Disclosing Empirical Relationships among Coefficient Alpha, Item *p*-value, and Point-Biserial Correlation

Yi-Fang Wu, PhD

To inform test developers on reliable test form construction with high internal consistency coefficients, this study investigated empirical relationships among coefficient alpha, item *p*-value (as an index of item difficulty), and point-biserial correlation (as an index of item discrimination). The increase or decrease of coefficient alpha was evaluated, especially for tests that varied in their average item *p*-values and average point-biserial correlations. This brief includes item- and test-level statistics that are closely related to coefficient alpha, describes the empirical test data and the resampling procedures used to obtain estimates of various statistics, and makes suggestions to practitioners for building highly reliable test forms.

Coefficient Alpha and Item-Level Statistics

For a single test administration, the reliability estimates of a test based on the internal structure of test/response data can be calculated using the inter-item covariances. The most wildly used reliability estimate to evaluate such internal consistency is coefficient alpha (Cronbach, 1951), which usually takes the following form:

$$\hat{\alpha} = (\frac{k}{k-1})(1 - \frac{\sum_{i=1}^{k} s_i^2}{s_x^2}) = (\frac{k}{k-1})(\frac{\sum_{i=1}^{k}\sum_{j\neq i}^{k} s_{ij}}{s_x^2}),$$

where *k* is the number of split parts (e.g., test items), s_x^2 is the sample variance of the total test score, s_i^2 is the sample variance of the *i*th part (e.g., *i*th item), and s_{ij} is the sample covariance between part *i* and part *j* (e.g., item *i* and item *j*). Several factors affect the reliability estimates of test scores, such as the length of the test, the degree of the homogeneity among items, the difficulty and discriminating power of items, the group variability, the guessing/chance errors, the administration instructions, and the scoring procedures.

To construct test forms with high reliability, psychometricians investigate the relationships between reliability and factors that test developers can have some control of including, but not limited to, test length, degree of the homogeneity among items, item difficulty, and item discrimination. It is almost inarguable that when a test has more items, the test score reliability is higher. Logically, the more samples of items of a given area of knowledge we include in a test, the more reliable the test will be. As proof, Ebel (1972) showed that lengthening a test increases the true score variance more rapidly than the error variance, which results in higher reliability estimates because, under classical test theory[1], test score reliability can be defined as

the ratio of the true score variance to the total/observed score variance (i.e., the sum of the true score variance and error variance). As for item-level statistics (e.g., item difficulty and discrimination), a collective set of items which can result in a broader spread of test scores will yield higher reliability estimates.

For test developers to better understand the relationships among test score reliability, item difficulty, and item discrimination, the remainder of this brief summarizes the use of statistical resampling procedures based on real student response data. It shows how the resulted synthetic test datasets were analyzed to demonstrate the empirical relationships among coefficient alpha, item *p*-values (as an index of item difficulty), point-biserial correlation (as an index of item discrimination), the test score standard deviation of the student sample (which indicates group variability), and inter-item correlation. Note that the sample standard deviation and inter-item correlation are explicitly shown in the formula of coefficient alpha (i.e., s_x^2 and $\sum_{i=1}^{k}\sum_{j\neq i}^{k} s_{ij}$).

Empirical Test Data and Resampling

To include test length as a factor in the current investigation, four fixed-length tests of multiple-choice items were used. For each test, one form was chosen and then the state samples were randomly selected from the same year's student database to perform resampling. The resampling procedures used to create the synthetic test datasets for computing statistics of interest were a combination of the group jackknife method (Efron, 1982; e.g., Haberman, Lee, & Qian, 2009; Wu & Li, 2012) and the bootstrap method (Efron, 1982).

As shown in Table 1, the original test lengths of the four tests were 30, 40, 50, and 60 items. The sample sizes of the state students' 0/1 response data are listed in the last column of the table. To demonstrate the steps for resampling and the creation of the synthetic datasets based on the empirical response data, the 30-item ACT EXPLORE® math test was used as an example in the following paragraphs.

First, the item *p*-values and the point-biserial corrections were computed. The items were then ordered from the easiest to the hardest. After ordering the items based on their *p*-values, to create the first synthetic dataset, the first five easiest items were removed from the original test. To create the second synthetic dataset, the second easiest item and the subsequent four easiest items (i.e., the second to sixth easiest items) were removed from the original test. The third easiest item and the subsequent four easiest items were then removed to create the third synthetic dataset. Following this pattern, 26 synthetic datasets were obtained from the original 30-item test, each of which included 25 items. The average item *p*-values of these synthetic tests increased (i.e., the average difficulty became easier) as a result of the resampling procedure.

The procedure described above began with ordering the items from the easiest to the hardest. It was then replicated but by ordering the items from the least discriminating to the most discriminating. That is, the first five least discriminating items were removed from the original test

to create a synthetic dataset; the second to sixth least discriminating items were then removed to create another dataset; and so on. From the original 30-item test, another 26 synthetic datasets (with decreasing average discriminating power) were created, and thus a total of 26*2 = 52 datasets, each with the 0/1 responses for *N* = 3,561, were available for analysis.

Table 1. Original Test Length, Synthetic Test Length, Number of Synthetic Datasets, and Sample Size

Test	Original Test Length	Size of Removed Item Block	Synthetic Test Length	Number of Synthetic Datasets	Sample Size Per Dataset
ACT EXPLORE® math	30	5	25	26*2 = 52	3,561
The ACT® reading	40	5	35	36*2 = 72	43,589
ACT PLAN® reading	50	5	45	46*2 = 92	244
The ACT math	60	5	55	56*2 = 112	43,626

Note: Sample size per dataset depended on the size of available data from a randomly selected state.

In addition to the sample size of the response data, Table 1 shows the resulting test length and the number of the synthetic datasets for other original test lengths under investigation. Note that five items were chosen to be the size of the removed item block because the resulting number of items (i.e., synthetic test length)—25, 35, 45, and 55—were reasonable and commonly-seen multiple-choice test lengths. The idea of removing items in blocks was borrowed from the grouped jackknife method, while the idea of selecting the same items to create different synthetic datasets came from the bootstrap method (e.g., sampling with replacement). Practicing the above approach resulted in as many datasets as possible based on the real response data, wherein different datasets had either increasing average item *p*-values or increasing item discrimination power. The resulting number of synthetic datasets is related to the original test length and the size of the removed item block[2], as illustrated at the end of this brief.

Analysis and Discussion

For each synthetic dataset, the following sample statistics were computed: the coefficient alpha, the raw score standard deviation (SD) (sample SD), the mean and SD of the item *p*-values (*p*-value mean and *p*-value SD), the mean and SD of the point-biserial correlations (r(pb) mean and r(pb) SD), and lastly, the mean of the inter-item correlations (*r(ij)* mean, where *r(ij)* represents the raw score correlation between the *i*th and *j*th items and $i \neq j$). Results are summarized in scatter plots in Figure 1.

Test Length

The horizontal axes in the scatter plots represent coefficient alpha (α); the vertical axis in each scatter plot represents each sample statistic of interest. The scatter plots in the same column are for a given test length (TL). As for the effect of test length on alpha, alpha increased when there were more items in a test; this fact could also be inferred from the range of the alpha

estimates across columns. For example, the upper bound of TL = 25, TL = 35, TL = 45, and TL = 55 were about .84, .88, .90, and .92, respectively.

Group/Sample Variability

As implied by the coefficient alpha formula and as shown in the first row of Figure 1, alpha was a function of group variability (i.e., sample SD)—the more variable the test-taker sample was, the higher the alpha estimate was. It was also noted that the sample variations of the curriculum-based, multiple-choice achievement tests were generally quite stable. For example, for a 35-item test, the sample SDs were around 7; for a 55-item test, the sample SDs were between 9 and 10. This sample SD information, along with the item difficulty and discrimination, could be helpful to inform desirable alpha values. Based on the understanding of the test-taker group variability, test developers can construct forms by adjusting the average item difficulty and/or item discrimination to reach a predefined and acceptable coefficient alpha value. This is elaborated further in the next section.

Item Difficulty, Item Discrimination, and Inter-Item Correlation

The second row in Figure 1 shows the relationship between coefficient alpha and average item difficulty (item *p*-value mean); the fourth row shows the relationship between coefficient alpha and average item discrimination (point-biserial correlation mean). For shorter tests (e.g., 25 or 35 items), the collinearity between the average item *p*-values and the coefficient alpha values was relatively more prominent compared to longer tests. Based on the aforementioned observations, increasing the average item difficulty for shorter tests boosted alpha.

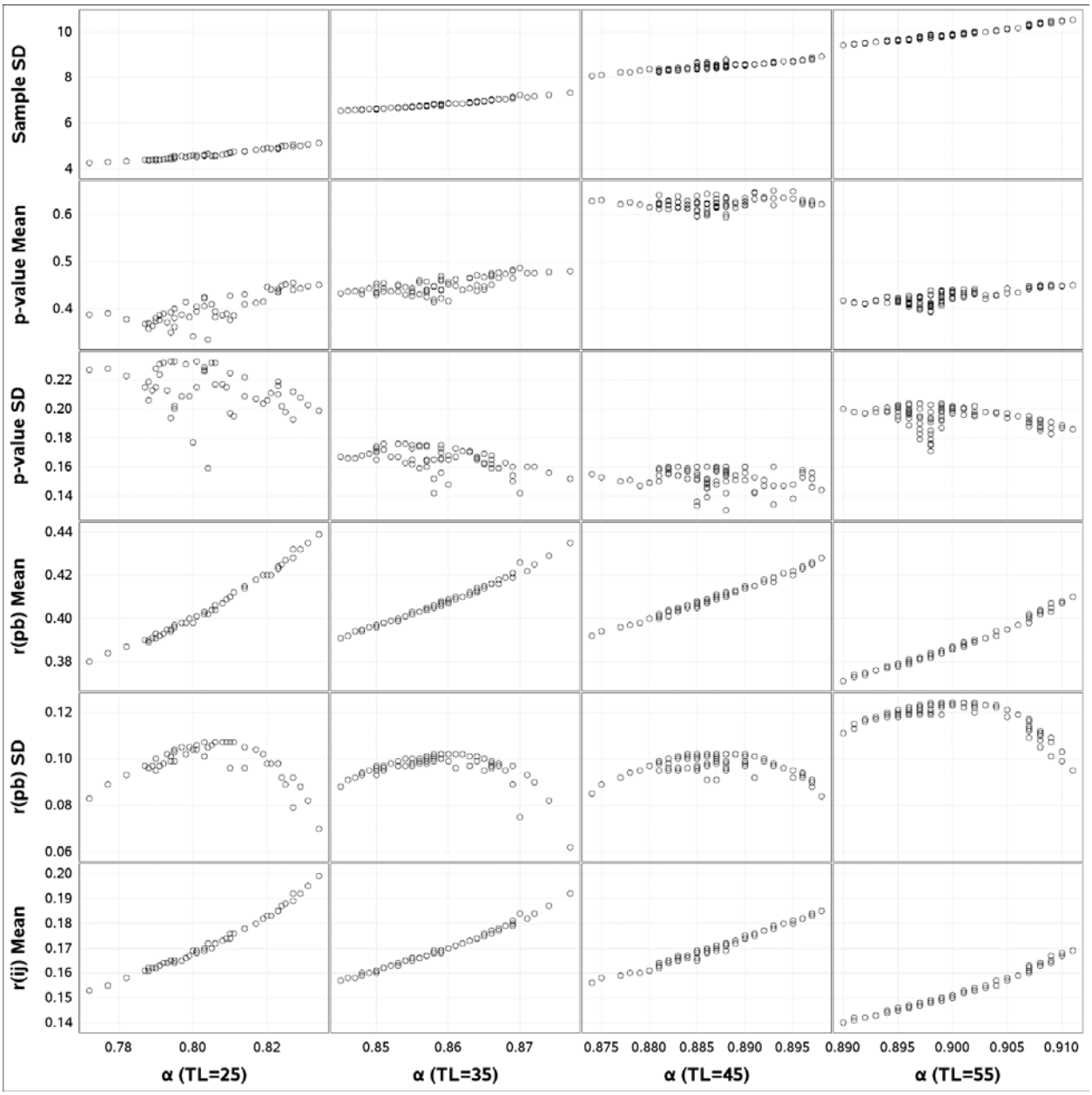
As expected, alpha became higher as the average point-biserial correlation increased. It was also found that alpha had an approximately linear relationship with the average point-biserial correlation. Consequently, one could expect a polynomial relationship between alpha and the variability of the point-biserial correlations as shown in the fifth row, which indicated that a test covering the widest range of item discriminations did not guarantee the highest alpha.

Regarding the relationship of coefficient alpha with inter-item dependency (i.e., r(ij) mean), alpha became higher with higher average inter-item correlation as expected. Higher inter-item correlations also imply stronger statistical dependence among test items and thus directly contribute to higher alpha estimates.

Lastly, the results in Figure 1 are informative because they suggest that test developers can select items of certain item difficulty and discrimination ranges to attain a predefined alpha value. For example, assume that we have a 25-item multiple-choice test that is moderately difficult with item *p*-values mostly ranging from .4 to .6. Also assume that the sample SD of this test is roughly 5 on the raw score scale. In this situation, we can anticipate that alpha can go above .8 if the average point-biserial correlations can be increased by .02.

In summary, the empirical relationship between coefficient alpha and *p*-values implies that to develop a test with a high alpha estimate (e.g., .8 or above), one should increase the average item difficulty to some extent, especially for shorter tests. More importantly, selecting items with good discriminating power can increase sample variability, which will lead to high alpha estimates regardless of test length. A strong collinearity in item discrimination and its relationships with alpha are empirically shown in Figure 1. While the spread of item *p*-values is often the main concern of test developers when constructing test forms, higher alphas are in fact more obtainable by relatively high average point-biserial correlations when test items have moderate to moderately high average item *p*-values.

Figure 1. Scatter Plots of Estimated Coefficient Alpha vs. Sample SD, *p*-value Mean, *p*-value SD, Point-Biserial Correlation Mean, Point-Biserial Correlation SD, and Inter-Item Correlation Mean for Various Test Lengths[3]



Notes: On the x-axis, α denotes coefficient alpha; TL is test length. On the y-axis, r(pb) denotes point-biserial correlation; r(ij) denotes inter-item correlation.

Notes

1. The interested reader may refer to Table 1 in Ebel (1972) for further explanation.
2. The following table provides an illustration of creating synthetic datasets when there were ten items in the original test and two items were dropped each time after ordering the items by item *p*-value.

Item Order	Set1	Set2	Set3	Set4	Set5	Set6	Set7	Set8	Set9
01/Easiest	X								
02	X	X							
03		X	X						
04			X	X					
05				X	X				
06					X	X			
07						X	X		
08							X	X	
09								X	X
10/Hardest									X

   The resulting number of synthetic datasets is 10 − 2 + 1 = 9; that is, the original test length (10) minus the size of the item block being removed (2) plus one.
3. For TL = 45, the estimates of the average item *p*-values were higher, and the SD of the *p*-values were smaller compared to the other test length conditions. For TL = 45, the original sample size for resampling was smaller and the samples could be a relatively capable sample (i.e., of better performing students) to produce the estimates. Nevertheless, the patterns to explain the relationship between coefficient alpha and item discrimination (as well as inter-item correlation) were still apparent and consistent with the other test length conditions.

References

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. doi:10.1007/BF02310555

Ebel, R. L. (1972). Why is a longer test usually a more reliable test? *Educational and Psychological Measurement, 32*(2), 249–253. doi:10.1177/001316447203200202

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Haberman, S., Lee, Y.-H., & Qian, J. (2009). Jackknifing techniques for evaluation of equating accuracy (ETS RR-09-39). Princeton, NJ: Educational Testing Service.

Wu, Y.-F. & Li, D. (2012, April). The standard errors of IRT true score equating based on internal and external anchors. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.

About the Author

Yi-Fang Wu, PhD is Senior Psychometrician in Psychometric Research. Her research interests include topics in psychometrics and educational measurement and statistics.

Acknowledgements

The author thanks Dr. NooRee Huh and Dr. Han Yi Kim for their helpful reviews of this brief. The author also thanks Ross Wagenhofer for his helpful editorial reviews and comments.

ABOUT ACT

ACT is a mission-driven, nonprofit organization dedicated to helping people achieve education and workplace success. Headquartered in Iowa City, Iowa, ACT is trusted as a national leader in college and career readiness, providing high-quality assessments grounded in over 60 years of research. ACT offers a uniquely integrated set of solutions designed to provide personalized insights that help individuals succeed from elementary school through career. Visit us online at www.act.org.