Self-concept | Self-concept/product congruity | S | 1 | 9 | 1 week | ≥ .7[e] | 352 |
| | | Ideal self-concept/product congruity | S | 1 | 9 | | ≥ .7 | |
| | | Purchase intentions | S | 1 | 9 | | ≥ .7 | |
| Lundstrom & Lamont (1976) | Consumer discontent | Attitudes toward business practices | L | 84 | 6 | 6 weeks | .79 | 154 |
| Villani & Wind (1975) | Personality | Personality traits: Sociable | L | 6 | 5 | 2 years | .72[f] | 504 |
| | | Relaxed | L | 4 | 5 | | .68 | |
| | | Internal control | L | 5 | 5 | | .48 | |
| Waung (1975) | Attitudes | Perceived instrumentality (for 3 prescription drugs) | S | 5 per drug | 7 | Immediate | range .38–1.0 (intrasubj.) | 55 |
| Wright (1975) | Cognitive resistance to advertising | Generalized self-confidence | S | 10 | 9 | 2 weeks | .75 | 160 |
| | | Information processing confidence | S | 10 | 9 | | .69 | |

[a] L = Likert type.
S = Semantic differential type.
[b] See also Best et al. (1976).
[c] The within-individual correlations between two replicates of 36 responses were averaged across subjects.
[d] Replicates of 15 cards listing various meal and dessert combinations were sorted into nine preference groups.
[e] Test-retest correlations (≥ .7) were used to determine which products would be included in the study. Twelve were selected for 179 males; seven were selected for 173 females.
[f] Test-retest correlations measured by simple sum of factor analysis items were .70, .72, .39 and measured by factor scores were .65, .55, .37 for the three scales respectively.

and internal consistency estimates and are included in both tables. Seven other studies employed test-retest and 10 others used internal consistency estimates.

In terms of content, 12 of the studies were in the attitude/behavioral intentions and personality/lifestyle areas. This finding can be explained because (1) these areas were the most heavily researched in the review period and (2) multi-item scales were available (or adaptable) from other disciplines in these areas. The majority of the studies used semantic differential or Likert-type items and "the magical number 7 plus or minus 2" accounted for the majority of points per item.

A difficult problem is the number of items necessary on a scale. Conceptually, the answer is simply enough items to measure the construct under investigation, and only that construct. Yet there are at least two practical problems in determining the appropriate number of items for marketing constructs. One is the problem of boredom and fatigue if too many items are included on a questionnaire to measure each construct. Although the greater the number of items,

## Table 5
### SUMMARY OF STUDIES INCLUDING INTERNAL CONSISTENCY RELIABILITY ESTIMATES

| *Authors* | *Study area* | *Nature of scales* | *Type of scale*[a] | *No. of items in scales* | *No. of points per item* | *Internal consistency estimate* | *Reliability coefficient* | *Sample size* |
|---|---|---|---|---|---|---|---|---|
| Best, Hawkins, & Albaum (1977) | Attitudes | Beliefs (about 5 department stores) | S | 10 per store | 6 | ANOVA<br>α | range .62–.71<br>range .56–.64 | 70 |
| Darden & Perreault (1975) | Lifestyle | 12 lifestyle covariates[b] | L | range 2–5 | 6 | split-half | range .68–.88 | 359 |
| Darden & Perreault (1976) | Lifestyle | 15 lifestyle scales[c] | L | range 2–4 | NR | split-half | range .52–.83 | 278 |
| Darden & Reynolds (1974) | Innovation | 13 AIO's[d] | L | range 4–10 | 6 | split-half | range .55–.89 | 154 |
| | | Apparel innovativeness | L | 9 | NR | split-half | .67 | |
| | | Grooming innovativeness | L | 6 | NR | split-half | .59 | |
| | | Home care innovativeness | L | 4 | NR | split-half | .61 | |
| Leavitt & Walton (1975) | Personality | Innovativeness | NR[e] | 40 | 5 | Spearman-Brown | .90 | 299 |
| Lundstrom & Lamont (1976) | Consumer discontent | Attitudes toward business practices | L | 84 | 6 | KR-20<br>split-half<br>Spearman-Brown | .88<br>.96<br>.94 | 226 |
| Moschis (1976) | Informal group influences | Group influence | L | 4 | 5 | split-half | .76 | 206 |
| | | Reflected appraisal | L | 4 | 5 | split-half | .65 | |
| | | Comparative appraisal | L | 3 | 5 | split-half | .65 | |
| Perry (1973) | Attitudes/personality | Attitudes toward: | | | | | | |
| | | Alcohol | L | 20 | 7 | α | .68 | 164 |
| | | Cigarettes | L | 20 | 7 | α | .70 | |
| | | Coffee | L | 20 | 7 | α | .65 | |
| Peter & Ryan (1976) | Perceived risk | Probability of loss | S | 6 per brand | 7 | α | range .72–.83 | 210[f] |
| | | Importance of loss (for six auto brands) | S | 6 per brand | 7 | α | range .55–.75 | |
| Ryan & Becherer (1976) | Personality | Personality traits: | | | | | | |
| | | Compliant | S | 10 | 6 | α | .724 | 175 |
| | | Aggressive | S | 15 | 6 | α | .680 | |
| | | Detached | S | 10 | 6 | α | .514 | |
| Ryan & Peter (1976) | Behavioral (purchase) intentions | Attitudinal influence | S | 4 per brand | 7 | α | .93, .94 | 97 |
| | | Social influence | L | 14 per brand | 7 | α | .86, .91 | |
| | | Purchase intentions (for 2 brands of toothpaste) | S | 3 per brand | 7 | α | .97, .98 | |
| Wilson, Mathews, & Harvey (1975) | Behavioral (purchase) intentions | Components of behavioral intentions model for six brands of toothpaste | S | NR | 7 | KR-20[g]<br>Lambda-3<br>α | range .665–.786<br>NR<br>NR | 162 |

[a] L = Likert type items.
S = Semantic differential type items.
NR for any entry = not reported.

[b] Scales were determined by employing factor analysis and labeled generalized self-confidence, opinion leadership, plan-ahead traveler, information seeker, camp traveler, relaxing traveler, first-class traveler, national traveler, jetsetter-vagabond traveler, historical traveler, sports-spectator, functional gregarious.

[c] Scales were determined by employing factor analysis and labeled price conscious, fashion conscious, dislikes housekeeping, community-minded, self-confidence, opinion leadership, information seeker, new brand trier, canned food user, dieter, financial optimist, wide horizons, arts enthusiast, patronage innovator, patronage opinion leader.

the higher the scale's reliability (see footnote 7), it may be necessary to prune extremely long scales to shorter forms. This step can be accomplished by selecting a subset of items which have high covariances.

A second problem is factorially complex constructs, i.e., constructs which contain multiple dimensions. For example, there are many types of perceived risk (e.g. financial, social, etc.) and perhaps a multi-item scale is needed for each type. The problem is even more complex in measurement of the attributes of various products, services, stores or brands. For example, in investigation of attitudes toward various brands, common formulations view the construct of attitudes as some additive combination of product or brand attributes. These attribute measures could be viewed in two ways: (1) as single-item components of the construct of attitudes or (2) as separate dimensions. If they are viewed as single-item components, then assessing the reliability of a scale composed of a set of attribute measures makes sense. However, if each attribute is viewed as a separate dimension, such as the taste versus price of a brand of toothpaste, then a different procedure is required. In this case a separate, multi-item, internally consistent scale would be required for each attribute in order to approach valid measurement of the attitude construct. Though the latter approach is more consistent conceptually and in terms of the measurement literature, developing multi-item subscales for each attribute may be tedious, as would filling out questionnaires composed of sets of such highly redundant items.

In terms of the retest periods used in the studies in Table 4, at least five of the retest periods were of such short duration that the initial measurement could have substantially affected remeasurement. An additional problem with test-retest was illustrated in many of the studies—the sample size (of subjects) had to be reduced because some subjects were not available for the retest.

In terms of the internal consistency estimates in Table 5, five studies used the split-half procedure which leaves open the question of whether different results would be obtained if the items had been split in half in another manner. Two studies employed the Spearman-Brown formula which is used primarily to determine what the reliability of the scale would be if the number of items were doubled.[11] One study used the analysis of variance approach and the remainder employed coefficient alpha.

Reviewing the reliability coefficients reported in the two tables raises the question of whether the scales demonstrated satisfactory levels of reliability. Though no hard and fast rules have been offered for evaluating the magnitude of reliability coefficients, Nunnally (1967, p. 226) suggests the following guidelines. In early stages of research, modest reliability in the range of .5 to .6 will suffice. For basic research, it is argued that increasing reliability beyond .8 is unnecessary because at that level correlations are attenuated very little by measurement error. Nunnally says that in applied settings, in contrast to basic research, a reliability of .9 is the minimum that should be tolerated and a reliability of .95 should be considered the desirable standard. Thus, as none of the studies reported were of an applied nature, these guidelines suggest that in almost all cases the scales demonstrated at least satisfactory reliability.[12] In marketing, guidelines are yet to be developed.

A final question is related to the sample sizes used in these studies because sampling error can make results appear better than they will be in subsequent studies. In other words, sampling errors provide the opportunity to take advantage of chance and such opportunities are related positively to the number of

[11] Although a test of double length is the most common application, the generalized Spearman-Brown formula can be used for any ratio of altered test length to the original length (Guilford, 1954, p. 354). The generalized formula is

$$r_{nn} = \frac{nr_{tt}}{1 + (n-1)r_{tt}}$$

where:

$r_{nn}$ = the reliability of the altered-length scale,
$n$ = any proportionate change in test length, and
$r_{tt}$ = the reliability of the original length scale.

[12] This is not to say that Nunnally's guidelines should be interpreted as absolute standards for marketing research. They are primarily concerned with the development of finely tuned measures of individual traits to be used for decisions about individual persons (e.g., GMAT tests). As most marketing research is not of this nature, lower levels of reliability may be acceptable in marketing research studies.

[d] Scales were determined by employing factor analysis and labeled interest-made fashions, interest-personal grooming, interest-home care, information seeking—male fashions, information seeking—personal grooming, information seeking—home care, opinion leadership—male fashions, opinion leadership—home care products, self-esteem, generalized self-confidence, attitude toward change, lifestyle venturesomeness.

[e] Not reported; a listing of items is available from the authors. The 40 items were also split into two groups of 20. Spearman-Brown and KR-20 for the forms were .84, .80 and .77, .77, respectively. Items were scored on five points and collapsed to two points for reliability assessment.

[f] 108 subjects for three brands of compact cars and 102 subjects for three brands of intermediate cars.

[g] In that the items are multipoint, the reported reliability coefficients are perhaps mislabeled as KR-20 rather than alpha.

items and related negatively to the number of subjects. A useful guideline suggests that for any type of item analysis (or multivariate analysis) there should be at least 10 times as many subjects as items or, in cases involving a large number of items, at least five subjects per item (Nunnally, 1967, p. 280). Of the studies reported in Tables 4 and 5, only five have at least a 10 to 1 ratio of subjects to items for each scale employed. Thus, many of the reported coefficients might be considerably smaller in subsequent research with adequate sample sizes.

## *CONCLUSIONS AND RECOMMENDATIONS*

This research was concerned with reliability and generalizability theory, measurement, and assessment practices in marketing. Given the state of the art of measurement in marketing, four recommendations are offered.

First, marketing researchers need to develop multi-item scales to measure constructs in the area. Most constructs by definition are too complex to be measured effectively with a single item, and multi-item scales are necessary for appropriate reliability and validity assessment. Despite encouraging signs of multi-item scale development found in this review, more and better multi-item scales need to be developed.

Second, the development of reliable scales presents a useful starting point for improving the quality of marketing research. If multi-item scales are developed which initially demonstrate low reliability, reliability often can be increased to acceptable levels by improving the clarity of the instructions, reducing ambiguity in the items, or by simply adding similar items to the scale.

Third, in reporting reliability coefficients, researchers should fully explain (1) relevant scale characteristics, (2) the procedure used to assess reliability and the source(s) of error which is treated, (3) appropriate references and previous reliability estimates (if any) for the scale, and (4) the interpreted meaning of the reliability coefficient. This information may help to overcome the problems of ambiguity in the area. For example, on the basis of the use of an internal consistency estimate and the researcher's background, internal consistency estimates have been referred to as measures of reliability, validity, homogeneity, and generalizability.

Last, coefficient alpha offers a useful and usable approach to assessing the reliability of measurement scales in marketing research. Though the development of alternative forms and multifacet generalizability studies will be needed for situations in which time and other facets of measurement need investigation, alpha can be fruitfully employed for scales containing a minimum of three items. Clearly, the development of reliable scales is a necessary condition for improving the quality of marketing research and theory.

## REFERENCES


Alexander, H. W. "The Estimation of Reliability When Several Trials are Available," *Psychometrika*, 12 (June 1947), 79–99.

Bentler, P. M. "A Lower-Bound Method for the Dimension-Free Measurement of Internal Consistency," *Social Science Research*, 1 (December 1972), 343–57.

Best, Roger, Del I. Hawkins, and Gerald Albaum. "The Role of Random Weights and Reliability in the Assessment of Multiattribute Attitude Models," *in* B. B. Anderson, ed., *Advances in Consumer Research*, Volume 3. Chicago: Association for Consumer Research, 1976, 88–91.

———, ———, and ———. "Reliability of Measured Beliefs in Consumer Research," *in* W. D. Perreault, ed., *Advances in Consumer Research*, Volume 4, Atlanta: Association for Consumer Research, 1977, 19–23.

Bettman, James R., Noel Capon, and Richard J. Lutz. "Multiattribute Measurement Models and Multiattribute Attitude Theory: A Test of Construct Validity," *Journal of Consumer Research*, 1 (March 1975), 1–15.

Bohrnstedt, George W. "Reliability and Validity Assessment in Attitude Measurement," *in* G. F. Summers, ed., *Attitude Measurement*. Chicago: Rand McNally, 1970, 81–99.

Brooker, George. "An Instrument to Measure Consumer Self-Actualization," *in* M. J. Schlinger, ed., *Advances in Consumer Research*, Volume 2. Chicago: Association for Consumer Research, 1975, 563–75.

Burt, C. "Test Reliability Estimated by Analysis of Variance," *British Journal of Statistical Psychology*, 8 (November 1955), 103–18.

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

Cornfield, J. and J. W. Tukey. "Average Values of Mean Squares in Factorials," *Annals of Mathematical Statistics*, 27 (December 1956), 907–49.

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

———, N. Rajaratnam, and G. C. Gleser. "Theory of Generalizability: A Liberalization of Reliability Theory," *British Journal of Statistical Psychology*, 16 (November 1963), 137–63.

———, G. C. Gleser, H. Nanda, and N. Rajaratnam. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York: John Wiley & Sons, Inc., 1972.

Darden, William R. and W. D. Perreault, Jr. "A Multivariate Analysis of Media Exposure and Vacation Behavior with Life Style Covariates," *Journal of Consumer Research*, 2 (September 1975), 93–103.

——— and ———. "Identifying Interurban Shoppers: Multiproduct Purchase Patterns and Segmentation Profiles," *Journal of Marketing Research*, 13 (February 1976), 51–60.

——— and Fred D. Reynolds. "Backward Profiling of Male Innovators," *Journal of Marketing Research*, 11 (February 1974), 79–85.

Gleser, G. C., L. J. Cronbach, and N. Rajaratnam. "Generalizability of Scores Influenced by Multiple Sources of Variance," *Psychometrika*, 30 (December 1965) 395–418.

Green, Paul E. and Michael T. Devita. "A Complementarity

Model of Consumer Utility for Item Collection," *Journal of Consumer Research,* 1 (December 1974), 56-67.

Guilford, J. P. *Psychometric Methods.* New York: McGraw-Hill Book Company, 1954.

Gulliksen, H. *Theory of Mental Tests.* New York: John Wiley & Sons, Inc., 1950.

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

Hiese, D. R. "Separating Reliability and Stability in Test-Retest Correlations," *American Sociological Review,* 34 (February 1969), 93-101.

Hoyt, C. "Test Reliability Estimated by Analysis of Variance," *Psychometrika,* 6 (June 1941), 153-60.

Jacoby, Jacob. "Consumer Research: Telling It Like It Is," *in* B. B. Anderson, ed., *Advances in Consumer Research,* Volume 3. Chicago: Association for Consumer Research, 1976, 1-11.

Kassarjian, Harold H. "Personality and Consumer Behavior: A Review," *Journal of Marketing Research,* 8 (November 1971), 409-18.

Kerlinger, F. N. *Foundations of Behavioral Research,* 3rd ed. New York: Holt, Rinehart and Winston, 1973.

Kuder, G. F. and M. W. Richardson. "The Theory of the Estimation of Test Reliability," *Psychometrika,* 2 (September 1937), 151-60.

Landon, E. Laird, Jr. "Self Concept, Ideal Self Concept, and Consumer Purchase Intentions," *Journal of Consumer Research,* 1 (September 1974), 44-51.

Lawlis, G. F. and E. Lu. "Judgment of Counseling Process: Reliability, Agreement, and Error," *Psychological Bulletin,* 78 (July 1972), 17-20.

Leavitt, Clark and John Walton. "Development of a Scale for Innovativeness," *in* M. J. Schlinger, ed., *Advances in Consumer Research,* Volume 2. Chicago: Association for Consumer Research, 1975, 545-54.

Lord, F. M. and M. R. Novick. *Statistical Theories of Mental Test Scores.* Reading, Massachusetts: Addison-Wesley, 1968.

Lundstrom, William J. and Lawrence M. Lamont. "The Development of a Scale to Measure Consumer Discontent," *Journal of Marketing Research,* 13 (November 1976), 373-81.

Moschis, George P. "Social Comparison and Informal Group Influences," *Journal of Marketing Research,* 13 (August 1976), 237-44.

Nunnally, J. *Psychometric Methods.* New York: McGraw-Hill Book Co., 1967.

Perry, Arnon. "Heredity Personality Traits, Product Attitude, and Product Consumption—An Exploratory Study," *Journal of Marketing Research,* 10 (November 1973), 376-9.

Peter, J. Paul and Michael J. Ryan. "An Investigation of Perceived Risk at the Brand Level," *Journal of Marketing Research,* 13 (May 1976), 184-8.

Rogers, Everett. "New Product Adoption and Diffusion," *Journal of Consumer Research,* 2 (March 1976), 290-301.

Ryan, Michael J. and Richard C. Becherer. "A Multivariate Test of CAD Instrument Construct Validity," *in* B. B. Anderson, ed., *Advances in Consumer Research,* Volume 3. Chicago: Association for Consumer Research, 1976, 149-54.

——— and E. H. Bonfield. "The Fishbein Extended Model and Consumer Behavior," *Journal of Consumer Research,* 2 (September 1975), 118-36.

——— and J. Paul Peter. "Two Operational Modifications for Improving the Delineation of Attitudinal and Social Influences on Purchase Intentions," *in* K. Bernhardt, ed., *Marketing: 1776-1976 and Beyond.* Chicago: American Marketing Association, 1976, 147-50.

Spearman, C. "The Proof and Measurement of the Association Between Two Things," *American Journal of Psychology,* 15 (January 1904), 72-101.

———. "Correlation Calculated from Faulty Data," *British Journal of Psychology,* 3 (October 1910), 271-95.

Tryon, R. C. "Reliability and Behavior Domain Validity: Reformation and Historical Critique," *Psychological Bulletin,* 54 (May 1957), 229-49.

Vigderhous, G. "Coefficient of Reliability Alpha," *Journal of Marketing Research,* 11 (May 1974), 194.

Villani, Kathryn E. A. and Yoram Wind. "On the Usage of 'Modified' Personality Trait Measures in Consumer Research," *Journal of Consumer Research,* 2 (December 1975), 223-8.

Waung, Sherren. "Explaining Behavior with Weighted, Unweighted and Standardized Attitude Scores," *in* M. J. Schlinger, ed., *Advances in Consumer Research,* Volume 2. Chicago: Association for Consumer Research, 1975, 345-55.

Wiley, D. E. and J. A. Wiley. "The Estimation of Measurement Error in Panel Data," *in* H. M. Blalock, Jr., ed., *Causal Models in the Social Sciences.* Chicago: Aldine-Atherton, 1971.

Wilson, David T., H. Lee Mathews, and James W. Harvey. "An Empirical Test of the Fishbein Behavioral Intentions Model," *Journal of Consumer Research,* 1 (March 1975), 39-48.

Wright, Peter. "Factors Affecting Cognitive Resistance to Advertising," *Journal of Consumer Research,* 2 (June 1975), 1-9.