Coefficient Alpha: An Engineer's Interpretation of Test Reliability

KIRK ALLEN
Department of Engineering Education
Purdue University

TERI REED-RHOADS
College of Engineering and
Department of Engineering Education
Purdue University

ROBERT A. TERRY
Department of Psychology
The University of Oklahoma

TERI J. MURPHY
Department of Mathematics
The University of Oklahoma

ANDREA D. STONE
Department of Mathematics
The University of Oklahoma

ABSTRACT

Reliability is a fundamental concept of test construction. The most common measure of reliability, coefficient alpha, is frequently used without an understanding of its behavior. This article contributes to the understanding of test reliability by demonstrating that questions which lower reliability are inconsistent with the bulk of the test, being prone to test-taking tricks and guessing. These qualitative characteristics, obtained from focus groups, provide possible causes of lower reliability such as poorly written questions (e.g., the correct answer looks different from the incorrect answers), questions where students must guess (e.g., the topic is too advanced), and questions where recalling a definition is crucial. Quantitative findings confirm that questions lower reliability when students who answer correctly have lower overall scores than students who answer incorrectly. This phenomenon is quantified by the "gap" between these students' overall scores, which is shown to be highly correlated with other item metrics. An increasing number of concept inventory tests are being developed to assess student learning in engineering. Scores and student comments from the Statistics Concept Inventory are used to make these judgments.

Keywords: concept inventory, statistics, test reliability

I. INTRODUCTION

Assessment plays a prominent role in the maturing field of Engineering Education (Special Report, *Journal of Engineering Education*, 2006). The focus on assessment is in part driven by ABET engineering accreditation changes, which focus on outcomes rather than fulfilling course requirements. In addition, emphasis is placed on assessment for a better understanding of learning systems and mechanisms, as well as diversity and inclusiveness (Special Report, *Journal of Engineering Education*, 2006). Heeding calls for the field to become more inter-disciplinary (Fortenberry, 2006) and rigorous (Streveler and Smith, 2006), an exposition on standard methods in educational research is apropos, benefiting both developers and users of assessment tools. This article sheds light on test reliability in a practical manner such that it can be understood and applied by those with little knowledge of psychometrics.

The concept of reliability is a cornerstone of test development and analysis. Nonetheless, a review of 220 articles from 22 educational research journals found that 46 percent of articles did not report any reliability evidence (Whittington, 1998). Reliability, analogous to precision, is inversely related to measurement error, which results from inconsistent response patterns and limits the generalizability of results from sample to population. Measurement error is a random phenomenon, as opposed to systematic error which can be controlled once identified (American Educational Research Association, 1999). Moreover, reliability is a property of the scores and not the test itself. Thus, it is advocated that researchers report reliability estimates for the data being analyzed rather than resort to previously reported estimates (Wilkinson, 1999).

There are several methods for assessing reliability, described in many textbooks on psychological testing (e.g., Hogan, 2003; Kline, 2005). The most commonly cited are test-retest, in which answer consistency is measured from one administration to the next; alternative forms, where subjects take two separate tests which are nearly identical in every aspect; and internal consistency, which measures the extent to which the test questions are highly correlated with each other. This article investigates the last one in that list, internal consistency, as measured by coefficient alpha.

There have been several attempts in recent years to shed light on coefficient alpha (Cortina, 1993; Streiner, 2003). While informative, these are often written from theoretical viewpoints. Little work has been presented that details how alpha behaves for real test data on a question-by-question basis. This article presents the background information for coefficient alpha and presents data from an instrument, the Statistics Concept Inventory (Statistics Concept Inventory, 2007; Stone et al., 2003), as an illustration of how alpha behaves.

II. RELIABILITY BACKGROUND

Among the first to quantify test reliability were Kuder and Richardson (1937). They comment that a reliability coefficient based on test-retest will often result in a reliability that is spuriously high due to material remembered on the second administration.

Further, increasing the time between administrations is impractical because subjects may gain knowledge in the interim. Therefore, they shifted their focus to internal consistency as a measure of reliability.

Kuder and Richardson focus on the concept of a split-half coefficient, in which the test is split in two parts and a correlation is calculated between those two parts. A test of length k has $({}_{k}C_{k/2}) \div 2$ ways to be split in two; the term is divided by 2 to remove redundancy. For a test with 10 items, there are 126 combinations. Each split-half will result in a different reliability. There are potential problems with deciding how to split the test and which is the most appropriate split. Depending on the split, the calculated reliability may be higher or lower than the "true" reliability. To overcome these problems, the authors derive several equations which arrive at unique values of the reliability coefficient.

The most often-cited result from Kuder and Richardson (KR) (1937) is their equation 20, sometimes called KR-20 and later dubbed "alpha" by Cronbach (1951). The KR-20 is so commonly used because it assumes dichotomous scoring (i.e., 0 for incorrect, 1 for correct), which is how most achievement tests are scored. The formula is given below in equation (1). The expression $\sum p_i q_i$ can be substituted in place of $k\,\overline{pq}$ to give a more general result.

$$\alpha = \frac{k}{k-1}\left(\frac{\sigma_t^2 - k\,\overline{pq}}{\sigma_t^2}\right) \tag{1}$$

where: α is the reliability of the test (denoted r_{tt} by Kuder and Richardson)
k is the number of questions (often referred to as items) on the test
σ_t^2 is the total score variance for the test
p_i is the proportion of students who answer item i correctly
q_i is the proportion of students who answer item i incorrectly
$\overline{pq}$ is the average p multiplied by the average q for the test, equivalent to assuming p is constant across all items.

Equation (1) was generalized by Cronbach (1951) as shown in equation (2), which allows any equally-weighted scoring method for test items, including common Likert-style scales. Although commonly referred to as Cronbach's alpha or coefficient alpha, this expression was derived independently by Guttman (1945) as well and is sometimes referred to as Guttman-Cronbach alpha in psychometric literature.

$$\alpha = \frac{k}{k-1}\left(\frac{\sigma_t^2 - \sum V_i}{\sigma_t^2}\right) = \frac{k}{k-1}\left(1 - \frac{\sum V_i}{\sigma_t^2}\right) \tag{2}$$

where: α is Cronbach's coefficient alpha (same meaning as r_{tt})
k is the number of questions (or items) on the test
$\sum V_i$ is the sum of the individual item variances
σ_t^2 is the total score variance for the test (denoted as V_t by Cronbach)

For dichotomously scored items, V_i reduces to $p_i q_i$ and the KR-20 equation is obtained. The derivation of this relationship is given in the Appendix. For Likert-style items seen in attitudinal surveys, the variance retains the standard variance formula available in statistics textbooks.

Statistical packages such as SPSS™ and SAS™ report "alpha if item deleted" which shows how coefficient alpha would change if a certain question were omitted. A "good" question will have a lower "alpha if item deleted" because deleting that question will lower the overall alpha. Each question's effect on alpha is measured by the "change in alpha," found by subtracting "alpha if item deleted" from the overall alpha, shown in equation (3). A "good" question will have a lower "alpha if item deleted" because removing that question would lower alpha; thus, the change in alpha will be positive.

$$\textit{Change in Alpha} = \textit{Overall Alpha} - \textit{Alpha if item deleted} \tag{3}$$

The simplest way to explain how a question will have a negative effect on alpha (i.e., a higher "alpha if item deleted") is to consider Cronbach's definition of alpha (equation 2). A "bad" question will lower the overall test variance (σ_t^2). This happens when students with low overall scores perform better on a question than students with high overall scores. This "squishes" the class together (smaller variance). When σ_t^2 decreases, the ratio $\sum V_i / \sigma_t^2$ increases. This ratio is then subtracted from 1, which lowers alpha.

For each item, the effect on total score variance is quantified by subtracting the average total score of those who answer the item incorrectly from the average total score of those who answer correctly. We call this value the "gap"; it is defined symbolically in equation (4).

$$Gap_i = \bar{x}_{Correct} - \bar{x}_{Incorrect} \tag{4}$$

where: Gap_i quantifies item i's effect on total variance (σ_t^2)
$\bar{x}$ refers to the mean total exam score
subscripts *Correct* and *Incorrect* refer to those students who answer item i correct and incorrect, respectively

The average inter-item correlation is also considered a measure of a question's reliability (Cortina, 1993). The inter-item correlation is the Pearson correlation coefficient (r) computed with the 0-1 scores for a pair of items. The average inter-item correlation is each item's average inter-item correlation with the other k-1 items. If a question has negative or low positive (close to zero) inter-item correlations, it does not "fit" with the rest of the questions. This will be shown to relate which students answer a question correctly.

It is even possible for the overall alpha to be negative. For example, if every student received the same total score on a test, σ_t^2 would be zero. As the test variance approaches zero, the ratio $\sum V_i / \sigma_t^2$ approaches infinity. When the calculation $1 - \sum V_i / \sigma_t^2$ is performed, alpha will approach negative infinity. (Note: $\sum V_i$ would only be zero also if every question were answered the same, either correctly or incorrectly, by every student.)

III. About The Data

A. The Statistics Concepts Inventory

The illustrative data were obtained using the Statistics Concepts Inventory (SCI). The SCI is a multiple choice instrument to assess student understanding of fundamental statistics concepts. It is part of a larger interest to develop such instruments in a range of engineering fields (Evans et al., 2003). The concept inventory movement was spurred by the development and successful implementation of the Force Concept Inventory (FCI) (Halloun and Hestenes, 1985; Hestenes, Wells, and Swackhamer, 1992). The FCI was developed as a pre-post- test to identify student misconceptions of Newtonian force when entering a physics course and check for gains upon completing the course. After many rounds of testing, it was discovered that students gain the most conceptual knowledge in interactive engagement courses, as opposed to traditional lectures (Hake, 1998).

The SCI was piloted during the Fall 2002 semester at the University of Oklahoma (OU) (Stone et al., 2003). The pilot version was constructed by first identifying topics to include using a faculty survey. Questions and multiple-choice answers were written by searching statistics textbooks and educational literature for examples which covered these topics. The researchers also used personal experience to develop additional questions.

The revision process included focus groups, analysis of correct and incorrect answer distributions, and expert opinions. Several new questions were generated from these processes. The data in this article were gathered from the second version of the SCI, which had 33 questions and was administered during summer 2003. Two sample questions from the test are shown in Figure 1.

B. Data Collection

The data in this study were gathered from four sources: (1) a statistics class in the College of Engineering at OU, with students having a background of at least three semesters of Calculus; (2) a statistics class in the Department of Mathematics at OU, primarily consisting of engineering students with a similar background as (1); (3) two groups of undergraduates participating in a summer research program in OU's School of Industrial Engineering, with backgrounds ranging from no statistics experience to several semesters of statistics; and (4) a statistics class in the College of Engineering at a four-year university outside OU, with a similar background to (1) and (2). An approved Institutional Review Board (IRB) protocol was utilized with all sources of data. The number of students in each group ranged from 14 to 39. Groups (1) and (2) took the instrument as a pre- and post-test. The data in this article are from the post-test.

IV. The Behavior of Alpha

A. A Macro View of Alpha

To understand the behavior of coefficient alpha, the components of Cronbach's formula (2) need to be analyzed first. Table 1 shows how alpha and its components vary across the four groups used in this study.

1. The following are temperatures for a week in August: 94, 93, 98, 101, 98, 96, and 93. By how much could the highest temperature increase without changing the median?

   a) Increase by 8°
   b) Increase by 2°
   c) It can increase by any amount (*correct*)
   d) It cannot increase without changing the median

2. A researcher performs a t-test to test the following hypotheses:

   $$H_0 : \mu \le \mu_0$$
   $$H_1 : \mu > \mu_0$$

   He rejects the null hypothesis and reports a p-value of 0.10. Which of the following must be correct?

   a) The test statistic fell within the rejection region at the significance level
   b) The power of the test statistic used was 90%
   c) Assuming the null is true, there is a 10% possibility that the observed value is due to chance (*correct*)
   d) The probability that the null hypothesis is not true is 0.10

Figure 1. Two items from the Statistics Concept Inventory.

Group	*n* (students)	*k* (items)	Overall α	σ_t^2	$\sum V_i$	Range*
(1) OU Math	14	33	0.8587	38.06	6.37	21
(2) OU Engr	24	33	0.8100	31.16	6.68	15
(3) OU REU	27	33	0.5983	14.99	6.29	16
(4) Outside	39	33	0.5781	14.69	6.46	17

*Range is the maximum score minus the minimum score.

Table 1. Macro view of alpha.

With the sum of individual variances ($\sum V_i$) approximately constant over the four groups, total test variance (σ_t^2) is seen as the most important component of alpha. For these groups, alpha varies inversely with the number of students for this data, but this is a coincidence when viewing the magnitude of the changes (going down the chart, alpha decreases by 0.05 then 0.21, but n increases by 10 then just 3). This pattern is not seen on data from subsequent administrations (refer to Table 4 near the end for more data). The range is included as a simplified estimate of variance, but it lacks explanatory power for alpha aside from the highest alpha having the highest range.

B. A Micro View of Alpha

The data used in this section are for the summer research students (group (3) in section III.B). They were selected for further illustration because a focus group was conducted with over half of these students, which allowed additional insight to be gained about why questions may be performing poorly in terms of reliability. Because the data for the other three groups are similar, their presentation would not add to the discussion or change the result except to reinforce the generalities of the data from group (3).

The most direct way to explain alpha is to first compare the change in alpha to the average inter-item correlation for each question as shown in Figure 2. Because both axes represent measures of a test's reliability, a strong correlation is expected. It is also important to show why certain questions have poor correlations. This is presented in terms of average inter-item correlation vs. gap and change in alpha vs. gap, shown in Figures 3 and 4, respectively. Gap is calculated using total score rather than percentage. Using percentage will change the scale of the x-axis, but the correlation will not change. On this 33 question test, one point of gap corresponds to 3 percent.

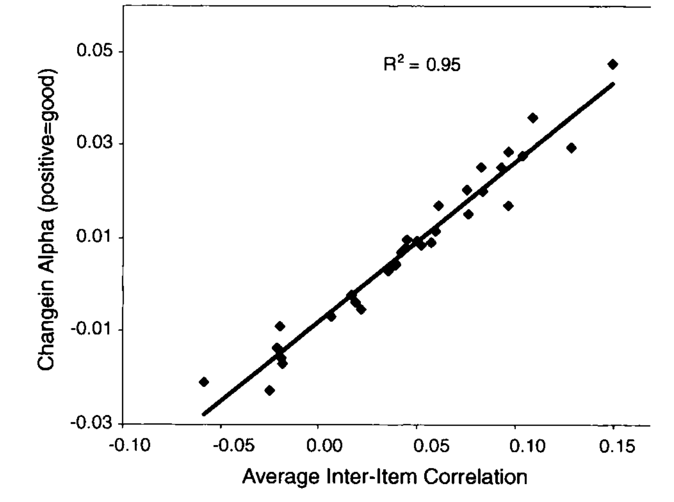

Figure 2. Relationship between change in alpha and average inter-item correlation.

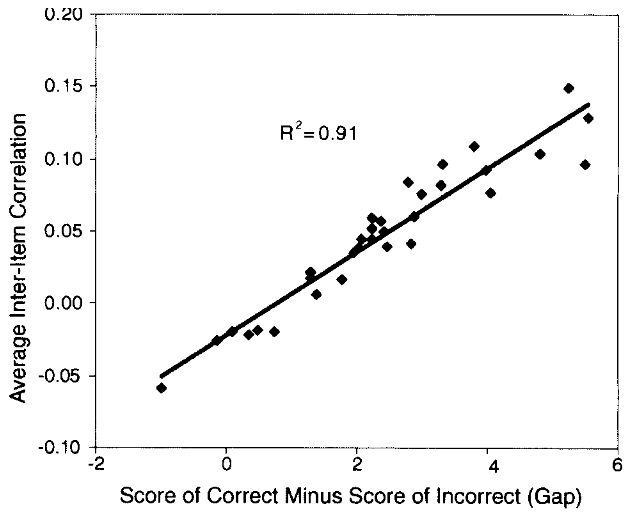

Figure 3. Relationship between average inter-item correlation and gap.

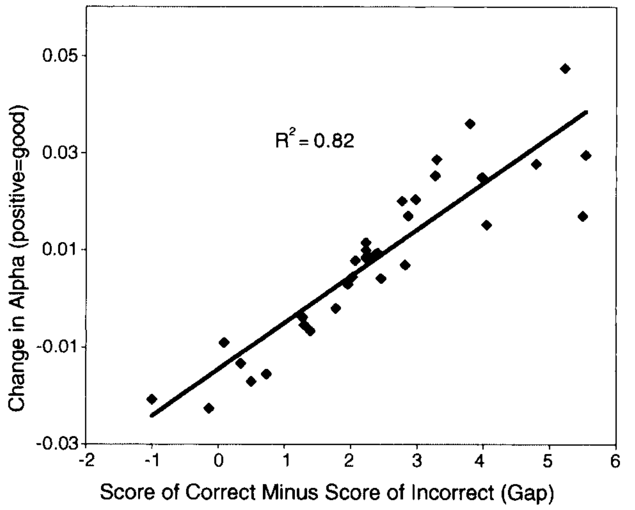

Figure 4. Relationship between change in alpha and gap.

These plots continue to show strong relationships between the variables. This matches the theoretical explanation presented in section II. Specifically, questions with a low or negative "gap" are those which lower the variance of the overall test score. Low total variance has been shown to be both mathematically and empirically the crucial component of coefficient alpha. Combined with what has been presented about the mathematical behavior of alpha, these graphs imply that a question's average inter-item correlation and, more directly, a question's "gap" are plausible causes of a question behaving poorly as measured by "alpha if item deleted."

Another measure to quantify the effectiveness of a question is the discriminatory index (Kelley, 1939). This statistic is calculated by comparing the average score on the item of the top quartile students to the bottom quartile students, where quartiles are defined by total exam score. (Example: 4th Q 60 percent of students correct, 1st Q 25 percent of students correct → For this item, Discriminatory index = 0.60 − 0.25 = 0.35) This statistic can also be shown to correlate highly with alpha if item deleted (Figure 5).

For this group of students, discriminatory index does not correlate as strongly as the "gap." However, for two other groups analyzed, change in alpha correlates more strongly with discriminatory index than with "gap." The lack of a consistent pattern limits further conclusions. Table 2 shows the correlations of alpha with the various other measures presented previously for three courses.

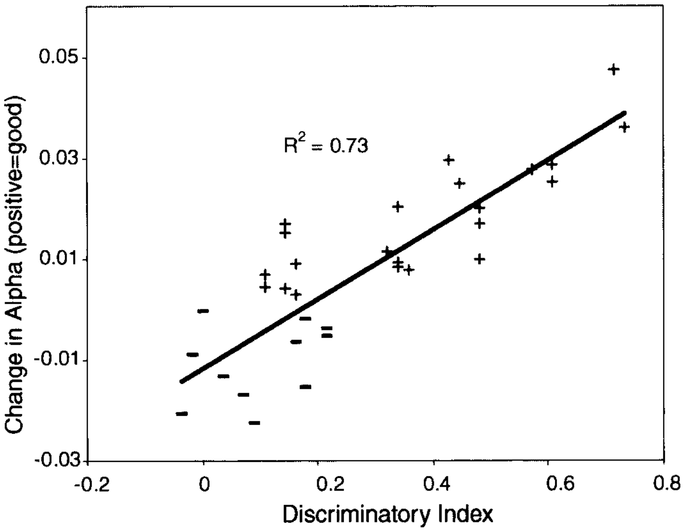

Figure 5. Relationship between change in alpha and discriminatory index.

Course	$\bar{r}$	Gap	Disc. Index	n
OU REU	0.977	0.905	0.854	27
OU Engr	0.991	0.877	0.918	24
OU Math	0.973	0.889	0.935	14
Outside	0.982	0.970	0.905	38

Key : $\bar{r}$ average inter-item correlation
n students in each class

Table 2. Correlation of alpha-if-deleted with various item metrics.

C. Explanation Using Comments from Focus Groups

Over half of the students who took the test attended a focus group where questions were discussed in detail. This allowed more scrutiny on a question-by-question basis. Using the comments of the focus groups, qualitative evidence can be obtained about what makes a question "bad" in terms of alpha. Table 3 presents the 10 worst questions in terms of "alpha if item deleted" (marked by a minus sign in Figure 5).

By evaluating these questions in such a manner, it is important to remember that alpha is a property of this set of scores and not of the test itself. The overall alpha and the "bad" questions will vary from class to class. This could be partly due to chance but also could indicate that one professor covered a topic whereas another did not or that topics were covered in different manners with varying results. These variations bring to light the difficulty of defining a target population and finding a representative, consistent sample. While the SCI has a target of statistical beginners, specifically those who are engineering majors, the varied backgrounds and classroom exposure make finding the precise target audience (i.e., those who have been exposed to all concepts) impractical.

The comments in Table 3 indicate that questions on which students guessed had a negative impact on alpha. This makes sense in light of the other data presented because one expects a question on which students guess to have a "gap" near zero. It is also likely that these questions measure some attribute other than statistical reasoning, such as test-taking ability or memory. This is plausible when compared with the effect that negatively correlated items have on alpha. These items do not appear to measure the same construct. When this happens, inter-correlations among items tend to be smaller. In other words, these questions are not internally consistent with the rest of the test.

D. The Big Picture

The reliability analysis is conducted after each round of test administration and used to guide revisions of the SCI. Table 4 shows the pre-test and post-test coefficient alpha for the combined course data from each semester. Moving down the chart, the test shows an increasing trend on the post-test, indicating the revisions are successful. The pre-test consistently has an alpha in the 0.69 range for the past four semesters. Moving across the chart for each semester, there is an increase in alpha from pre-test to post-test on three of five occasions. One expects a pre-test to be subject to more guessing and test-taking tricks than a post-test, which would explain the lower pre-test alphas. However, lack of a consistent pattern (i.e., post-test having lower alpha on two of five administrations) suggests there are additional sources which lower total test variance, such as student knowledge becoming more standardized as a result of instruction.

V. Conclusion

This paper provides insight into the behavior of one measure of reliability, coefficient alpha, from a theoretical vantage and extends this to data from a real test. High variance of scores is the key component needed to attain a high coefficient alpha. Focus groups conducted with students after taking the test corroborate the reliability results by showing that there are several possible causes for questions that adversely affect alpha-guessing, the use of test-taking skills, or when recalling a definition is necessary. In general, these "bad" questions do not conform to the material on the test and have high "alpha if item deleted" values, which are highly correlated with average inter-item correlations and the discriminatory index.

Coefficient alpha can play a prominent role as a tool to aid in the revision of questions and thus improving the overall reliability of a test. The derived metric "alpha if item deleted" indicates which

Rank	Question Topic	Possible problem	Student comments	Change in Alpha
33	Meaning of *p*-value	Too many symbols; definition recall	Most students guessed	−0.0227
32	Meaning of *p*-value	Definition recall; one answer nearly correct but wrong by one word	Several students guessed; strong distracter misled others	−0.0208
31	68-95-99 rule for normal	Requires remembering a rule	People who got it correct say they just remember the rule	−0.0170
30	Parent distribution of a sample	*No useful comments*	n/a	−0.0157
29	Calculating standard deviation	Depends on attention to detail	They think it is easy as long as you read carefully	−0.0134
28	*t*-distribution	*No useful comments*	n/a	−0.0009
27	Sample vs. population	Poorly written: incorrect choices looked different from correct choice	One student chose the correct answer for incorrect reasons	−0.0067
26	Design of experiment	Advanced topic	Most students guessed	−0.0053
25	Variability of a histogram	Students do not understand the graphs	Most students discussed lack of understanding	−0.0038
24	Central tendency	Term "central tendency" possibly confusing	One mentioned being confused by the term	−0.0022

Note: These questions ranked highest on alpha if deleted, therefore are considered the "worst" questions relative to the remaining questions.

Table 3. Ten worst questions in terms of "alpha if item deleted."

		Post-Test					
Semester	Pre-Test Alpha	Alpha	*n* (students)	*k* (items)	σ_t^2	$\sum V_i$	Range
Fall 2002	n/a	0.5957	174	32	15.11	6.39	22
Summer 2003	0.7434	0.6965	66	33	18.91	6.64	19
Fall 2003	0.6915	0.7031	241	34	18.10	6.63	21
Spring 2004	0.6979	0.7203	91	35	18.72	7.14	18
Fall 2004	0.6943	0.6692	107	37	16.76	7.85	19
Spring 2005	0.6852	0.7600	59	39	22.29	8.22	19

Table 4. Coefficient Alpha (with post-test components) for the SCI across six semesters.

questions are not conforming to the overall conceptual framework of the test. The results presented here indicate that this metric can be used to aide in item revision or deletion, especially when coupled with focus group discussion. Coefficient alpha and "alpha if item deleted" should simply be considered tools in the test-writer's toolbox. Once reliability has been established, other judgments, such as those based on validity, are still important in evaluating the appropriateness of test items.

Acknowledgments

The researchers wish to acknowledge the support provided by a grant from the National Science Foundation, Division of Undergraduate Education, Assessment of Student Achievement program (DUE-0206977). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. In addition, the authors would like to express their thanks to the reviewers for their constructive comments that have led to an enhanced manuscript.

References

American Educational Research Association. 1999. *Standards for educational and psychological testing.* Washington, D.C.: American Educational Research Association.

Cortina, J. M. 1993. What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology* 78 (1): 98–104.

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

Evans, D. L., G. L. Gray, S. Krause, J. Martin, C. Midkiff, B. M. Notaros, M. Pavelich, D. Rancour, T. R. Rhoads, P. Steif, R.A. Streveler, and K. Wage. 2003. Progress on concept inventory assessment tools. Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference, Boulder, CO, Session T4G-8.

Fortenberry, N. L. 2006. An extensive agenda for engineering education research. *Journal of Engineering Education* 95 (1): 3–5.

Guttman, L. 1945. A basis for analyzing test-retest reliability. *Psychometrika* 10:255–82.

Hake, R. 1998. Interactive-engagement vs. traditional methods: A six-thousand student survey of mechanics test data for introductory physics courses. *American Journal of Physics* 6 (1): 64–75.

Halloun, I., and D. Hestenes. 1985. The initial knowledge state of college physics students. *American Journal of Physics* 53 (11): 1043–55.

Hestenes, D., M. Wells, and G. Swackhamer. 1992. Force concept inventory. *The Physics Teacher* 30 (March): 141–58.

Hogan, T. P. 2003. *Psychological testing.* New York: John Wiley & Sons, Inc.

Kelley, T. 1939. The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology* 30:17–24.

Kline, T. J. B. 2005. *Psychological testing.* Thousand Oaks, CA: Sage Publications, Inc.

Kuder, G. F., and M. W. Richardson. 1937. The theory of the estimation of test reliability. *Psychometrika* 2 (3): 151–60.

Special report: The research agenda for the new discipline of engineering education. 2006. *Journal of Engineering Education* 95 (4): 259–61.

Statistics Concept Inventory homepage. 2007. https://engineering.purdue.edu/SCI, accessed May 2007.

Stone, A., K. Allen, T. R. Rhoads, T. J. Murphy, R. L. Shehab, and C. Saha. 2003. The Statistics Concept Inventory: A pilot study. Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference, Boulder, CO, Session T3D-6.

Streiner, D. L. 2003. Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment* 80 (1): 99–103.

Streveler, R. A., and K. A. Smith. 2006. Conducting rigorous research in engineering education. *Journal of Engineering Education* 95 (2): 103–05.

Whittington, D. 1998. How well do researchers report their measures? An evaluation of measurement in published educational research. *Educational and Psychological Measurement* 58 (1): 21–37.

Wilkinson, L., and The Task Force on Statistical Inference. 1999. Statistical methods in psychology journals. *Psychological Bulletin* 54 (8): 594–604.

Authors' Biographies

Kirk Allen is a post-doctoral researcher in the Department of Engineering Education at Purdue University He earned his Ph.D. in Industrial Engineering from the University of Oklahoma in 2006 and a B.S. in Chemical Engineering in 2000. He is a member of the American Society for Engineering Education.

Address: Department of Engineering Education, 400 Centennial Mall Drive, Room 212, West Lafayette, IN 47907-2016; telephone: (+1) 765.496.9526; fax: (+1) 765.496.1180; e-mail: allenk@purdue.edu.

Teri Reed-Rhoads is assistant dean of Engineering for Undergraduate Education and associate professor of Engineering Education at Purdue University. Her B.S. in Petroleum Engineering is from the University of Oklahoma. She received her M.B.A. from the University of Texas of the Permian Basin and her Ph.D. in Industrial Engineering from Arizona State University in 1999. Her research areas include statistics education, concept inventory development, assessment and evaluation of learning and programs, recruitment and retention, diversity, equity, and P-12 engineering education outreach.

Address: Department of Engineering Education, 400 Centennial Mall Drive, Room 212, West Lafayette, IN 47907-2016; telephone: (+1) 765.494.4966; fax: (+1) 765.496.1180; e-mail: trhoads@purdue.edu.

Robert A. Terry is an associate professor in the Department of Psychology at the University of Oklahoma. He earned his Ph.D. at the University of North Carolina at Chapel Hill in 1989. His research interests are in psychometric theory and industrial/organizational psychology.

Address: Department of Psychology, 455 W. Lindsey Street, Dale Hall Tower, Room 705, Norman, OK, 73019-2007; telephone: (+1) 405.325.4511; fax: (+1) 405.325.4737; e-mail: rterry@ou.edu.

Teri J. Murphy is an associate professor of Mathematics at the University of Oklahoma. Her B.S. in Mathematics is from Kent State University. She received her M.S. degrees in Mathematics and Applied Mathematics and her Ph.D. in Mathematics Education from the University of Illinois at Urbana-Champaign. Her research area is undergraduate mathematics education.

Address: Department of Mathematics, 601 Elm Ave., PHSC 423, Norman, OK, 73019-0315; telephone: (+1) 405.325.4071; fax: (+1) 405.325.7484; e-mail: tjmurphy@ou.edu.

Andrea Stone completed her Ph.D. in Mathematics with a specialization in research in undergraduate curriculum and pedagogy at the University of Oklahoma in 2006. She earned an M.S. in applied mathematics (1998) and a B.S. in mathematics (1997) from the University of Notre Dame. She now works as a consultant in the area of test analysis.

Address: 5204 Schuyler Ct., Columbia, MO 65202; e-mail: adstone@centurytel.net.

Appendix: Derivation of variance for dichotomous scoring

This relationship is derived using the basic definition of population variance:

$$\sigma^2 = V_i = \frac{\Sigma(x_i - \mu)^2}{n}$$

where: x_i are the individual observations (0 or 1)
μ is the population mean (p_i for each question)
n is the total number of observations (students)

For dichotomously scored data, the sum portion of the variance equation can be broken down into the 0 and 1 scores:

For 0 scores on a question: $\Sigma(x_i - \mu)^2 = (0 - p_i)^2 q_i n = p_i^2 q_i n$

The term $(0 - p_i)^2$ represents the fact that 0 is the value of each observation (x_i) and that the overall mean for each question is p_i. The term $q_i n$ accounts for summing all incorrect scores for that question (the proportion incorrect multiplied by the total number).

For the correct students, the same logic holds in calculating Vi, but now each x_i is 1 and the total number of correct students is $p_i n$.

For 1 scores on same question: $\Sigma(x_i - \mu)^2 = (1 - p_i)^2 p_i n = q_i^2 p_i n$

Combining the 0 and 1 portions and dividing by n yields the total variance for an individual question (V_i):

$$V_i = \frac{p_i^2 q_i n + q_i^2 p_i n}{n}$$

The next step is to divide out the n's and re-arrange the numerator:

$$V_i = p_i q_i (p_i + q_i)$$

The term $p_i + q_i$ is the sum of the proportion correct plus the proportion incorrect. This must total 1. Therefore, the final result for each question's variance is:

$$V_i = p_i q_i$$