**Journal of Psychoeducational Assessment**
Volume 27 Number 5
October 2009 409-420

10.1177/0734282908330937
http://jpa.sagepub.com
hosted at
http://online.sagepub.com

# The Average Distance Between Item Values

## A Novel Approach for Estimating Internal Consistency

Edward D. Sturman
*State University of New York, Plattsburgh*
Robert A. Cribbie
Gordon L. Flett
*York University*

This article presents a method for assessing the internal consistency of scales that works equally well with short and long scales, namely, the average proportional distance. The method provides information on the average distance between item scores for a particular scale. In this article, we sought to demonstrate how this relatively simple statistic could be calculated and present examples that show its advantage over traditional methods. Simulation and empirical tests were conducted to establish standards for the average proportional distance of scores. The implications for test construction are discussed with a particular emphasis on the advantages of developing shorter scales for psychological and educational research.

***Keywords:*** *internal consistency; reliability; personality; test construction*

The reliability or consistency of a particular scale can be assessed with a variety of methods, including test–retest reliability (the temporal stability of scores between one administration and another), alternate forms reliability (comparing scores on equivalent versions of a test), split-half methods (comparing scores on half of a scale to the other half), and internal consistency reliability (see Murphy & Davidshofer, 1998). It is with this latter form of reliability that the present article is concerned. Internal consistency reliability can be broadly defined as the extent to which items on a particular scale are related to one another (see Clark & Watson, 2003). There currently exist several methods for assessing the internal consistency reliability of scales, although Cronbach's alpha is by far the most widely used. Cronbach (1951) originally described this statistic as the mean of all possible split-half reliabilities for a particular scale.

Cronbach's alpha is generally regarded as a very useful statistic to assess internal consistency reliability. However, Clark and Watson (2003) have argued that Cronbach's alpha is an imperfect measure of internal consistency reliability as values are a function of the average interitem correlations and the number of items, which should be irrelevant. With a large number of items, it is difficult to not achieve high values. For example, Cortina (1993) showed that it was possible to achieve high Cronbach's alpha values if one increased the

number of items on a scale, even if interitem correlations were low and the scale was multidimensional.

One conclusion that can be drawn from these examples is that the calculation of coefficient alpha is less than optimal in certain assessment situations. Indeed, in one of his final papers, Cronbach outlined some of his concerns about the use of coefficient alpha to calculate internal consistency reliability, and he cautioned that there are several measurement issues and concerns that cannot be resolved by relying on this particular statistic (see Cronbach & Shavelson, 2004). Other authors have noted similarly that other methods and criteria need to be considered when evaluating the internal consistency reliability of a set of scale items (e.g., Streiner, 2003a, 2003b).

Another popular measure of internal consistency is the interitem correlation. As noted by Briggs and Cheek (1986), the mean interitem correlation is a very informative statistic in evaluating the homogeneity of a scale and its degree of bandwidth and fidelity. However, a problem with the interitem correlation as a measure of internal consistency is that it may yield high values even when item means are far apart.

One potential solution is to turn reliability on its head so that we instead focus on the average distance between scores on scale items. In the present article, we put forward a method that provides a numerical value for the typical agreement between scale items. As a descriptive statistic, it is similar to a standard deviation in terms of the intuitive understanding it affords. It provides researchers with the actual degree of difference in responses to various items on a scale and not their similarity in rank (as with interitem correlations). Thus, differences in mean scores on various items, which would not necessarily affect correlation coefficients, will be detected with this method. Its primary advantage over Cronbach's alpha is that it is not influenced by the number of items on a scale.

This approach is therefore a departure from classical reliability theory, which holds that reliability is increased with a greater number of items. Granted, some redundancy is required to evaluate the reliability of a particular scale, but there is a point of diminishing returns that scale developers regularly cross. When scale developers use many items to tap a single, narrow construct, this can serve only to increase Cronbach's alpha at the expense of precision and parsimony. Indeed, Burisch (1984a, 1984b) has argued that a parsimonious approach to scale construction is needed given that repetitive questioning may lead to boredom and fatigue, which will increase error variance. This issue is becoming increasingly salient with a number of authors developing short personality scales in recent years. The move toward shorter scales makes obvious sense from the perspective of time and cost. It also makes sense when administering scales to clinical participants with diminished attentional capacity. We can expect this trend to continue as shorter scales enable greater accessibility to a larger population through Internet testing (where long scales would hinder responding). There is also little evidence that longer tests yield data that are more valid than the data from shorter tests (Burisch, 1984b, 1997).

The proposed method involves simply calculating the average distance between scores on the items that constitute a scale. With two items, this would involve calculating a difference score for each case and taking an average of the difference scores. With three or more items, the absolute difference between items would be calculated, and the differences would be averaged. For instance, when using the method on a three-item scale, the absolute

difference between Items 1 and 2, 2 and 3, and 1 and 3 would be calculated. These differences would then be averaged. An average deviation of 0 would indicate perfect agreement between the items on a scale. We can also calculate the average distance as a proportion by simply dividing the average distance by the number of response options (minus one) for a given scale. Doing so allows us to establish standards and compare one scale to another on internal consistency, regardless of the number of response options.

We would argue that this statistic is important in evaluating internal consistency not only for short scales but also for longer scales that consist of many items. Although it is simple, the information it provides would seem to be a basic first step in determining the extent to which the items on a particular scale deviate from each other. In testing this method, we sought to compare it to established measures of internal consistency reliability (i.e., Cronbach's alpha and mean interitem correlations). We also recognized the need to establish a set of standards that could be used by researchers in the field. To this end, we conducted a simulation study of the statistic where we varied the number of response options (or item categories) and used interitem correlations as a comparator. We also used simulation to compare average distance to Cronbach's alpha with scales of different lengths and interitem correlations. The standards that emerged from the simulation studies were then examined in relation to empirical data.

## Average Proportional Distance (APD) and Interitem Correlations

An important consideration with novel methods of quantifying internal consistency is how the new methods relate to previously established measures. For example, an important consideration in understanding average distance measures of internal consistency is to understand how they relate to interitem correlations.

We simulated random normal variates with specified interitem correlations and number of item categories to explore how the number of item categories and the population correlation coefficient between the items relate to the values of the average distance and APD statistics. Specifically, two random normal variates ($N$ = 100,000) were generated with population correlation coefficient $\rho$, where $\rho$ was set at .4, .5, .6, .7, .8, or .9. The random normal deviates were then converted to categorical variables with c categories, where c was set at 3, 5, or 7. The categories were created such that the expected number of responses in each category would be equal. Average distance and APD statistics were then computed from the categorical variables.

Table 1 displays the average distance and APD statistics as a function of the number of item categories, the population correlation coefficient between items, and the correlation between the sample categorical variables. As expected, the correlation between the categorical variables was lower than the correlation between the underlying continuous variables, and the difference between the correlations is larger with a smaller number of item categories. The average distance statistics increase as the correlation between the items decrease. As can be seen in Table 1, there is a slight penalty for having fewer response options as the item responses do not reflect the true underlying process as accurately as with a greater number of response options. This same penalty applies to the sample interitem correlations.

**Table 1**
**Average Distance Statistics for the Number of Item Categories and the Correlation Between Items With an Equal Probability for Each Item Category**

| Number of Item Categories | *r* (Cont) | *r* (Cat) | Avg. Dist. | Avg. Prop. Dist |
|---|---|---|---|---|
| 3 | .900 | .769 | 0.272 | .136 |
| | .800 | .669 | 0.384 | .192 |
| | .700 | .570 | 0.469 | .234 |
| | .600 | .482 | 0.542 | .271 |
| | .500 | .400 | 0.603 | .302 |
| | .400 | .316 | 0.666 | .333 |
| 5 | .900 | .854 | 0.483 | .121 |
| | .800 | .750 | 0.681 | .170 |
| | .700 | .645 | 0.842 | .210 |
| | .600 | .550 | 0.979 | .245 |
| | .500 | .448 | 1.100 | .275 |
| | .400 | .357 | 1.212 | .303 |
| 7 | .900 | .872 | 0.684 | .115 |
| | .800 | .766 | 0.973 | .162 |
| | .700 | .663 | 1.202 | .200 |
| | .600 | .566 | 1.390 | .232 |
| | .500 | .466 | 1.560 | .261 |
| | .400 | .372 | 1.722 | .287 |

Note: *r* (cont) = correlation between the underlying continuous variables; *r* (cat) = empirical correlation between the discrete (categorized) variables; Avg. Dist. = average distance statistic; Avg. Prop. Dist. = average proportional distance statistic.

To determine the internal consistency for Cronbach's alpha and the APD as a function of the number of, and correlation between, items, we simulated data ($N = 100{,}000$) with a known interitem population correlation ($r = .1, .3, .5$, and $.7$) for scales with five response options and 3, 10, 25, 50, or 100 items. The interitem population correlation represents the correlation between the underlying continuous variables (i.e., before discretization). The results are displayed in Table 2. The inflation of Cronbach's alpha with the number of items is clearly seen with this simulation. For example, when a scale contained 100 items and the items were poorly correlated with one another (i.e., $r = .10$), the scale nevertheless exhibited a high value for Cronbach's alpha (.898). In contrast, APD showed only minute changes in connection with the number of items.

Exploring the relationship between the interitem correlations and the APD also suggests that cutoffs for the APD could be qualified in a similar manner to those used for correlations. For example, if we assume that an acceptable correlation among scale items (specifically the correlation among the underlying continuous variables) is somewhere between .6 or .7 (i.e., 36% to 49% shared variability), then the following recommendations can be made for qualifying the APD statistic: very good/excellent: 0 to .20; acceptable: .20 to .25.

**Table 2**
**Internal Consistency for Cronbach's Alpha and the Average Proportional Distance, as a Function of the Correlation Between Items and the Number of Items**

| *R* Between Items | | | | |
|---|---|---|---|---|
| Cont. | Dich. | Number of Items | APD | Alpha |
| .10 | .08 | 3 | .379 | .151 |
| | | 10 | .380 | .443 |
| | | 25 | .380 | .681 |
| | | 50 | .380 | .812 |
| | | 100 | .380 | .898 |
| .30 | .27 | 3 | .330 | .349 |
| | | 10 | .330 | .707 |
| | | 25 | .330 | .866 |
| | | 50 | .331 | .928 |
| | | 100 | .330 | .963 |
| .50 | .45 | 3 | .275 | .467 |
| | | 10 | .275 | .803 |
| | | 25 | .275 | .916 |
| | | 50 | .276 | .957 |
| | | 100 | .277 | .978 |
| .70 | .65 | 3 | .211 | .563 |
| | | 10 | .212 | .852 |
| | | 25 | .212 | .939 |
| | | 50 | .211 | .970 |
| | | 100 | .211 | .985 |

Note: Cont. = population correlation between underlying continuous items; Dich. = population correlation between items after discretization; APD = average proportional distance; Alpha = Cronbach's alpha; Number of response options = 5.

As always, it is important to point out that generic cutoffs are only appropriate when very little information is available for qualifying the meaning of the average distance statistics. More specifically, when previous information is available regarding the size of the average distance statistics, this should always be used to qualify the nature of any sample average distance statistic. For example, certain short scales will be designed such that the interitem correlations are expected to be very large, whereas others may be tapping more distinct information and the interitem correlations are not expected to be as large. This information is much more valuable in qualifying the nature of the average distance statistics than generic cutoffs. Furthermore, some studies are designed to compare the responses of different samples to the scale items, and again, the differences in the average distance statistics across the samples will be more informative than any generic cutoffs.

Although, in this article, we simulated data where there was an equal probability of responses falling into any of the response categories, many times there will be more responses for certain categories. As the APD measures distance, if most responses fall in a

small number of related categories (e.g., on a five-item scale most people select the Options 1 or 2, with few selecting Options 3, 4, or 5), then the distance between items (and hence the APD) will decrease relative to when there are a similar number of responses for each option. However, if most responses fall in the extreme categories (e.g., on a five-item scale most people select the Options 1 or 5, with few selecting Options 2, 3, or 4), then the distance between items (and hence the APD) will increase relative to when there are a similar number of responses for each option.

## Empirical Testing of Average Distance Standards

### Example 1

Several versions of the Center for Epidemiologic Studies–Depression Scale (CES-D; Radloff, 1977) were tested in relation to various measures of internal consistency including average distance. Specifically, average distance was compared to the mean interitem correlation and Cronbach's alpha (Cronbach, 1951) for each version of the CES-D to determine (a) whether average distance varied from instrument to instrument in the same manner as other statistics of internal consistency reliability and (b) as a preliminary test of the proposed cut points. Fortunately, several short scales have been derived from the CES-D, but we also created our own versions, which were informed by factor analysis. The sample consisted of 320 undergraduate university students (112 men, 208 women) with an average age of 18.9 years ($SD$ = 2.3). Data pertaining to the CES-D were available for 319 cases.

*Measures*

*CES-D.* The CES-D (Radloff, 1977) is a widely used self-report measure originally designed to assess levels of depression in the general population. Radloff (1977) tested the psychometric properties of the scale and found that it achieved a high degree of internal consistency reliability ranging from coefficient alpha levels of 0.85 in the general population to 0.90 in a clinical sample. Radloff found four factors in the CES-D: somatic symptoms, depressed affect, positive affect, and interpersonal problems.

Several versions of the CES-D have been developed as short screening tools for depression. For example, Shrout and Yager (1989) selected 5 items from the CES-D that best predicted membership in depressed versus nondepressed groups. This scale demonstrated somewhat low Cronbach's alpha (.66 and .76 for a community and patient sample, respectively), but Shrout and Yager (1989) made a good point that items should not be selected on the basis of alpha values but rather on the basis of which items yield greater validity (the axiom that reliability is necessary for validity remains true, but Shrout and Yager argued that Cronbach's alpha is not a perfect measure of reliability). In contrast, Melchior, Huba, Brown, and Reback (1993) devised 8-item and 4-item versions of the CES-D by excluding items that would lead to the smallest decrements in Cronbach's alpha. As a result, their scales exhibited fairly high Cronbach's alpha at .86 and .81 for the 8-item and 4-item versions, respectively. The 8-item and 4-item scales correlated with the full 20-item version at .93 and .87, respectively. In the present study, we used the Shrout and Yager (1989) 5-item

CES-D and the 8-item and 4-item versions developed by Melchior et al. (1993). In addition, we developed 3-item scales that were deemed to be unidimensional or multidimensional (as determined by factor analysis—see below).

*Results*

A principal component analysis was conducted to create short unidimensional and multidimensional scales from the CES-D. Three components with eigenvalues of more than 1 were obtained. The first component seemed to correspond most closely to depressed affect and accounted for 37.38% of the total variance. Items dealing with a lack of positive affect loaded highest on the second component, which accounted for 10.74% of the total variance. The third component seemed to represent an interpersonal focus although factor loadings were quite weak, limiting interpretation. This third factor accounted for only 5.51% of the total variance. The results are generally in accord with past factor analyses of the CES-D, which tend to yield four factors—depressed affect, lack of positive affect, interpersonal problems, and somatic symptoms (for a review, see Shafer, 2006).

A short 3-item version of the CES-D was created by selecting the 3 items with the highest factor loadings on the first component (Items 3, 6, and 18; see Table 3). This short version was therefore thought to represent a unidimensional construct (depressed affect) and could be expected to possess adequate internal consistency. A second, multidimensional version was created by taking the 3 items with the highest factor loadings on each of the three components (Items 6, 8, and 15). This scale could be expected to demonstrate low internal consistency. We should note that we do not endorse either of these scales as valid screens for depressive symptoms, and they were developed for the sole purpose of demonstration.

As shown in Table 4, all measures of internal consistency reliability were high for the three-item unidimensional version of the CES-D and low for the three-item multidimensional version. According to the criteria proposed earlier, the three-item unidimensional version could be said to possess very good/excellent internal consistency (i.e., APD < .20). Cronbach's alpha could also be described as high for this scale. Although there are no strict standards for the mean interitem correlation, the unidimensional three-item scale had the highest interitem correlations compared to every other version of the CES-D. In contrast, the three-item multidimensional version of the CES-D exhibited the lowest Cronbach's alpha and mean interitem correlation, and the highest average distance statistics. According to the Cronbach's alpha value, the scale would be categorized as having poor internal consistency reliability in this sample. Similarly, it did not meet the criterion for an adequate scale using the APD standard (i.e., <.250). Therefore, the APD method converged with other measures of internal consistency reliability for the three-item scales and could discriminate between internally consistent and noninternally consistent scales.

All three statistics of internal consistency reliability were in line with each other for the four-item version of the CES-D. However, differences between the methods could be seen in regards to the five-item and eight-item scales. Although mean interitem correlations were consistent with the average distance statistics, Cronbach's alpha was highest for the longer scale. Specifically, the eight-item scale had the highest Cronbach's alpha values of any version of the CES-D, yet according to both the mean interitem correlations and average distance methods it ranked third among all of the scales. We should also mention that the

**Table 3**
**Factor Loadings for the CES-D**
**(Principal Components Analysis)—Example 1 (*n* = 319)**

| CES-D Item Number | Component 1 | Component 2 | Component 3 |
|---|---|---|---|
| 1 | .578 | –.208 | –.305 |
| 2 | .487 | –.200 | –.024 |
| 3 | .744[a] | –.010 | –.297 |
| 4 | .292 | .646 | .221 |
| 5 | .527 | –.222 | –.448 |
| 6 | .825[a,b] | .022 | –.168 |
| 7 | .374 | –.396 | .042 |
| 8 | .302 | .703[b] | –.092 |
| 9 | .709 | .041 | .228 |
| 10 | .679 | –.073 | .161 |
| 11 | .567 | –.219 | –.069 |
| 12 | .511 | .626 | –.201 |
| 13 | .546 | –.286 | .254 |
| 14 | .737 | –.068 | .076 |
| 15 | .560 | –.054 | .533[b] |
| 16 | .548 | .619 | –.015 |
| 17 | .660 | –.101 | .067 |
| 18 | .784[a] | –.046 | –.082 |
| 19 | .740 | –.041 | .341 |
| 20 | .694 | –.107 | –.163 |

Note: CES-D = Center for Epidemiologic Studies–Depression Scale.
a. Items used to create a unidimensional three-item CES-D.
b. Items used to create a multidimensional three-item CES-D.

**Table 4**
**Measures of Internal Consistency for Various**
**Scales Derived From the CES-D—Example 1 (*n* = 319)**

| Scale | Cronbach's Alpha | Mean Interitem Correlation | Average Distance | Average Proportional Distance |
|---|---|---|---|---|
| Three-item CES-D (unidimensional)[a] | .855 | .662 | .537 | .179 |
| Three-item CES-D (multidimensional)[a] | .475 | .232 | .882 | .294 |
| Four-item CES-D[b] | .847 | .583 | .622 | .207 |
| Five-item CES-D[c] | .803 | .450 | .757 | .253 |
| Eight-item CES-D[b] | .882 | .486 | .705 | .235 |

Note: CES-D = Center for Epidemiologic Studies–Depression Scale.
a. Derived from principal components analysis.
b. Derived from Melchior, Huba, Brown, and Reback (1993).
c. Derived from Shrout and Yager (1989).

unidimensional three-item scale clearly had the highest mean interitem correlation and lowest APD, yet it ranked second in terms of Cronbach's alpha. Therefore, at least in this example, the APD was closer to the interitem correlations than Cronbach's alpha and seemed to be a more sensitive statistic.

## Example 2: Replication of Empirical Findings

A second dataset was used to test the replicability of the empirical findings above. This dataset consisted of 170 female undergraduate students who had completed the CES-D. Cases with intermediate values on the CES-D were removed leaving 152 participants (mean age = 20.0 years, *SD* = 2.23). As in the previous example, the CES-D was factor analyzed to produce unidimensional and multidimensional scales. The four-item, eight-item, and five-item versions of the CES-D were also compared on the various internal consistency statistics.

### *Results*

The factor analysis produced a different solution than the previous example such that five components had eigenvalues greater than 1, rather than three components. However, the first factor was essentially the same with Items 3, 6, and 18 having the highest factor loadings (see Table 5). As we had done previously, these 3 items were used to create a 3-item unidimensional scale. The items with the highest loadings on each of the first three components were used to create a 3-item multidimensional scale. Again, we were not interested in the validity of this scale but rather sought to compare an internally consistent scale to one lacking internal consistency.

In comparing Tables 4 and 6, it is apparent that the findings were remarkably consistent in both samples. For example, APD was exactly the same for the multidimensional measure (even though it was composed of different items). The values for Cronbach's alpha and the mean interitem correlation were also similar to those obtained for the multidimensional measure in the first sample. All three statistics strongly suggest that this version of the CES-D lacked internal consistency. In contrast, the APD for the unidimensional measure was below .200, which would be considered very good/excellent according to the proposed standards. This judgment was supported by the fact that this scale also had the highest interitem correlations in both samples.

A similar pattern of results between the first and second samples was found for the other versions of the CES-D as well. For example, the five-item scale (Shrout & Yager, 1989) would fall below the threshold on APD for adequate internal consistency. The mean interitem correlation for this scale was also relatively low. Cronbach's alpha for the five-item scale was lower than in the first sample but could still be interpreted as adequate. As before, Cronbach's alpha tended to favor the eight-item version, even though it was not the most internally consistent scale according to the APD and mean interitem correlations. The average distance method and mean interitem correlations were in line with each other such that each scale would be ranked in the same manner by both methods.

Therefore, the results found with the first sample were replicated in the second sample. The only major difference was that the four-item scale (Melchior et al., 1993) would be

**Table 5**
**Factor Loadings for the CES-D**
**(Principal Components Analysis)—Example 2 (*n* = 152)**

| CES-D Item Number | Component 1 | Component 2 | Component 3 |
|---|---|---|---|
| 1 | .394 | .295 | .255 |
| 2 | .398 | .148 | .340 |
| 3 | .734[a] | .098 | .029 |
| 4 | .459 | –.600[b] | –.173 |
| 5 | .507 | .338 | –.322 |
| 6 | .776[a] | –.077 | .034 |
| 7 | .454 | .292 | –.126 |
| 8 | .420 | –.513 | –.006 |
| 9 | .601 | –.155 | –.271 |
| 10 | .605 | .256 | –.252 |
| 11 | .539 | .066 | .147 |
| 12 | .623 | –.503 | .137 |
| 13 | .519 | .200 | .383 |
| 14 | .678 | .137 | .015 |
| 15 | .469 | .014 | –.038 |
| 16 | .634 | –.474 | .027 |
| 17 | .471 | .055 | .645[b] |
| 18 | .809[a,b] | .111 | .131 |
| 19 | .607 | .097 | –.386 |
| 20 | .597 | .231 | –.358 |

Note: Only the first three factors are presented. CES-D = Center for Epidemiologic Studies–Depression Scale.
a. Items used to create a unidimensional three-item CES-D.
b. Items used to create a multidimensional three-item CES-D.

**Table 6**
**Measures of Internal Consistency for Various**
**Scales Derived From the CES-D—Example 2 (*n* = 152)**

| Scale | Cronbach's Alpha | Mean Interitem Correlation | Average Distance | Average Proportional Distance |
|---|---|---|---|---|
| Three-item CES-D (unidimensional) | .817 | .598 | .522 | .174 |
| Three-item CES-D (multidimensional) | .511 | .277 | .882 | .294 |
| Four-item CES-D | .814 | .519 | .591 | .197 |
| Five-item CES-D | .719 | .339 | .755 | .252 |
| Eight-item CES-D | .849 | .413 | .668 | .223 |

Note: CES-D = Center for Epidemiologic Studies–Depression Scale.
a. Derived from principal components analysis.
b. Derived from Melchior et al. (1993).
c. Derived from Shrout and Yager (1989).

categorized differently according to the standards proposed here. Specifically, the APD in the first example was slightly above the threshold for very good/excellent internal consistency (i.e., <.20) but met this standard in the second example. However, in real terms, the APD values for this scale were highly similar in both examples with a difference of only .01. Thus, we should caution about being overly strict with the criteria. This caution extends to judgments of the five-item version as well, which should not be viewed as having inadequate internal consistency but rather that it may have some problems in this regard.

## Conclusions

In the present article, we sought to demonstrate that the APD method may be a practical alternative to other measures of internal consistency as it can be used in scales with varying lengths and response options. The present investigation also established the standards of internal consistency for the APD method, and these standards were shown to discriminate between scales. In other words, internally consistent scales were clearly distinguishable from those lacking internal consistency, and these judgments were supported by other statistics, such as Cronbach's alpha and mean interitem correlations.

Although the average interitem correlation and APD produce similar results, we view the APD as having two distinct advantages over the former method: (a) It is an intuitive measure of the difference between scores on items, and (b) it penalizes large mean differences across items. To expand on these points, the APD provides the researcher with direct information regarding differences in the scores across items. Other methods of internal consistency do not provide this information except indirectly. In regards to mean differences between items, it is expected that subjects will score similarly across items, and the APD is reduced if certain items produce disparate raw scores. This is not necessarily the case for interitem correlations, which may be high even when the items have disparate raw scores.

Until recently, social scientists erred on the side of redundancy in scale construction as longer scales were more likely to achieve satisfactory alpha levels. Yet, as demonstrated in this article, Cronbach's alpha will have higher values with increasing numbers of items, even if the interitem correlations are small. In the present study, we were also able to demonstrate that Cronbach's alpha could be high, given enough items, even when the correlation between items was modest. The move to shorter, less-redundant measures is warranted particularly because short scales have been shown to be as valid as longer measures (Burisch, 1984a, 1997). Indeed, as the population increasingly shifts toward an older demographic and the number of Internet-based studies increases, shorter scales for many psychological variables will be required. Where longer scales are necessary, as is the case for broad constructs, we would suggest the use of statistics of internal consistency such as the APD method, which will give a true measure of the agreement or lack of agreement between items. We do not advocate replacing established methods such as interitem correlations and Cronbach's alpha but rather supplementing them with the APD method. All of these measures provide different and potentially useful information about scale consistency and reliability, and it is difficult to imagine a situation in which additional information of this sort is not beneficial to researchers.

# References

Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality, 54*, 107-148.

Burisch, M. (1984a). Approaches to personality inventory construction: A comparison of merits. *American Psychologist, 39*, 214-227.

Burisch, M. (1984b). You don't always get what you pay for: Measuring depression with short and simple versus long and sophisticated scales. *Journal of Research in Personality, 18*, 81-98.

Burisch, M. (1997). Test length and validity revisited. *European Journal of Personality, 11*, 303-315.

Clark, L. A., & Watson, D. (2003). Constructing validity: Basic issues in objective scale development. In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research* (pp. 207-231). Washington, DC: American Psychological Association.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and application. *Journal of Applied Psychology, 78*, 98-104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*, 391-418.

Melchior, L. A., Huba, G. J., Brown, V. B., & Reback, C. J. (1993). A short depression index for women. *Educational and Psychological Measurement, 53*, 1117-1125.

Murphy, K. R., & Davidshofer, C. O. (1998). *Psychological testing: Principles and applications.* Upper Saddle River, NJ: Prentice-Hall.

Radloff, L. S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385-401.

Shafer, A. B. (2006). Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. *Journal of Clinical Psychology, 62*, 123-146.

Shrout, P. E., & Yager, T. J. (1989). Reliability and validity of screening scales: Effect of reducing scale length. *Journal of Clinical Epidemiology, 42*, 69-78.

Streiner, D. L. (2003a). Being inconsistent about consistency: When coefficient alpha does and doesn't matter. *Journal of Personality Assessment, 80*, 217-222.

Streiner, D. L. (2003b). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*, 99-103.