A Meta-analysis of Cronbach's Coefficient Alpha

ROBERT A. PETERSON*

Despite some limitations, Cronbach's coefficient alpha remains the most widely used measure of scale reliability. The purpose of this article was to empirically document the magnitudes of alpha coefficients obtained in behavioral research, compare these obtained values with guidelines and recommendations set forth by individuals such as Nunnally (1967, 1978), and provide insights into research design characteristics that may influence the size of coefficient alpha. Average reported alpha coefficients ranged from .70 for values and beliefs to .82 for job satisfaction. With few exceptions, there were no substantive relationships between the magnitude of coefficient alpha and the research design characteristics investigated.

There is virtual consensus among researchers that, for a scale to be valid and possess practical utility, it must be reliable. Conceptually, reliability is defined as "the degree to which measures are free from error and therefore yield consistent results" (Peter 1979, p. 6). As such, the reliability of a scale places a limit on its construct validity.

However, despite its importance, there is surprisingly little guidance in the literature as to what constitutes "acceptable" or "sufficient" reliability for research purposes. Table 1 contains illustrative recommendations regarding minimally acceptable reliability. Although the recommendations differ somewhat, they share two commonalties. First, they indicate that the required degree of reliability is a function of the research purpose, whether the research is exploratory, applied, or so forth. For example, a scale in the preliminary stages of development is generally not thought to require the reliability of one used to discriminate between groups or of one being used to make decisions about individuals. Second, none of the recommendations have an empirical basis, a theoretical justification, or an analytical rationale. Rather, they appear to reflect either "experience" or intuition.

Of the recommendations contained in Table 1, those of Nunnally (1967, 1978) are the most widely referenced, either in support or criticism of an obtained reliability coefficient. For example, both Churchill (1979) and Peter (1979), in widely cited articles, endorsed Nunnally's recommendations. More generally, in the last dozen years Nunnally has been cited in support of obtained reliability coefficients more than 50 times in the *Journal of Marketing Research* and three dozen times in the *Journal of Consumer Research.* It is interesting, though, that Nunnally changed his reliability recommendations from his 1967 edition of *Psychometric Theory* in his 1978 edition. In 1967, he recommended that the minimally acceptable reliability for preliminary research should be in the range of .5 to .6, whereas in 1978 he increased the recommended level to .7 (without explanation).

Purpose

The purpose of conducting the present research was twofold. First, it was to empirically ascertain and document the magnitudes of reliability coefficients actually obtained in empirical studies and compare these coefficients with the recommendations set forth in Table 1. A second purpose was to determine whether relationships respectively exist between the magnitude of a reliability coefficient and selected individual difference constructs and research design characteristics. If reliability coefficients systematically vary across well-defined individual difference constructs or research design characteristics, it might be possible to derive "comparison standards" analogous to the base rates developed by Peterson, Albaum, and Beltramini (1984) in their investigation of effect sizes in consumer behavior experiments. Such comparison standards would complement existing guidelines like those in Table 1 and should facilitate the interpretation of reliability coefficients obtained in empirical studies.

In addition to being motivated by a lack of consistent, generalizable information regarding the magnitude of reliability coefficients typically obtained in behavioral research, the present article was motivated in part by

*Robert A. Peterson holds the John T. Stuart III Centennial Chair in Business Administration and is the Charles E. Hurwitz Fellow at the IC^2 Institute, University of Texas at Austin, Austin, TX 78712. This research was supported in part by the IC^2 Institute. The opinions expressed in the article are those of the author and do not necessarily reflect the views of the institute.

© 1994 by JOURNAL OF CONSUMER RESEARCH, Inc. • Vol. 21 • September 1994
All rights reserved. 0093-5301/95/2102-0012$2.00

Copyright © 1994. All rights reserved.

TABLE 1

SELECTED RECOMMENDED RELIABILITY LEVELS

Author	Situation	Recommended level
Davis (1964, p. 24)	Prediction for individual	Above .75
	Prediction for group of 25–50	.5
	Prediction for group over 50	Below .5
Kaplan and Saccuzzo (1982, p. 106)	Basic research	.7–.8
	Applied research	.95
Murphy and Davidshofer (1988, p. 89)	Unacceptable level	Below .6
	Low level	.7
	Moderate to high level	.8–.9
	High level	.9
Nunnally (1967, p. 226)	Preliminary research	.5–.6
	Basic research	.8
	Applied research	.9–.95
Nunnally (1978, pp. 245–246)	Preliminary research	.7
	Basic research	.8
	Applied research	.9–.95

the research of Churchill and Peter (1984) and bears many similarities to their work (see also Peter and Churchill 1986). Even so, this research differs from that of Churchill and Peter in several regards. First, there is a subtle difference in the objectives, and consequently the scope, of the two endeavors. Although both studies can be described as meta-analyses, Churchill and Peter focused on investigating "the effects of research design on reliability estimates" (p. 360), whereas the current focus was somewhat broader in that it attempted to provide base rate comparison standards at a construct level.

The difference in objectives resulted in three notable study differences. First, Churchill and Peter included six different reliability coefficients in their analyses. Because of the admonitions of Guilford (1965, chap. 17) that different reliability coefficients are not conceptually or numerically comparable, the present investigation focused on only one reliability coefficient, Cronbach's (1951) coefficient alpha. Second, the present study was significantly larger in scope in terms of the journals reviewed, the years covered, and the number of reliability coefficients analyzed. For example, Churchill and Peter limited their investigation to marketing-related journals over 12 years, whereas the present research included both marketing and psychology journals over 33 years. This resulted in nearly 30 times the number of reliability coefficients analyzed by Churchill and Peter. Third, the two studies differed in their sampling approaches. Churchill and Peter followed what might be termed the Mansfield and Busse sampling approach, whereas the present research followed a traditional Glassian sampling approach (Bangert-Drowns 1986). Churchill and Peter selectively sampled at most two reliability coefficients from each study reviewed, the largest and the smallest reported reliability coefficients, whereas the present study used *all* alpha coefficients found in a reviewed study.

Coefficient Alpha

Oversimplifying somewhat, there are two general categories of reliability coefficients, those based on longitudinal data (e.g., the test-retest reliability coefficient) and those based on cross-sectional data (e.g., internal consistency reliability coefficients and equivalence reliability coefficients). By far the most commonly used reliability coefficient is coefficient alpha, an estimator of internal consistency.

Coefficient alpha was developed by Cronbach (1951) as a generalized measure of the internal consistency of a multi-item scale. It is formulated as

$$\alpha = \left(\frac{k}{k-1}\right)(1 - \sum_{i=1}^{k} \sigma_i^2/\sigma_s^2)$$

or

$$\frac{k\bar{r}}{1 + \bar{r}(k-1)},$$

where k is the number of items in the scale, σ_i^2 is the variance of item i, σ_s^2 is the variance of the scale, and $\bar{r}$ is the average interitem correlation.

Whether by acclamation (see, e.g., Churchill 1979; Gerbing and Anderson 1988; Peter 1979) or citation, coefficient alpha has effectively become the measure of choice for estimating the reliability of a multi-item scale. Indeed, coefficient alpha has become one of the foundations of measurement theory. Because it is a generalized intraclass correlation coefficient, coefficient alpha can be derived from the theory of true and error scores, as well as from the domain sampling model. According to the *Social Science Citation Index*, Cronbach's 1951 article has been referenced in more than 2,200 articles in the last 20 years. Not only is coefficient alpha the most widely used estimator of reliability, but also it has been the subject of considerable methodological and analytical attention (see, e.g., Cortina 1993). Therefore, focusing on coefficient alpha should not detract from the generality of the research. Instead, it should improve the usefulness of the research, because there is no heterogeneity in the data due to the presence of other reliability coefficients.

METHODOLOGY

To obtain a large number and wide representation of alpha coefficients, an extensive literature review was undertaken. A census of eight psychology- and marketing-related journals was conducted, beginning with

Copyright © 1994. All rights reserved.

the year 1960 (the year in which citations of Cronbach's work began to appear) and ending with 1992. This meant that every article published in these journals was systematically and individually examined to locate alpha coefficients. In addition, a convenience sample of selected issues of 16 other journals (e.g., *Journal of Advertising Research, Journal of Business Research, Educational and Psychological Measurement, Journal of Educational Psychology,* and others) and two conference proceedings (American Marketing Association, Association for Consumer Research) were examined for alpha coefficients. Finally, a sample of unpublished manuscripts (manuscripts that had been rejected at least once for publication) were examined for alpha coefficients. The "harvesting" of alpha coefficients and the coding of individual difference constructs and research design characteristics were done by the author and two experienced research assistants, with the bulk of the coding being done by the latter two individuals after an extensive pretest and a training session. A comparison of the research assistants' coding consistency for a sample of research design characteristics produced an interrater agreement coefficient of .92 (Perreault and Leigh 1989). Hence, coding was judged to be consistently done (as would be expected given the relatively straightforward nature of the task).

Table 2 contains the sources of the alpha coefficients analyzed in the study, as well as the years they were reviewed, the type of search undertaken, the number of coefficients obtained from each, and the mean and median alpha coefficient observed for each source. Across the sources, 4,286 alpha coefficients were harvested. All alpha coefficients were independent in the sense that they were estimated from distinct sets of items. For example, if alpha coefficients were reported for both an "overall" or composite scale and one or more subscales of that scale, only the subscale alpha coefficients were retained for analysis. Eliminating the composite scale from the meta-analysis minimized the possibility of interdependencies among the alpha coefficients.

Several aspects of Table 2 merit brief mention. First, as is apparent from the table, it was not possible to conduct a census of all journals back to 1960 because some, such as the *Journal of Consumer Research,* did not begin publication until after that date. Second, as previously noted, journals were purposely selected to provide a representative array of journals, research domains, and alpha coefficients. Finally, unpublished manuscripts consisted of manuscripts submitted to the journals and conference proceedings listed in Table 2. Manuscripts classified as unpublished can only be so defined for the submissions evaluated. Some of the manuscripts classified as unpublished may well have been previously rejected or subsequently published in an outlet unknown to the present researcher. Therefore, any interpretation of the alpha coefficients obtained from this category must take this caveat into consideration.

More than 33,000 articles, proceedings papers, and rejected manuscripts were individually examined during the course of data collection. From these, alpha coefficients were obtained from 832 different articles, proceedings papers, and manuscripts reporting data from 1,030 samples consisting of more than 300,000 individuals. Thus, on average there were 5.1 alpha coefficients per reviewed publication or manuscript and 4.1 alpha coefficients per sample. To obtain a conceptually coherent pool of alpha coefficients, only those alpha coefficients reported for rating scales designed to measure individual difference constructs such as personality, attitude, and opinion in nonspecial populations were included in the analysis. Excluded were alpha coefficients reported for forced-choice scales (e.g., constant-sum scales), scales used to measure interrater agreement, and scales developed or designed for special populations (e.g., institutionalized individuals).

Constructs Measured

All harvested alpha coefficients were categorized according to the underlying construct being measured. In most instances this was accomplished by simply accepting the construct designation employed in a study. However, in certain instances it was necessary to infer the appropriate construct category on the basis of nomenclature and terminology used.

After an extensive review of the behavioral literature and a preliminary analysis, 42 categories of constructs were initially constructed for classification purposes. Because of overlap among the categories and the small numbers of alpha coefficients in some of the categories, the number of categories was ultimately collapsed to 20, which included a miscellaneous category. As might be expected, the categories varied in terms of their level of generality and the number of scales they contained. The specific construct categories used are reported in Table 3. The largest category consisted of attitude constructs, with 699 alpha coefficients. The smallest category contained constructs relating to expectation, with 37 alpha coefficients.

Research Design Characteristics

For each alpha coefficient harvested, information was sought on 12 research design characteristics in addition to source and year of publication. The research design characteristics examined in this article have been posited for more than 65 years as influencing the size of a reliability coefficient (see e.g., Symonds 1928). Each is briefly described below.

Although measurement theory does not consider the effect of *sample size* on the magnitude of an alpha coefficient, Churchill and Peter (1984) observed a negative relationship between the two. Because they were not

Copyright © 1994. All rights reserved.

TABLE 2

SOURCES OF ALPHA COEFFICIENTS

Source	Time period covered	Data collection	Number of α's	Mean α	Median α
AMA/ACR *Proceedings*	1971–1992	Sample	113	.76	.77
Journal of Applied Psychology	1960–1992	Census	670	.79	.81
Journal of Consumer Research	1974–1992	Census	166	.80	.81
Journal of Marketing	1960–1992	Census	238	.76	.78
Journal of Marketing Research	1964–1992	Census	639	.76	.79
Journal of Personality and Social Psychology	1960–1992	Census	724	.76	.79
Journal of Personality Assessment	1960–1992	Census	586	.77	.80
Journal of the Academy of Marketing Science	1972–1992	Census	387	.75	.76
Psychological Reports	1960–1992	Census	418	.76	.79
Other journals[a]	1970–1992	Sample	30	.79	.82
Unpublished manuscripts[b]	1980–1992	Sample	315	.76	.77

[a] See text for illustrative journals.
[b] See text for description of unpublished manuscripts.

able to explain their finding, the present research restudied the relationship.

Type of sample was operationalized as college student, consumer, businessperson, "mixed" (more than one type), or "cannot tell" (sample unclassifiable). Churchill and Peter (1984) hypothesized that college student samples should evince higher scale reliabilities than should noncollege student samples because students should be more experienced in completing questionnaires and perhaps more educated. However, their hypothesis was not supported. Given their finding and a lack of conceptual support for the hypothesis of different alpha coefficients for different types of samples, type of sample was not expected to influence the size of an alpha coefficient.

Two research design characteristics have been studied extensively for their effects on the magnitudes of reliability coefficients—the *number of categories,* points, or intervals in a scale item, and the *number of items* in a scale. The effect of the number of categories on the size of a reliability coefficient has long been debated in the literature, with, for example, Bendig (1954) and Jacoby and Matell (1971) concluding that the magnitude of a reliability coefficient is independent of the number of scale categories and Komorita and Graham (1965) and Lissitz and Green (1975) concluding the opposite. With the exception of the Churchill and Peter (1984) research, which found a positive relationship between the number of scale categories and the size of a reliability coefficient, prior research on the number-of-categories issue either relied on relatively small samples or used a simulation approach. It was anticipated that the present research would resolve these conflicting findings through its synthesis of a large body of alpha coefficients.

The formula for coefficient alpha implies that the larger the number of items in a scale, the greater its reliability. This relationship is generally taken for granted in the literature, and Churchill and Peter's findings corroborated it. Even so, careful inspection of the literature on this subject reveals that the reliability of a scale is expected to increase with an increase in the number of items *only under certain conditions* that relate to the homogeneity of individual item variances. Contrary to popular belief, it is not clear that simply increasing the number of items in a scale will guarantee that its reliability will also increase. Consequently, the present research again addressed the issue from the perspective of a large body of alpha coefficients.

Analogous to the Churchill and Peter research, the present study investigated the effect of *scale type, format,* and *nature* on the magnitude of an alpha coefficient. Type of scale was operationalized by whether the scale consisted of traditional Likert items (i.e., declarative statements with a five-category "agree-disagree" response format) or semantic differential items (i.e., seven-category bipolar items). Scale format was operationalized as the specific labeling of scale-item categories (i.e., whether only endpoints were labeled, whether numerical or verbal labels were used on inner categories, or whether it was impossible to label categories). The nature of the scale was operationalized by whether there was an odd or even number of scale-item categories or whether it was impossible to tell how many categories there were. Given the findings obtained by Churchill and Peter, no differences in the magnitudes of alpha coefficients were expected as a function of scale format or scale nature. However, it was anticipated that scales consisting of semantic differential items would exhibit larger alpha coefficients than would scales consisting of Likert items if a relationship exists between the number of categories in a scale item and the magnitude of an alpha coefficient (simply because of the difference in the number of item categories).

The final characteristic studied that Churchill and Peter also investigated was *mode of scale administration.* Although Churchill and Peter did not offer a hypothesis regarding mode of scale administration and the magnitude of a reliability coefficient, in the present

Copyright © 1994. All rights reserved.

TABLE 3

ALPHA COEFFICIENTS EXHIBITED BY SELECTED INDIVIDUAL DIFFERENCE CONSTRUCTS

Construct	N	Mean α	Median α	95% Confidence interval for α_μ	Quartile First	Quartile Third
Attitude	699	.76	.79	±.010	.69	.86
Conflict/stress	378	.78	.81	±.012	.73	.87
Cognition/knowledge	74	.81	.84	±.029	.75	.91
Emotion (affect, mood, etc.)	234	.80	.84	±.016	.75	.89
Expectation	37	.73	.81	±.058	.58	.87
Intention	46	.81	.84	±.043	.73	.93
Involvement/commitment	94	.79	.80	±.025	.72	.87
Lifestyle/interest	65	.74	.77	±.031	.65	.84
Motivation	86	.76	.78	±.029	.68	.87
Perceived risk	50	.75	.75	±.024	.70	.83
Perception	601	.77	.79	±.010	.70	.86
Performance (job-related)	89	.81	.83	±.028	.74	.90
Personality	544	.75	.79	±.012	.69	.85
Preference	57	.80	.81	±.024	.76	.86
Reported behavior	235	.71	.72	±.017	.63	.82
Satisfaction (job)	174	.82	.83	±.013	.77	.88
Satisfaction (other)	135	.79	.83	±.018	.75	.89
Self-confidence/self-esteem	102	.76	.79	±.020	.71	.82
Value/belief	297	.70	.73	±.017	.63	.86
Miscellaneous[a]	289	.76	.78	±.016	.69	.86

[a] Includes such constructs as loyalty, innovativeness, and importance, each with fewer than 30 alpha coefficients.

study it was expected that self-administered scales would exhibit larger alpha coefficients than would scales that were not self-administered because of the likelihood of there being less ambiguity and confusion associated with scale items that an individual could physically view and complete at his or her own pace. In their analysis, Churchill and Peter did not find a relationship between the administration mode and the magnitude of the reliability coefficients they analyzed.

Four research design characteristics not directly investigated by Churchill and Peter were also studied. One such characteristic was *scale orientation*, which indicated whether a scale was stimulus- or respondent-centered or both. Because respondent-centered scales have been the focus of more measurement attention than stimulus-centered scales (Cox 1980), it was anticipated that the former would exhibit larger alpha coefficients than the latter. A second research design characteristic studied was the *nature of the construct* represented by a scale, whether it was designated primarily as a dependent or an independent variable or as both or whether it was impossible to designate it at all. It was postulated that, because of the likelihood that a greater emphasis would be placed on a dependent variable during the conceptual and operational development of a study, larger alpha coefficients would be exhibited by dependent variables than by independent variables.

The third research design characteristic not directly investigated by Churchill and Peter was the *type of research*, whether a scale was developed specifically for the reported research, whether it was simply being applied, or whether this could not be determined. (Although Churchill and Peter did study a design characteristic they termed "source of scale," it is not directly comparable to the design characteristic investigated here.) If the scale was developed, a fourth research design characteristic was investigated. Specifically, an analysis of developed scales was conducted to determine whether there were scale items deleted during the development process and, if so, whether the *number of items deleted from a scale* influenced the size of its alpha coefficient.

It is unfortunate that, because of reporting practices, it was not always possible to capture the desired research design characteristics of every study reviewed. Hence, not all alpha coefficients harvested were included in every analysis.

RESULTS

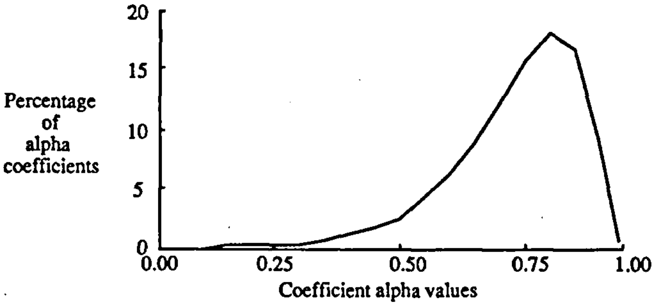
Figure 1 illustrates the distribution of the 4,286 alpha coefficients harvested. The coefficients ranged from .06 to .99 with a mean of .77 and a median of .79. As can be observed from the figure, the coefficients were relatively tightly grouped (SE = .002), and there was a slight negative skew (sk = −1.15) to the distribution.

Seventy-five percent of the observed alpha coefficients were .70 or greater, 49 percent were .80 or greater, and 14 percent were .90 or greater. These three values correspond to Nunnally's 1978 recommendations (pp. 245–246) for minimally acceptable reliability levels for preliminary, basic, and applied research, respectively.

Copyright © 1994. All rights reserved.

FIGURE 1

PERCENTAGE DISTRIBUTION OF 4,286 ALPHA COEFFICIENTS



In general, the majority of reported alpha coefficients surpass the minimal standards recommended in Table 1.

As mentioned previously, Table 2 contains the number and mean value of alpha coefficients found in each source examined. The table reveals that the mean alpha coefficient observed in the psychology-related journals ($\bar{\alpha}$ = .77) was not significantly different from that observed in the marketing-related journals and proceedings examined ($\bar{\alpha}$ = .76). Further, because the mean alpha coefficients for the published studies ($\bar{\alpha}$ = .77) and the unpublished studies ($\bar{\alpha}$ = .76) were not significantly different, all alpha coefficients were included in the remaining analyses.

It is instructive to note that the mean observed alpha coefficient, .77, is very similar to the mean observed by Churchill and Peter, .75, in their review of reliability coefficients reported in a sample of marketing-related publications. Alpha coefficients reported prior to 1976 averaged .71; those reported after 1976 averaged .77. Although this difference is statistically significant ($p < .001$), the reason for it is not clear. The difference may be due to a variety of factors, including methodological improvements and/or reporting practices, or it simply may reflect changing standards as reflected by Nunnally's (1967, 1978) recommendations.

For each of the constructs investigated, the number of alpha coefficients observed, the mean and median alpha coefficients, the approximate 95 percent confidence interval for the mean (Feldt, Woodruff, and Salih 1987), and the first and third alpha coefficient quartiles are reported in Table 3. No systematic relationship between the type of construct measured and the magnitude of coefficient alpha was observed. Mean alpha coefficients ranged from .70 for values and beliefs to .82 for job satisfaction. Differences between alpha coefficients in excess of |.04| for pairs of constructs are generally statistically significant at .05. However, because of large differences in sample sizes, caution must be exercised when interpreting mean alpha coefficient differences.

Table 4 reports the number of and the mean and median values of alpha coefficients obtained for different levels of the research design characteristics investigated. In addition, approximate 95 percent confidence intervals are presented for each of the means, and the first and third quartiles of coefficient alpha are reported. The table reveals that, with the exception of the type of research (scale development or scale application) and scale format, there were statistically significant differences ($p < .001$) between at least two levels of each of the remaining research design characteristics. In virtually every instance, though, significant differences were due to the large sample sizes employed rather than to large differences among or between alpha coefficients. There are few substantively significant or practically meaningful differences. For only three research design characteristics—the number of scale-item categories, the number of scale items, and the self- versus interviewer-administration mode—were differences of |.05| or greater observed in mean alpha coefficients.

Although statistically significant, the relationship between coefficient alpha and the number of scale-item categories was not especially strong. A regression analysis produced an r^2 of only .01, which indicated that only 1 percent of the variance in coefficient alpha was explained or accounted for by the number of categories in a scale item (the Churchill and Peter analysis produced an r^2 of .05). The major difference in mean alpha coefficients was between a scale item with two categories ($\bar{\alpha}$ = .70) and scale items with more than two categories ($\bar{\alpha}$ = .77).

Analogous to the relationship between the number of item categories and coefficient alpha, the relationship between the number of items and coefficient alpha was not especially strong. A regression analysis of the relationship resulted in an r^2 of .10, which was the same as that obtained by Churchill and Peter. The major difference in mean alpha coefficients was between scales with two or three items ($\bar{\alpha}$ = .73) and those with more than three items ($\bar{\alpha}$ = .78). Scales with 11 or more items exhibited the largest alpha coefficients, .81 on average.

A 2 × 2 ANOVA conducted on the number of item categories (two categories, three or more categories) and number of items (two or three items, four or more items) revealed no statistically significant interaction effect. The simple effects (shown in Fig. 2 as means), though, suggest that researchers should consider avoiding scales with less than four items when there are only two categories per item. (The limited number of alpha coefficients in the two-category and two or three item cell precludes extensive analyses and suggests that caution be used when making inferences.)

As expected, scales that were self-administered exhibited larger alpha coefficients ($\bar{\alpha}$ = .77) than did those that were administered by an interviewer ($\bar{\alpha}$ = .72). This finding, though, must be tempered by the disparate sample sizes and relatively small number of alpha coefficients for the interviewer mode of administration. Contrary to expectations, however, respondent-centered scales exhibited slightly smaller alpha coefficients ($\bar{\alpha}$ = .76) than did stimulus-centered scales ($\bar{\alpha}$ = .79), de-

Copyright © 1994. All rights reserved.

TABLE 4

RELATIONSHIP BETWEEN COEFFICIENT ALPHA AND SELECTED RESEARCH DESIGN CHARACTERISTICS

					Quartile	
Construct	N	Mean α	Median α	95% Confidence interval for α_μ	First	Third
Sample size:[a]						
<100	1,028	.76	.80	±.009	.68	.87
100–199	1,169	.78	.80	±.007	.71	.87
200–299	696	.78	.80	±.008	.72	.86
300 or more	1,265	.75	.77	±.007	.68	.84
Not given	128	.76	.80	±.003	.67	.86
Type of sample:[a]						
College students	1,741	.77	.80	±.007	.70	.87
Consumers	879	.74	.76	±.009	.66	.84
Businesspersons	1,130	.77	.80	±.007	.70	.86
Mixed	450	.78	.79	±.010	.72	.86
Cannot tell	86	.76	.77	±.025	.69	.87
Number of scale categories:[a]						
Not given	667	.76	.78	±.009	.68	.85
2	221	.70	.74	±.022	.63	.82
3	158	.78	.80	±.020	.69	.88
4	305	.76	.78	±.014	.69	.86
5	1,319	.77	.79	±.007	.71	.86
6	249	.75	.77	±.016	.68	.84
7	991	.78	.82	±.009	.72	.88
8 or more	376	.77	.81	±.015	.70	.88
Number of items:[a]						
Not given	342	.77	.79	±.015	.71	.86
2	307	.73	.75	±.016	.63	.83
3	519	.73	.75	±.012	.65	.84
4	503	.76	.78	±.011	.68	.85
5	441	.78	.79	±.011	.71	.86
6	377	.75	.77	±.013	.69	.84
7	210	.76	.78	±.016	.69	.85
8	183	.73	.77	±.025	.65	.85
9	117	.80	.83	±.017	.74	.89
10	311	.74	.78	±.015	.65	.85
11 or more	976	.81	.83	±.007	.76	.89
Scale type:[a]						
Likert	828	.76	.79	±.009	.70	.86
Semantic differential	372	.80	.82	±.013	.74	.89
Scale format:						
Only endpoints labeled	811	.77	.80	±.010	.69	.87
Numerical values on inner categories	553	.77	.80	±.011	.69	.86
Verbal values on inner categories	1,869	.77	.80	±.006	.71	.86
Cannot tell	1,053	.76	.78	±.008	.68	.85
Nature of scale:[a]						
Odd number of item categories	2,756	.78	.80	±.005	.70	.86
Even number of item categories	863	.74	.76	±.010	.70	.86
Cannot tell	667	.76	.78	±.009	.68	.85
Administration mode:[a]						
Self	4,064	.77	.79	±.004	.70	.86
Interviewer	153	.72	.75	±.028	.65	.85
Not given	69	.77	.78	±.029	.70	.87
Scale orientation:[a]						
Respondent-centered	3,017	.76	.79	±.002	.69	.86
Stimulus-centered	1,206	.79	.81	±.004	.73	.88
Both	63	.79	.78	±.018	.72	.87
Nature of construct:[a]						
Dependent	919	.79	.82	±.008	.73	.89
Independent	1,897	.77	.79	±.006	.70	.86
Cannot tell/both	1,470	.75	.78	±.007	.67	.85
Type of research:						
Scale development	1,270	.77	.79	±.007	.70	.86
Scale application	2,978	.77	.79	±.005	.70	.86
Cannot tell	38	.84	.84	±.029	.78	.91

[a] Relationship significant at $p < .001$.

Copyright © 1994. All rights reserved.

FIGURE 2

RELATIONSHIP BETWEEN COEFFICIENT ALPHA AND NUMBER OF SCALE ITEMS AND ITEM CATEGORIES

		Number of scale items	
		2 or 3	4 or more
Number of categories in item	2	$\bar{\alpha}$ = .62 (n = 23)	$\bar{\alpha}$ = .71 (n = 186)
	3+	$\bar{\alpha}$ = .74 (n = 710)	$\bar{\alpha}$ = .78 (n = 2,536)

spite the fact that respondent-centered scales averaged nearly two items more per scale than stimulus-centered scales.

Table 4 reveals that the average alpha coefficient did not vary as a function of whether a scale was being developed or applied (i.e., the scale had been developed previously). Table 5 indicates, though, that when a scale was developed, its alpha coefficient was significantly related to the number of items that were eliminated during the developmental process (r^2 = .18). Four hundred ninety-three scales that were reported in the reviewed literature as being developed had information regarding the number of items eliminated during the developmental process. It is apparent that eliminating items significantly increased the average alpha coefficient of a scale. (Scales with more than 30 items eliminated during the developmental process tended to be very long scales that were essentially used as item pools to develop new "short form" scale versions.) Eliminating even one item from a scale during the developmental process (usually on the basis of a statistical criterion as opposed to a theoretical one) significantly increased coefficient alpha from .70 to a minimum of .77. In general, eliminating items during scale development increased coefficient alpha to an average of .81.

DISCUSSION AND CONCLUSIONS

The results of this study provide at least tentative answers to a variety of frequently asked questions regarding coefficient alpha. First, the results provide empirical evidence as to what is a "typical" alpha coefficient or, more precisely, what constitutes "high" or "low" alpha coefficients relative to previously obtained coefficients, both in general and for specific constructs and selected research design characteristics. Until this study (with the possible exception of the Churchill and Peter [1984] work), no empirical standards existed against which obtained alpha coefficients could be systematically compared. Researchers attempting to interpret an obtained alpha coefficient previously only had recommendations such as those offered in Table 1 or were forced to rely on experience or intuition. The present research provides empirical standards that permit direct comparisons. By comparing an observed alpha coefficient with coefficients reported in Tables 3 and 4 that were obtained under similar circumstances, "actuarial-type" insights are available regarding the magnitude of the observed coefficient.

TABLE 5

EFFECT OF ELIMINATING SCALE ITEMS DURING SCALE DEVELOPMENT ON COEFFICIENT ALPHA

Number of items eliminated[a]	*N*	Mean α
None	234	.70
1–3	64	.79
4–10	69	.80
11–30	63	.77
More than 30	63	.87

[a] Relationship significant at $p < .001$.

Across the 4,286 alpha coefficients, 1,030 samples, and 832 studies investigated, the mean coefficient alpha was .77. Seventy-five percent of the observed alpha coefficients were .70 or greater. These values compare favorably with the recommendations set forth in Table 1 for preliminary or basic research. This agreement between the recommendations and reported alpha coefficients is neither surprising nor likely to be coincidental. Because the recommendations have effectively become sacrosanct, it can be persuasively argued that reported alpha coefficients (especially those that are associated with developed scales) are, on average, in large measure a function of the recommendations.

Only 14 percent of the observed alpha coefficients reached or exceeded .90, the threshold generally recommended for applied research by authorities such as Nunnally. However, the implications of this latter finding are not altogether clear. It may be that the recommended threshold levels for applied research are unrealistically high for consumer behavior and marketing research. Or, since most behavioral researchers characterize their research as basic (especially those comparing their alpha coefficients with the recommended standards), the absence of coefficients reaching the .90 standard may be of little consequence. Indeed, the relative absence of alpha coefficients at or above .90 may actually reflect good research practice. Boyle (1991), for instance, has argued that scales exhibiting very high alpha coefficients (e.g., above .90) should be avoided, because they simply imply a high level of item redundancy, not scale reliability.

An investigation of alpha coefficients .90 or greater revealed that, relative to alpha coefficients less than .90,

Copyright © 1994. All rights reserved.

TABLE 6

SUMMARY COMPARISON OF CHURCHILL AND PETER (1984) AND PRESENT RESEARCH FINDINGS

	Relationship with coefficient alpha	
Research design characteristic	Churchill and Peter	Present research
Sample size	Sample size negatively related to alpha	No substantive relationship
Type of sample	No relationship	No substantive relationship
Number of scale-item categories	Number of item categories positively related to magnitude of alpha	Scale items with two categories exhibited smaller alphas than those with more than two categories
Number of items in scale	Positive relationship between number of items and size of alpha	Scales with two or three items exhibited smaller alphas than those with more than three items
Scale type	No relationship	No substantive relationship
Scale format	No relationship	No relationship
Scale nature	No relationship	No substantive relationship
Administration mode	No relationship	Interviewer administration produced lower alpha than did self-administration
Scale orientation	Not studied	No substantive relationship
Nature of construct	Not studied	No substantive relationship
Type of research	Not directly studied	No relationship
Number of items deleted during scale development	Not studied	Positive relationship between the number of items deleted and the magnitude of alpha

their originating scales were more likely to consist of *more* items with *more* categories, to be derived from smaller samples, to be stimulus-centered, and to have been developed to measure constructs employed as dependent variables. Succinctly stated, there appears to be a systematic relationship between the origin of an alpha coefficient and whether it is "sufficiently high" (in excess of .90) to warrant being deemed acceptable for applied research.

Second, the present research generally corroborated the findings of Churchill and Peter (1984; Peter and Churchill 1986) to the extent that it documented that coefficient alpha is relatively robust and is not subject to dramatic fluctuations as a consequence of research design characteristics. As summarized in Table 6, despite their underlying differences, the Churchill and Peter research and the present research obtained similar results for six of the eight design characteristics investigated in common. The two studies differed in their findings regarding the impact of sample size and the mode of administration on the magnitude of coefficient alpha, with Churchill and Peter finding a relationship for the former but not for the latter. Because Churchill and Peter had no a priori hypothesis regarding the relationship between sample size and coefficient alpha and were unable to explain the obtained relationship post hoc, it may have simply been an anomaly, as the present research suggests. The lack of a significant relationship between the mode of administration and the magnitude of coefficient alpha in the Churchill and Peter study may have been a consequence of their sample size and/or operationalization of the design characteristic. The finding in the present research that self-administered scales exhibited larger alpha coefficients than interviewer-administered scales is intuitively logical. Hence, the mode of administration warrants attention in future research as an influencer of coefficient alpha.

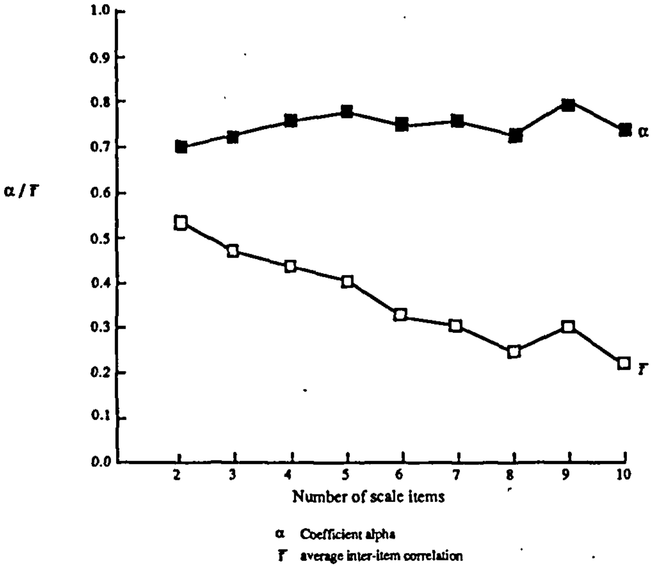
Furthermore, the present research confirms previous findings from simulation studies (see, e.g., Jenkins and Taber 1977; Lissitz and Green 1975) and empirical demonstrations (see, e.g., Bendig 1954; Jacoby and Matell 1971) that, with the exception of scale items possessing only two response categories, the number of categories in a scale item is essentially unrelated to the magnitude of coefficient alpha. Thus, the research adds to the body of knowledge relating to the "optimal" number of scale categories (see, e.g., Cox 1980) in that it shows that to increase reliability (as estimated by coefficient alpha) it is necessary to assume a strategy other than simply increasing the number of scale-item categories.

Perhaps the most interesting relationship studied was that between the number of items in a scale and the magnitude of coefficient alpha. Theoretically, the larger the number of items in a scale, the more reliable will be the scale (see, e.g., Nunnally 1978, p. 243). However,

Copyright © 1994. All rights reserved.

FIGURE 3

RELATIONSHIPS BETWEEN NUMBER OF SCALE ITEMS AND COEFFICIENT ALPHA AND AVERAGE INTERITEM CORRELATION



as Table 3 indicates, the observed relationship deviated considerably from theory; only 10 percent of the variance in coefficient alpha could be attributed to the number of scale items. On average, coefficient alpha does not appear to systematically increase once there are more than three items in a scale. This could be due to the heterogeneity in coefficient alpha values within a particular number of scale items because of differences in constructs being measured, "noise" due to scale type and format differences, sampling errors, and so forth. To a large extent, though, the observed relationship probably reflects a decrease in the average interitem correlation as the number of items in a scale is increased.

As its formula indicates, ceteris paribus, coefficient alpha varies as a joint function of the number of items and the average interitem correlation. In particular, according to the formula, coefficient alpha should increase as the number of items and the average interitem correlation increase. It is interesting, though, that in the present study coefficient alpha and the average interitem correlation were inversely related. As shown in Figure 3, the average interitem correlation declined monotonically as the number of scale items increased, whereas coefficient alpha increased slightly as the number of scale items increased.

Consider the mean alpha coefficient observed for three-item scales, .73. The average interitem correlation corresponding to this alpha is .47. If this average interitem correlation were to be applied to a nine-item scale, the expected alpha coefficient would be .89. However, the observed alpha coefficient was .80. The difference between the expected and the observed alpha coefficient values is due to the size of the average interitem correlation observed for nine-item scales, .31. This result suggests that researchers attempting to increase the magnitude of an alpha coefficient should concentrate on the quality of items included in a scale and not simply on the quantity of items.

In conclusion, this article has documented the magnitudes of alpha coefficients obtained in behavioral research over the past three decades and has demonstrated that, with few exceptions, the magnitudes appear to be more of a function of the construct being measured than of the characteristics of the underlying research design. It is hoped that this article will provide useful information for those researchers constructing or evaluating multi-item scales. Moreover, by implication, it is hoped that the article will stimulate researchers to report more information on the scales used in their studies. Many of the articles and papers examined for this investigation did not contain sufficient information about a scale to permit an informed judgment as to its potential usefulness or application. For example, the format of 22 percent of the scales was not identified, and for 16 percent of the scales examined no mention was made of the number of item categories. Such statistics suggest the need for at least minimal reporting standards regarding research design characteristics when publishing scale-related research.

[*Received September 1991. Revised October 1993. Kent B. Monroe served as editor for this article.*]

REFERENCES

Bangert-Drowns, Robert L. (1986), "Review of Developments in Meta-analytic Method," *Psychological Bulletin,* 99 (May), 388-399.

Bendig, A. W. (1954), "Reliability and the Number of Rating Scale Categories," *Journal of Applied Psychology,* 38 (February), 38-40.

Boyle, Gregory J. (1991), "Does Item Homogeneity Indicate Internal Consistency or Item Redundancy in Psychometric Scales?" *Personality and Individual Differences,* 12 (March), 291-294.

Churchill, Gilbert A., Jr. (1979), "A Paradigm for Developing Better Measures of Marketing Constructs," *Journal of Marketing Research,* 16 (February), 64-73.

——— and J. Paul Peter (1984), "Research Design Effects on the Reliability of Rating Scales: A Meta-analysis," *Journal of Marketing Research,* 21 (November), 360-375.

Cortina, Jose M. (1993), "What Is Coefficient Alpha? An Examination of Theory and Applications," *Journal of Applied Psychology,* 78 (February), 98-104.

Cox, Eli P., III (1980), "The Optimal Number of Response Alternatives for a Scale: A Review," *Journal of Marketing Research,* 17 (November), 407-422.

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

Davis, Frederick B. (1964), *Educational Measurements and Their Interpretation,* Belmont, CA: Wadsworth.

Feldt, Leonard S., David J. Woodruff, and Fathi A. Salih (1987), "Statistical Inference for Coefficient Alpha," *Applied Psychological Measurement,* 11 (March), 93-103.

Copyright © 1994. All rights reserved.

Gerbing, David W. and James C. Anderson (1988), "An Updated Paradigm for Scale Development Incorporating Unidimensionality and Its Assessment," *Journal of Marketing Research*, 25 (May), 186–192.

Guilford, J. P. (1965), *Fundamental Statistics in Psychology and Education*, 4th ed., New York: McGraw-Hill.

Jacoby, Jacob and Michael S. Matell (1971), "Three-Point Likert Scales Are Good Enough," *Journal of Marketing Research*, 8 (November), 495–500.

Jenkins, G. Douglas, Jr. and Thomas D. Taber (1977), "A Monte Carlo Study of Factors Affecting Three Indices of Composite Scale Reliability," *Journal of Applied Psychology*, 62 (November), 392–398.

Kaplan, Robert W. and Dennis P. Saccuzzo (1982), *Psychological Testing: Principles, Applications, and Issues*, Monterey, CA: Brooks/Cole.

Komorita, Samuel S. and William K. Graham (1965), "Number of Scale Points and the Reliability of Scales," *Educational and Psychological Measurement*, 25 (November), 987–995.

Lissitz, Robert W. and Samuel B. Green (1975), "Effect of the Number of Scale Points on Reliability: A Monte Carlo Approach," *Journal of Applied Psychology*, 60 (February), 10–13.

Murphy, Kevin R. and Charles O. Davidshofer (1988), *Psychological Testing: Principles and Applications*, Englewood Cliffs, NJ: Prentice-Hall.

Nunnally, Jum C. (1967), *Psychometric Theory*, 1st ed., New York: McGraw-Hill.

——— (1978), *Psychometric Theory*, 2d ed., New York: McGraw-Hill.

Perreault, William D., Jr. and Lawrence E. Leigh (1989), "Reliability of Nominal Data Based on Qualitative Judgments," *Journal of Marketing Research*, 26 (May), 135–148.

Peter, J. Paul (1979), "Reliability: A Review of Psychometric Basics and Recent Marketing Practices," *Journal of Marketing Research*, 16 (February), 6–17.

——— and Gilbert A. Churchill, Jr. (1986), "Relationships among Research Design Choices and Psychometric Properties of Rating Scales: A Meta-analysis," *Journal of Marketing Research*, 23 (February), 1–10.

Peterson, Robert A., Gerald Albaum, and Richard F. Beltramini (1984), "A Meta-analysis of Effect Sizes in Consumer Behavior Experiments," *Journal of Consumer Research*, 12 (June), 97–103.

Symonds, Percival M. (1928), "Factors Influencing Test Reliability," *Journal of Educational Psychology*, 19 (February), 73–87.

Copyright © 1994. All rights reserved.