MARKETING SCIENCE
Vol. 24, No. 2, Spring 2005, pp. 294–301
ISSN 0732-2399 | EISSN 1526-548X | 05 | 2402 | 0294

informs

DOI 10.1287/mksc.1040.0097


# Results on the Standard Error of the Coefficient Alpha Index of Reliability

Adam Duhachek
Department of Marketing, Kelley School of Business, Indiana University, Bloomington, Indiana 47405, aduhache@indiana.edu

Anne T. Coughlan
Northwestern University, Evanston, Illinois 60208, a-coughlan@kellogg.northwestern.edu

Dawn Iacobucci
Department of Marketing, Wharton School, University of Pennsylvania, 3730 Walnut Street, Philadelphia, Pennsylvania 19428, iacobucci@wharton.upenn.edu

In this research, we investigate the behavior of Cronbach's coefficient alpha and its new standard error. We systematically analyze the effects of sample size, scale length, strength of item intercorrelations, and scale dimensionality. We demonstrate the beneficial effects of sample size on alpha's standard error and of scale length and the strengths of item intercorrelations (effects that are substitutes in their benefits) on both alpha and its standard error. Our findings also speak to this adage: Heterogeneity within the item covariance matrix (e.g., through multidimensionality or poor items) negatively impacts reliability by decreasing the precision of the estimation. We also examined the question of "equilibrium" scale length, showing the conditions for which it is optimal to add no items, or one, or multiple items to a scale. In terms of "best practices," we recommend that researchers report a confidence interval or standard error along with the coefficient alpha point estimate.

*Key words*: measurement; survey research; reliability; coefficient alpha
*History*: This paper was received July 9, 2003, and was with the authors 6 months for 2 revisions; processed by Roland Rust.

Measurement and scaling are integral to research in marketing (Laurent et al. 1995, Finn 1992, Kalwani and Silk 1983). In the last 10 years, measurement was among the five most popular topics in the marketing journals (e.g., along with advertising and branding, Malhotra et al. 1999). The 500 marketing scales documented in Bearden and Netemeyer (1999) and Bruner and Hensel (1994) also attest to the centrality and importance of measurement in marketing. *Marketing Science* researchers frequently report survey data (e.g., Brown 1999, Chen and Rao 2002, Dhar et al. 1999, Häubl and Trifts 2000, Hoch et al. 1999, Kahn and Luce 2003, Lynch and Ariely 2000, Novak et al. 2000, Rao and Mahi 2003, Rust et al. 1999, Soman and Cheema 2002).

Whether researchers are developing scales or using established ones, readers expect authors to report a reliability index. While other indices exist, Peterson (1994) tallies that by far the most frequently reported reliability index is Cronbach's coefficient alpha.

This paper contributes to the marketing measurement literature in two manners. First, we derive a standard error for alpha and encourage scholars to report confidence intervals about alpha, rather than merely the alpha point estimate, because confidence intervals always convey greater information. Second, we examine the analytical behavior of alpha and its standard error and confidence limits with respect to their component factors, including sample size, scale length, strength of item correlations, and scale heterogeneity.

## Cronbach's Coefficient Alpha

Cronbach's coefficient alpha is widely known and defined as follows (Cronbach 1951):[1]

$$\alpha = \frac{p}{p-1}\left[1 - \frac{\sum \sigma_i^2}{\sigma_T^2}\right], \tag{1}$$

where $p$ is the number of items in the scale ($p \geq 2$); $\sigma_i^2$ is the variance of the $i$th item, $i = 1, 2, \ldots p$; and $\sigma_T^2$ is the variance of the entire test, $\sigma_T^2 = \sum_{i=1}^{p} \sigma_i^2 + \sum_{i \neq} \sum_j \sigma_{ij}$. Early attempts to account for a statistical distribution for alpha were predicated on the restrictive (and empirically unobserved) assumption of compound symmetry (equal item variances and covariances) and were not robust (Barchard and Hakstian 1997). The simple practice was adopted that

[1] See Cooil and Rust (1994) and Rust and Cooil (1994) for analogous indices for qualitative judgments and Rossiter (2002) for a general measurement framework.

a computed alpha is compared to a conventional, but arbitrary, threshold $\hat{\alpha} \geq 0.70$ (Nunnally and Bernstein 1994, p. 265). By that standard, the coefficients reported in the *Marketing Science* articles just cited are certainly respectable (cf. $\bar{\alpha} = 0.77$).

## Alpha's Standard Error

Recent research has affirmed that Equation (1) is the form of the maximum likelihood estimator of alpha based on a standard assumption of multivariate normality (van Zyl et al. 2000). As $n \to \infty$, $\sqrt{n}(\hat{\alpha} - \alpha)$ is distributed normal with mean zero and variance:

$$Q = \left[\frac{2p^2}{(p-1)^2(j'Vj)^3}\right] \cdot \left[(j'Vj)(trV^2 + tr^2V) - 2(trV)(j'V^2j)\right], \quad (2)$$

where $n$ represents sample size, $\hat{\alpha}$ is the MLE of $\alpha$, $V$ is the population covariance matrix among the items, and $j$ is a $px1$ vector of 1s (van Zyl et al. 2000).

Armed with a variance, we derive a standard error to create 95% confidence intervals:[2]

$$\hat{\alpha} \pm (1.96)\left(\sqrt{\frac{Q}{n}}\right). \quad (3)$$

As for any index, point estimates are not satisfactory when sufficient statistical theorizing has provided means of inferential tests. In the next section, we illustrate how a single index can be misleading and why the supplemental information of a confidence interval is necessary.

## More Marketing Relevance

One reason confidence intervals are informative is that researchers with lower alphas need not always be tentative in their interpretations, depending on factors such as sample size. Similarly, a higher alpha estimate, but accompanied by a high variance, would limit researchers in their data interpretations. For example, consider an alpha that does not exceed the 0.70 threshold, say, $\alpha = 0.65$. Depending on sample size, e.g., $n = 30, 100, 200$, or 500, the standard errors would be, respectively, 0.103, 0.057, 0.040, 0.025, yielding 95% confidence intervals of [0.450, 0.856], [0.542, 0.764], [0.575, 0.732], and [0.603, 0.703]. Note that in each case, the confidence interval includes 0.70. (Of course, the lower bounds are also informative, and $n = 30$ is rather problematic.)

At the other extreme, consider a larger, apparently acceptable alpha, say $\alpha = 0.77$. For sample sizes of $n = 30$, 50, 100, 200, the standard errors are 0.070, 0.054, 0.038, 0.027, and the confidence intervals [0.629, 0.903], [0.660, 0.872], [0.691, 0.841], [0.713, 0.819]. These confidence intervals indicate that even though 0.77 seems to exceed the cutoff of 0.70, sample size needs to reach 200 before the lower bound surpasses the requisite threshold. For smaller samples (30, 50, 100), researchers should not be overly confident in reporting a "reliable" scale due to the point estimate, given that these confidence ranges indicate rather wide margins of error.

As a final illustration, we have borrowed data from a published article (Iacobucci et al. 2003) in which there exists a five-item scale measuring customer satisfaction. In one sample, $\alpha = 0.67$, and the standard error for alpha is 0.07 ($n = 50$). The confidence interval on alpha is [0.53, 0.81]. The authors correlate the customer satisfaction scale with measures of likelihood of repeat purchase ($r = 0.24$) and perceptions of value ($r = 0.16$). Neither correlation is significant, but adjusting the correlations for reliability indicates that the true relationships are in the ranges of [0.26, 0.33] and [0.18, 0.22]. The latter confidence interval falls below significance in its entirety, but the first interval contains the $p < 0.05$ cutoff of a correlation of 0.28, thus suggesting a stronger link between the constructs. Thus, the interpretation for the marketing researcher is refined. This richer information is all the more important for such "real-world" (i.e., usually messy) data, where it is not unusual for alphas to be low. Data collection efforts can be expensive, e.g., spanning multiple countries. When important practical or theoretical questions are on the line, the confidence intervals provide two extremes of a "what if" scenario, with an optimistic upside and a cautious downside.

Hopefully, we have convinced the reader of the importance of alpha, confidence intervals about alpha, and the central role of measurement for marketing researchers. We now present two analyses. In Analysis 1, we study the factors that affect alpha and the standard error of alpha. We analytically investigate these statistics, isolating sample size, number of items composing the scale, and the correlations among the items. In Analysis 2, we examine the effect of scale covariance heterogeneity, i.e., multidimensionality, on both alpha and its standard error.

## Analysis 1. Analysis of the Components: $n, p, r$

We begin by analytically investigating the influence on alpha and its standard error of the three component factors: sample size ($n$), scale length ($p$), and item intercorrelations ($r$). In the simple case where all item intercorrelations are equal to $\bar{r}$ (or simply, $r$), we can express alpha (called here $\alpha_1$) and its standard error, $SE_1$, as

$$\alpha_1 = \frac{p\bar{r}}{1 + \bar{r}(p-1)}; \quad (4)$$

$$SE_1 = \sqrt{\frac{Q_1}{n}}, \quad (5)$$

[2] Readers may obtain SAS or SPSS programming code from the authors to compute standard errors and confidence intervals to assess the reliability of their own scales.

**Table 1 Comparative-Static Effects on Alpha and Its Standard Error**

| | Sign of comparative-static effect on: | |
|---|---|---|
| Of these factors | $\alpha_1$ | $SE_1$ |
| $n$ | $\frac{\partial \alpha_1}{\partial n} = 0$ | $\frac{\partial SE_1}{\partial n} < 0$ |
| $p$ | $\frac{\partial \alpha_1}{\partial p} > 0$ | $\frac{\partial SE_1}{\partial p} < 0$ |
| $r$ | $\frac{\partial \alpha_1}{\partial r} > 0$ | $\frac{\partial SE_1}{\partial r} < 0$ |

where

$$Q_1 = \frac{2p(1-\bar{r})^2}{(p-1)[1+\bar{r}(p-1)]^2}. \quad (6)$$

We establish the following results (results follow in the appendix) and proofs are elaborated in a fuller (Technical Appendix to be posted at http://mktsci.pubs.informs.org).

Proposition 1. *The comparative-static sign effects of $n$, $p$, and $r$ on alpha and its standard error are as shown in Table* 1.

Proposition 2. *$p$ and $r$ are complements in increasing the estimate of alpha for*

$$r < \frac{1}{1+p}$$

*and substitutes in increasing alpha for*

$$r > \frac{1}{1+p}.$$

*Furthermore, $n$, $p$, and $r$ are substitutes in decreasing the standard error of alpha. Formally,*

$$\frac{\partial^2 \alpha_1}{\partial p\, \partial r} > 0 \text{ for } r < \frac{1}{1+p}, \quad \text{and} \quad < 0 \text{ for } r > \frac{1}{1+p};$$

$$\frac{\partial^2 SE_1}{\partial n\, \partial p} > 0; \quad \frac{\partial^2 SE_1}{\partial n\, \partial r} > 0; \quad \frac{\partial^2 SE_1}{\partial p\, \partial r} > 0.$$

The findings in Proposition 1 have important implications. First, a focus solely on alpha and not its standard error would lead a researcher to ignore the important effect exerted by $n$ on the *precision* of the alpha estimate. Taking the standard error into account clearly should impel a researcher to seek larger sample sizes. Further, a longer scale (increasing $p$) both increases alpha and makes the estimate more precise (holding constant $n$ and $r$).[3]

[3] Drolet and Morrison (2001) suggest practitioners might not wish to absorb the cost of longer scales for what might be only incremental information. Measurement theorists have long acknowledged that single-item scales are conservative in being less reliable than multi-item scales. Longer scales are advised because a priori one does not know whether the empirical relationship will be sufficiently strong as to be detected by the single item. Furthermore, it is important, of course, to acknowledge that reliability is only one tool in assessing the goodness of a scale (cf. Rossiter 2002); from a substantive perspective, it would not be good practice to enhance alpha by creating redundant items.

Proposition 2 carries the comparative-static analysis further to look at interaction effects between the parameters. It shows that $p$ and $r$ are substitutable ways to increase alpha for high enough values of $p$ and $r$ (in the online Technical Appendix at http://mktsci.pubs.informs.org, we provide a table of critical values), and thus the incremental positive effect of either one on alpha diminishes, the higher the other parameter. Similarly, $n$, $p$, and $r$ are all substitutable in their effects on the standard error: Each parameter's effect is negative, but *less so* the higher either of the other parameters. Survey researchers typically do not exert direct control over the size of $r$, but this analysis shows that if items are likely to exhibit low intercorrelations (e.g., per past research), researchers can get a proportionately greater benefit from enhancing scale length or sample size.

## Analysis 2. Investigating Item Covariance Heterogeneity

The recently derived standard error of alpha is the first to not require the restrictive assumption of compound symmetry, so we are interested in determining the degree to which violations of this assumption (common in real data) might affect alpha's standard error. Analysis 1 provides a comparative benchmark by assuming that all interitem correlations are constant, i.e., $\bar{r}$ (or $r$), representing item homogeneity (i.e., parallel test forms). We examine two scenarios of covariance heterogeneity. First, we consider the frequently encountered occasion that an investigator has a scale that is largely good but contains one bad item, i.e., one whose correlation is lower with the other items. A classic step in scale development is to compute item-total correlations as initial diagnostics to detect items that appear to be poor indicators of the construct. For example, for four items—the first item being the culprit—the pattern of item intercorrelations would resemble

$$\begin{bmatrix} 1 & cr & cr & cr \\ cr & 1 & r & r \\ cr & r & 1 & r \\ cr & r & r & 1 \end{bmatrix}, \quad 0 \le c < 1.$$

The degree of heterogeneity is greater the smaller $c$ is. Conversely, as $c$ approaches 1, the heterogeneity

vanishes (the item intercorrelations are homogeneous per Analysis 1).

This covariance heterogeneity condition yields formulae for $Q$ and alpha:

$$A_2 = p + 2(p-1)(cr) + (p-1)(p-2)r,$$
$$B_2 = [1 + (p-1)(cr)^2] + (p-1)[(cr)^2 + 1 + (p-2)r^2],$$
$$C_2 = 2(p-1)[2cr + (p-2)(cr)r] + (p-1)(p-2)[(cr)^2 + 2r + (p-3)r];$$

then

$$Q_2 = \left[\frac{2p^2}{(p-1)^2 A_2^3}\right][A_2(B_2 + p^2) - 2p(B_2 + C_2)], \quad (7)$$

and

$$\alpha_2 = \frac{p}{p-1}\left[1 - \frac{p}{p + 2(p-1)(cr) + (p-1)(p-2)r}\right]. \quad (8)$$

For the second scenario of item covariance heterogeneity, we posit a similarly frequently encountered empirical condition. Instead of using a single underlying factor yielding the results on the $p$ items, researchers frequently use multifaceted scales. Thus, we examine the condition that the item intercorrelations result from a multiple (two) factor structure. For example, for $p = 4$ items, there would exist two clusters of highly correlated items, with lower interfactor correlations:

$$\begin{bmatrix} 1 & r & cr & cr \\ r & 1 & cr & cr \\ cr & cr & 1 & r \\ cr & cr & r & 1 \end{bmatrix}, \quad 0 \le c < 1.$$

Heterogeneity is greater as $c$ decreases, and as $c$ approaches 1, heterogeneity vanishes. For this second scenario,

$$A_3 = p + \tfrac{1}{2}p^2(cr) + pr\left(\tfrac{1}{2}p - 1\right),$$
$$B_3 = p\left\{1 + \left[\tfrac{1}{2}p(cr)^2\right] + \left[\left(\tfrac{1}{2}p - 1\right)(r^2)\right]\right\},$$
$$C_3 = p\left(\tfrac{1}{2}p - 1\right)\left\{(2r) + \left[\left(\tfrac{1}{2}p\right)(cr)^2\right] + \left[\left(\tfrac{1}{2}p - 2\right)(r^2)\right]\right\} + \left(\tfrac{1}{2}p^2\right)\left\{(2cr) + [(p-2)r(cr)]\right\}.$$
$$Q_3 = \left[\frac{2p^2}{(p-1)^2 A_3^3}\right][A_3(B_3 + p^2) - 2p(B_3 + C_3)],$$

which simplifies to:

$$Q_3 = \frac{4p\{-2(r-1)^2 + p[2 - 2r(c+1) + r^2(c^2+1)]\}}{(p-1)^2[2 + (r(cp + p - 2))]^2}, \quad (9)$$

$$\text{and } \alpha_3 = \frac{p}{p-1}\left[1 - \frac{p}{p + \frac{1}{2}p^2(cr) + pr\left(\frac{1}{2}p - 1\right)}\right]. \quad (10)$$

This covariance heterogeneity condition considers the case of multidimensional scales. Alpha is a measure of internal consistency reliability, not an assessment of scale unidimensionality, so scales with more than one underlying factor can still yield high levels of alpha. Coefficient alpha is intended only for homogeneous items, i.e., those measuring a single construct. However, it is clear in its formulation that whether the items achieve the status of internal consistency, it is the extent of covariability that drives the size of $\alpha$, along with $p$, the scale length. Accordingly, we have Proposition 3.

Proposition 3. *Heterogeneity due to the presence of a single poor item in a scale has an adverse (negative) effect on alpha, and an adverse (positive) effect on the standard error of alpha, holding p and r constant*:

$$\frac{\partial \alpha_2}{\partial c} > 0, \quad \textit{and} \quad \frac{\partial SE_2}{\partial c} < 0.$$

Proposition 4. *Heterogeneity due to multiple underlying factors also adversely (negatively) affects alpha and adversely (positively) affects the standard error of alpha, holding p and r constant*:

$$\frac{\partial \alpha_3}{\partial c} > 0, \quad \textit{and} \quad \frac{\partial SE_3}{\partial c} < 0.$$

These results have important practical implications. Conventional advice maintains that heterogeneity among items is not desired in scale construction, and we demonstrate this effect analytically here. These results yield the first analytical evidence of the deleterious effect of covariance heterogeneity. Traditional arguments have been theoretical, but now we see that confidence intervals widen, providing analytical support for the popular notion that there is a cost affiliated with generating multidimensional scales.

## Adding an Item to a Scale

In this section, we demonstrate what happens when a researcher adds an item to a scale with the intention of enhancing alpha through increasing $p$, yet inadvertently adds a "bad" item, thus adversely affecting alpha and its standard error. Under what conditions will the added item yield a stronger alpha and/or a tighter confidence interval versus e.g., the overall weaker $r$ yielding a lower alpha? We find that the critical value of $c$ that preserves alpha when making this addition is

$$c_1^* = \frac{1 + p + r(p-1)}{2(1 + p - r)} = \frac{1}{2} + \frac{rp}{2(1 + p - r)}; \quad \frac{1}{2} < c_1^* \le 1. \quad (11)$$

Adding a "bad" item with a $c$ value greater than $c_1^*$ causes the alpha estimate to *increase*; with a $c$ value less

than $c_1^*$, alpha *decreases*. For example, when starting with a three-item scale ($p = 3$) with an interitem correlation of 0.3 ($r = 0.3$), so that the baseline alpha is 0.563, adding one more bad item will just keep alpha constant if $c = 0.622$. This $c$ implies a correlation between the bad item and the other items of 0.187 (0.3 times 0.622), reflecting the fact that adding the one bad item has both a good effect *and* a bad effect: It is bad to include an item with lower interitem correlation, but it is good to add another item to the scale (because it increases the value of alpha). To keep alpha constant, then, the value of $c \cdot r$ need not be as high as $r$ itself.

This alpha-preserving $c_1^*$ is increasing in both $p$ and $r$. This result is sensible: It is harder to improve upon a longer scale than a shorter one, holding $r$ constant. Similarly, for any given $p$, it is more difficult to improve upon a scale that starts out with a high $r$ than with a low $r$. We also show in a complete numerical analysis (in the online Technical Appendix at http://mktsci.pubs.informs.org) that adding a bad item that preserves alpha tightens the confidence interval as well.

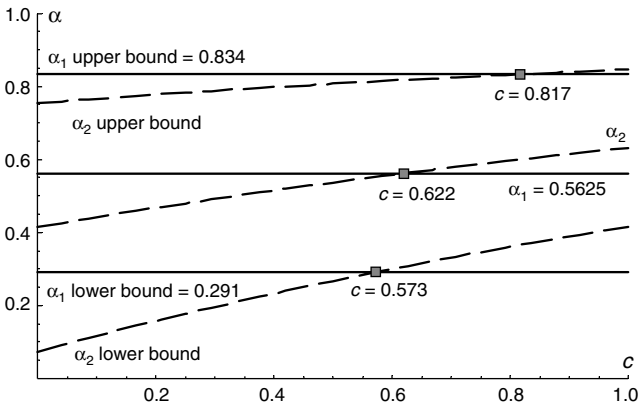
In a final illustration, we find the lower bound of alpha's 95% confidence interval can rise even if adding a bad item slightly *decreases* the alpha estimate. For example, when $p = 3$ and $r = 0.3$, the baseline-case alpha is 0.5625 and the critical value of $c$ is 0.622. Assuming $n = 30$, we find that:

- For $0 < c < 0.573$, alpha and the entire 95% confidence interval with one bad item ($p = 4$) are *lower* than with the baseline alpha ($p = 3$);
- For $0.573 < c < 0.622$, the upper bound with one bad item, and alpha itself, are still *lower*, but the lower bound is now *higher*, than in the baseline case;
- For $0.622 < c < 0.817$, the upper bound with one bad item is *lower*, but the lower bound and alpha are *higher* in value than in the baseline case;
- For $0.817 < c \leq 1$, alpha and the entire interval are *higher* in value in the case of one bad item than in the baseline case.

In short, for $0.573 < c < 0.622$, the addition of the bad item causes the alpha estimate to *fall* but causes the lower bound of the 95% confidence interval to *rise* (i.e., improve); in this region, the beneficial effect of adding another item has a stronger impact on the confidence interval than does the "badness" of the item added. Clearly, however, a "bad enough" item (i.e., $c < 0.573$) will cause alpha to decrease and the lower bound to decrease as well. We show this graphically in Figure 1.

These insights begin to indicate that more (items) is not necessarily better. Thus, in the next and final analysis, we consider how a researcher might decide *how many* items to add to the scale.

**Figure 1 Alpha and 95% Confidence Intervals: Baseline vs. One Bad Item**




*Notes.* This graph assumes $p = 3$, $r = 0.3$, $n = 30$. The baseline alpha ($\alpha_1$) is 0.5625, and the 95% confidence interval is [0.291, 0.834] ($\alpha_1$ lower bound and $\alpha_1$ upper bound). Adding one bad item results in the dotted curves labeled $\alpha_2$, $\alpha_2$ lower bound, and $\alpha_2$ upper bound.

## How Many Items Are Enough?

Given a choice, how many items should a researcher include in a scale?[4] What is implied for the "equilibrium" number of items in a scale? First, consider the criterion of maximally improving the estimate of coefficient alpha in a scale. We have established the critical values of $c$ for which alpha increases when adding one the baseline case to the "one bad item" case—call this "Illustration 1." We also derive the critical values for which alpha improves when adding two items that move the baseline case to one of the heterogeneous "two underlying factors" case (call this "Illustration 2"), as well as the scenario in which adding one item causes movement from the one bad item case to the two underlying factors case (Illustration 3). The results for $p$ that maximize alpha depend jointly on $c$, $r$, and $p$. Figure 2 shows the results for $p = 3$, and Figure 3 shows $p = 7$.

Figure 2 delineates four regions of the $\{r, c\}$ space. Analysis of the researcher's possible choices leads to the following result:

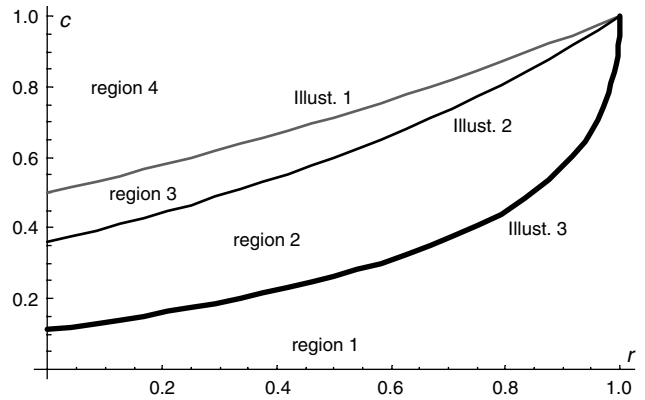
Result 1. When the researcher's goal is to maximize the size of alpha (and the initial scale has three items), the equilibrium number of items is as follows:

- the baseline $p$ items, if the $\{r, c\}$ pair lies in either region 1 or 2;
- the two underlying factors ($p = 5$) scale if the $\{r, c\}$ pair lies in either region 3 or 4.

It is never an equilibrium in this situation to add just one bad item to the baseline scale, given the opportunity to add another item that creates a second underlying factor. The result rests on the relative size of $c$ and $r$, given the low number of original items in

[4] Similar computational analyses are discussed by Rust and Cooil (1994) for the qualitative case.

**Figure 2** $\{r, c\}$ **Values That Keep Alpha Constant for Illustrations 1, 2, and 3 (for $p = 3$)**



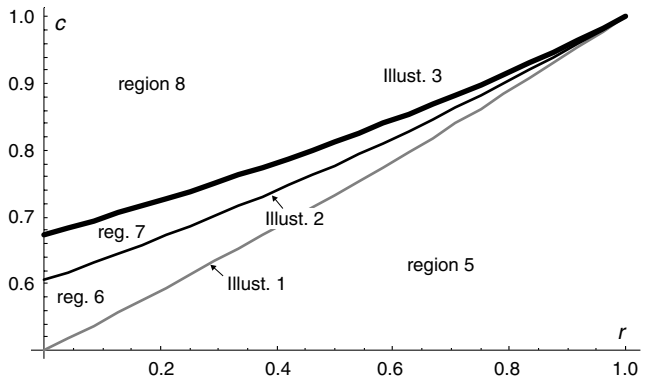

*Notes.* The curve Illustration 1 depicts $\{r, c\}$ pairs for which moving from the baseline case to the one bad item case holds alpha constant. Illustration 2 moves from the baseline case to the two underlying factors case, and Illustration 3 depicts moving from the one bad item case to the two underlying factors case. Points above each curve are $\{r, c\}$ pairs for which alpha increases when the indicated change is made; for points below each curve, alpha decreases.

the scale ($p = 3$). When $c$ is high enough relative to $r$, the beneficial effect of adding items swamps the negative effect of creating a scale with two underlying factors. However, if $c$ is too low relative to $r$, it is best for the researcher to stay with the original three-item scale.

In comparison, consider Figure 3, which depicts the $p = 7$ analysis:

This figure also has four regions, but the alpha-constant curves are arrayed in the opposite order from those for $p = 3$. Analysis of regions 5–8 in Figure 3 generates the following result.

**Figure 3** $\{r, c\}$ **Values That Keep Alpha Constant for Illustrations 1, 2, and 3** ($p = 7$)




*Notes.* The bottom curve is the locus of $\{r, c\}$ pairs that keep alpha constant when moving from the baseline case to "one bad item" ("Illust. 1"). The middle curve moves from baseline to "two underlying factors" ("Illust. 2"). The top curve moves "one bad item" to "two factors" ("Illust. 3").

Result 2. When the researcher's goal is to maximize the size of alpha (and the initial scale has seven items), the equilibrium number of items is:

- the baseline seven items, if the $\{r, c\}$ pair lies in region 5;
- the eight-item "one bad item" scale in regions 6 and 7; and
- the nine-item "two underlying factors" scale in region 8.

These analyses demonstrate that longer scales are not always better when the goal is to maximize alpha. This is particularly true when $c$ is low relative to $r$, especially when $p$ is large. There is a cost of adding items to a scale if they lower the overall interitem correlations, which swamps the benefit of a longer scale. The results show clearly that $p$, scale length, is an insufficient statistic to indicate whether or not adding more items is a good idea.

## Conclusion

These analyses should have clear applicability to researchers in marketing. Analysis 1 demonstrates that alpha's standard error is inversely related to sample size; thus, researchers seeking to improve the predictive ability of their scales can do so indirectly through increasing sample size. Analysis 1 also demonstrates the beneficial effects of scale length ($p$) on alpha and its standard error, analytically validating the Spearman-Brown formula (Ghiselli 1964). These findings show that scale length has both direct (via its influence on alpha) and indirect (via its influence on the standard error) measurement benefits to the researcher. Finally, the effects of item intercorrelations on both the standard error and alpha are dramatic; stronger correlations among the items drastically reduce the standard error and increase alpha.

Analysis 1 also proves analytically that $p$ and $r$ are substitutes in their beneficial (positive) effects on alpha, and that $p$, $r$, and $n$ are substitutes in their beneficial (negative) effects on alpha's standard error. Researchers seeking to improve their scales have multiple means of doing so, and at low-to-moderate levels of alpha or standard error, the marginal gains are substantial. Our findings also speak to a long-observed measurement reliability adage: Heterogeneity within the covariance matrix negatively impacts reliability. Specifically, it decreases the precision of the alpha estimate. Our analysis also provides insights to the researcher considering adding items to a scale. It is not always alpha enhancing to add items; it depends on the length of the original scale and their correlations. The items added must be of increasingly high quality (in terms of their correlations with the original items), the higher either $p$ or $r$ is to improve alpha. We examined the question of "equilibrium" scale length,

showing the conditions defining when it is optimal to add no items, multiple items, or just one to a scale, given a choice among these options.

Taken together, the results from these analytical demonstrations have profound ramifications for measurement and marketing science. These findings represent a clear advancement over contemporary measurement practice and directly implicate myriad research questions for marketing scholars. Of course, per measurement theory (cf. Rossiter 2002), we would not exhort researchers to lengthen scales by writing redundant items. Doing so elicits respondent boredom and misses the opportunity to enhance predictive validity by including additional facets of the behavior to be predicted; we would urge that the reliability of each facet measured be noted. Extensions of the current research could examine the variance with smaller samples given its asymptotic nature (but note that the statistic behaves reasonably quickly, e.g., with $n = 50$ and 30 if $p$ or $r$ are high), or with variant distributions (e.g., nonnormal or asymmetric distributions when $r$ is very low or high) to probe robustness further. We close with some prescriptions regarding the assessment of reliability and the reporting of coefficient alpha.

## Best Practices

Researchers should consider several key points in assessing the reliability of their measures:

(1) *Item correlation and scale length are critical, influencing both alpha and its standard error.*

(2) *Sample size has a significant effect on the precision of the estimation of alpha. A focus only on alpha, and not on its standard error, fails to capture this important effect and is misguided.*

(3) *Covariance heterogeneity compromises measurement (heretofore not demonstrated analytically); it has a deleterious effect on the standard error, widening the confidence interval.*

Confidence intervals or hypothesis tests are *de rigueur* in statistical analyses and the same standard should also be applicable to measurement, rather than assessing alpha against the familiar though arbitrary rule of thumb of 0.70. Once inferential statistics are available, it becomes no longer sufficient to subjectively judge reliability solely on the basis of a point estimate. Accordingly, *reporting of the familiar alpha point estimate should be supplemented with the standard error or confidence interval around alpha.*

### Acknowledgments

The authors are grateful to Professor Steve Shugan, the AE and the reviewers, participants at the European ACR meeting, and seminar attendees from the marketing departments at the University of Colorado, University of Michigan, Northwestern University, Notre Dame, and the University of Southern California for their helpful feedback on this research.

## Appendix

Proposition 1.

$$\frac{\partial\alpha}{\partial n}=0 \quad \frac{\partial\alpha}{\partial p}=-\frac{r(r-1)}{[1+r(p-1)]^2}>0 \quad \frac{\partial\alpha}{\partial r}=\frac{p}{[1+r(p-1)]^2}>0$$

$$\frac{\partial SE}{\partial n}=-\frac{\sqrt{Q/n}}{2n}<0 \quad \frac{\partial Q}{\partial p}=-\frac{2(r-1)^2[1+r(2p^2-p-1)]}{(p-1)^2[1+r(p-1)^3}<0$$

$$\frac{\partial Q}{\partial r}=-\frac{4p^2(r-1)}{(p-1)[1+r(p-1)]^3}<0$$

Proposition 2.

$$\frac{\partial^2\alpha_1}{\partial p\,\partial r}=\frac{1-r-pr}{[1+(p-1)r]^3}, \quad >0 \text{ for } r<\frac{1}{1+p}, \quad \text{and } <0 \text{ for } r>\frac{1}{1+p}.$$

$$\frac{\partial^2 SE_1}{\partial n\,\partial p}=\frac{(r-1)^2[1+r(2p^2-p-1)]}{2n\sqrt{n}(p-1)^2[1+r(p-1)]^3}>0;$$

$$\frac{\partial^2 SE_1}{\partial n\,\partial r}=\frac{p^2(r-1)}{n\sqrt{n}(p-1)[1+r(p-1)]^3}>0;$$

$$\frac{\partial^2 Q}{\partial p\,\partial r}=\frac{4p(r-1)(2rp^2-p+2(1-r))}{(p-1)^2[1+r(p-1)]^4}>0;$$

$$\frac{\partial^2 SE_1}{\partial p\,\partial r}=\frac{4p(r-1)(2rp^2-p+2(1-r))}{2\sqrt{n}(p-1)^2[1+r(p-1)]^4}>0.$$

Proposition 3.

$$\frac{\partial\alpha_2}{\partial c}=\frac{2rp^2}{[p+2cr(p-1)+r(p-1)(p-2)]^2}>0,$$

$$\frac{\partial Q_2}{\partial c}=\left[\frac{1}{(p-1)(p^2r+2cpr-3pr+p-2cr+2r)^4}\right]\times\{8p^2r[-4r^3(c-1)^2+p^4r(r-1)(r(c-1)+1)+2pr(c-1)(r^2(5c-6)-r(c-5)-3)-p^2r(c-1)(r^2(8c-13)-r(c-19)-9)+p^3r(r^3(c^2-4c+3)+r^2(c^2+9c-11)+r(5-3c)-1)]\}<0.$$

Proposition 4.

$$\frac{\partial\alpha_3}{\partial c}=\frac{2rp^2}{(p-1)[2+r(p+cp-2)]^2}>0,$$

$$\frac{\partial SE_3}{\partial c}=\frac{8p^2r[-2r(1-c)(1-r)-p[2-r(1+c)+r^2(1-c)]]}{(p-1)^2[2+r(cp+p-2)]^3}<0.$$

## References

Barchard, Ken A., Ralph Hakstian. 1997. The robustness of confidence intervals for coefficient alpha under violation of the assumption of essential parallelism. *Multivariate Behavioral Res.* **32** 169–191.

Bearden, William O., Richard G. Netemeyer. 1999. *Handbook of Marketing Scales: Multi-Item Measures for Marketing and Consumer Behavior Research*, 2nd ed. Sage, Thousand Oaks, CA.

Brown, Christina L. 1999. "Do the right thing": Diverging effects of accountability in a managerial context. *Marketing Sci.* **18**(3) 230–246.

Bruner, II, Gordon C., Paul J. Hensel. 1994. *Marketing Scales Handbook: A Compilation of Multi-Item Measures*. American Marketing Association, Chicago, IL.

Chen, Haipeng (Allan), Akshay R. Rao. 2002. Close encounters of two kinds: False alarms and dashed hopes. *Marketing Sci.* **21**(2) 178–196.

Cooil, Bruce, Roland T. Rust. 1994. Reliability and expected loss: A unifying principle. *Psychometrika* **59**(2) 203–216.

Cronbach, Lee J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* **16** 297–334.

Dhar, Sanjay K., Claudia González-Vallejo, Dilip Soman. 1999. Modeling the effects of advertised price claims: Tensile versus precise claims? *Marketing Sci.* **18**(2) 154–177.

Drolet, Aimee L., Donald G. Morrison. 2001. Do we really need multiple-item measures in service research? *J. Service Res.* **3**(5) 196–204.

Finn, Adam. 1992. Recall recognition and the measurement of memory for print advertisements: A reassessment. *Marketing Sci.* **11** 95–100.

Ghiselli, Edwin E. 1964. *Theory of Psychological Measurement.* McGraw Hill, New York.

Häubl, Gerald, Valerie Trifts. 2000. Consumer decision making in online shopping environments: The effects of interactive decision aids. *Marketing Sci.* **19**(1) 4–21.

Hoch, Stephen J., Eric T. Bradlow, Brian Wansink. 1999. The variety of an assortment. *Marketing Sci.* **18**(4) 527–546.

Iacobucci, Dawn, Doug Grisaffe, Adam Duhachek, Alberto Marcati. 2003. FAC-SEM: A methodology for modeling factorial structural equations models, applied to cross-cultural and cross-industry drivers of customer evaluations. *J. Service Res.* **6**(1) 3–23.

Kahn, Barbara E., Mary Frances Luce. 2003. Understanding high-stakes consumer decisions: Mammography adherence following false-alarm test results. *Marketing Sci.* **22**(3) 393–410.

Kalwani, Manohar, Alvin Silk. 1983. On the reliability and predictive validity of purchase intention measures. *Marketing Sci.* **1**(3) 243–286.

Laurent, Gilles, Jean-Noel Kapferer, Francoise Roussel. 1995. The underlying structure of brand awareness scores. *Marketing Sci.* **14**(3) G170–G179.

Lynch, Jr., John G., Dan Ariely. 2000. Wine online: Search costs affect competition on price, quality, and distribution. *Marketing Sci.* **19**(1) 83–103.

Malhotra, Naresh K., M. Peterson, S. B. Kleiser. 1999. Marketing research: A state-of-the-art review and directions for the twenty-first century. *J. Acad. Marketing Sci.* **27**(2) 160–183.

Novak, Thomas P., Donna L. Hoffman, Yiu-Fai Yung. 2000. Measuring the customer experience in online environments: A structural modeling approach. *Marketing Sci.* **19**(1) 22–42.

Nunnally, Jum C., Ira H. Bernstein. 1994. *Psychometric Theory*, 3rd ed. WCB/McGraw-Hill, New York.

Peterson, Robert A. 1994. A meta-analysis of Cronbach's coefficient alpha. *J. Consumer Res.* **21**(September) 381–391.

Rao, Akshay R., Humaira Mahi. 2003. The price of launching a new product: Empirical evidence on factors affecting the relative magnitude of slotting allowances. *Marketing Sci.* **22**(2) 246–268.

Rossiter, John R. 2002. The C-OAR-SE procedure for scale development in marketing. *Internat. J. Res. Marketing* **19**(4) 305–335.

Rust, Roland T., Bruce Cooil. 1994. Reliability measures for qualitative data: Theory and implications. *J. Marketing Res.* **31**(February) 1–14.

Rust, Roland T., J. Jeffrey Inman, Jianmin Jia, Anthony Zahorik. 1999. What you *don't* know about customer-perceived quality: The role of customer expectation distributions. *Marketing Sci.* **18**(1) 77–92.

Soman, Dilip, Amar Cheema. 2002. The effect of credit on spending decisions: The role of the credit limit and credibility. *Marketing Sci.* **21**(1) 32–53.

van Zyl, J. Martin, N. Heinz, D. G. Nel. 2000. On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika* **65** 271–280.