The Statistics Concepts Inventory: Developing a Valid and Reliable Instrument

Kirk Allen[1], Andrea Stone[2], Teri Reed Rhoads[1], Teri J. Murphy[2]
University of Oklahoma
School of Industrial Engineering[1]
Department of Mathematics[2]

Abstract

The Statistics Concepts Inventory (SCI) is currently under development at the University of Oklahoma. This paper documents the early stages of assessing the validity, reliability, and discriminatory power of a cognitive assessment instrument for statistics. The evolution of test items on the basis of validity, reliability, and discrimination is included. The instrument has been validated on the basis of content validity through the use of focus groups and faculty surveys. Concurrent validity is measured by correlating SCI scores with course grades. The SCI currently attains concurrent validity for Engineering Statistics courses, but fails to do so for Mathematics Statistics courses. Because the SCI is targeted at Engineering departments, this is a good starting point, but the researchers hope to improve the instrument so that it has applicability across disciplines. The test is shown to be reliable in terms of coefficient alpha for most populations. This paper also describes how specific questions have changed as a result of answer distribution analysis, reliability, discrimination, and focus group comments. Four questions are analyzed in detail: 1) one that was thrown out, 2) one that underwent major revisions, 3) one that required only minor changes, and 4) one that required no changes.

Introduction

The concept inventory movement was spurred by the development and successful implementation of the Force Concept Inventory[1,2]. The FCI was developed as a pre-post test to identify student misconceptions when entering a 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[3].

The success of the FCI prompted researchers to develop instruments in other fields. In light of recent ABET accreditation standards which focus on outcomes rather than simply fulfilling seat time requirements, many engineering fields have begun to development concept inventories[4].

The pilot study of the Statistics Concept Inventory (SCI), conducted in Fall 2002, contained 32 questions and examined differences in scores due to gender and academic discipline[5]. The study

compared five groups, students majoring in mathematics, physical sciences, engineering, life sciences, and social sciences. The groups' average scores ranked from highest to lowest in that order. The results show that discipline has a small effect on SCI score. The only statistically significant differences were between mathematics and social science majors, with no differences between any other pair. The difference found was more likely due to other causes than to discipline. Gender had a significant effect, with males outperforming females. SCI scores were also positively correlated with mathematics experience and a statistics attitudinal measure[6].

The second stage in the development of the SCI aims to improve the validity, reliability, and discriminatory power of the instrument. This paper documents the evolution of the instrument in a broad sense and describes how the questions were modified to help these changes come about.

Validity

Validity refers to the extent that an instrument measures what it claims to measure. Validation is the process of accumulating evidence that supports this claim. It is an on-going process in that the instrument must be constantly evaluated as its uses and needs evolve[7]. There are many types of validity (e.g., face validity, concurrent validity, predictive validity, incremental validity, construct validity[8]). However, they are not mutually exclusive. The researchers focused on content and concurrent validity because they are broad and are described in most psychometric textbooks. Predictive validity is also mentioned because it relates to students' pre-conceptions. The SCI's validity is measured in terms of its target audience, introductory statistics courses in engineering departments. This section describes the content validity, concurrent validity, and predictive validity of the SCI.

Content Validity

Content validity refers to the extent to which items are (1) representative of the knowledge base being tested and (2) constructed in a "sensible" manner[7]. Achieving content validity requires finding an adequate sampling of possible topics and incorporating value judgments as to which topics to include.

As a starting point for item construction, the researchers searched textbooks and statistics journals and used personal experience to identify important concepts. The Advanced Placement (AP) Statistics course outline[9] was used as a guide to ensure breadth of coverage. Next, a survey was used to verify the appropriateness of topics and to fill gaps in coverage. The survey was sent to all faculty members in the College of Engineering at the University of Oklahoma during the Spring 2001 semester. The respondents were asked to rank the importance of statistics topics for their curricular needs. The scale ranged from 1 (not at all important) to 4 (very important), along with the option of "No opinion" if the topic was unfamiliar. Respondents were instructed to list additional items if they felt something was missing. The responses indicate that no major topic was omitted from the original list. Twenty-three faculty members completed the survey. Simultaneously, a literature search was conducted to identify misconceptions in statistics. Both journal articles (see [10,11,12,13,14]) and textbooks (see [15,16]) were utilized.

After the first round of test administration, items were analyzed in several ways. First, answer distributions were examined to find choices which were consistently not chosen. These options were either thrown out or revised. Second, focus groups were conducted to gain insight from students as to why they chose certain answers and identify other choices to serve as distracters. Items were revised based on these results. Several new items were constructed in a similar manner as described above.

This revised SCI was administered in Summer 2003. A similar revision process was conducted as described above, and new items were constructed where necessary. In addition, specific effort was made to identify poorly written questions. It is necessary to identify poorly written questions because the SCI is not intended to identify good test-takers. Each question was evaluated on the basis of seven criteria identified by Gibb[17] that may lead students with good test-taking skills to figure out the answer. The criteria are listed below:

1. Phrase-Repeat: Correct answer contains a key sound, word, or phrase that is contained in the question's stem.
2. Absurd Relationship: Distracters are unrelated to the stem.
3. Categorical Exclusive: Distracters contain words such as "all" or "every."
4. Precise: Correct answer is more precise, clear, or qualified than the distracters.
5. Length: Correct answer is longer than the distracters.
6. Grammar: Distracters do not match the verb tense of the stem, or there is not a match between articles ("a", "an", "the").
7. Give-Away: Correct answer is given away by another item in the test.

These problems were revealed both through focus groups and the researchers' analysis. Information on the application of these criteria is provided in the section Item Analysis.

Concurrent Validity

Concurrent validity is "assessed by correlating the test with other tests"[8]. The "other test" is the overall course percentage grade, which is correlated with the SCI Pre-Test, SCI Post-Test, SCI Gain (Post minus Pre), and SCI Normalized Gain (Gain as a percentage of the maximum possible Gain). Data was available for two courses: 1) Applied Statistical Methods in the Mathematics Department (Math), and 2) Applied Engineering Statistics in the College of Engineering (Engr). The correlation coefficients are presented in Table 1.

Table 1: Correlation of SCI Scores with Overall Course Grade(%), Summer 2003 (p-values in parenthesis)

Course	SCI Pre	SCI Post	SCI Gain	SCI Norm.Gain
Math #1 (n=12)	r = -0.392 (p = 0.207)	r = -0.023 (p = 0.944)	r = 0.318 (p = 0.314)	r = 0.288 (p = 0.365)
Engr (n=22)	r = 0.360 (p = 0.109)	r = 0.593** (p = 0.005)	r = 0.511* (p = 0.018)	r = 0.604** (p = 0.004)

** significant at 0.01, * significant at 0.05

These results indicate that the SCI attains concurrent validity for the Engr class but not for the Math class. In fact, the Math course has negative correlations on the Pre-Test and Post-Test. When gains are considered, the Math course yields moderate correlations, but they are not significant. While this could be random, it is encouraging that the correlations are positive. It should also be noted that this class has a small sample size. The Engineering course yields a moderate (but not significant) correlation on the Pre-Test and significant positive correlations on the three other measures. The correlation is strongest with Normalized Gain.

The failure of the Math course to attain concurrent validity is disappointing, though perhaps not surprising. The SCI was constructed to serve as an assessment instrument for Engineering Statistics courses. In general, mathematics courses are taught from a more theoretical perspective, while engineering courses are typically more applied. The use of a different textbook and different topic coverage may also contribute. If the results were reversed (low correlations for Engr), this would be a concern. Future work will aim to improve the SCI so that it is valid across all statistics courses (e.g., Mathematics, Engineering, Psychology).

For Fall 2003, grade data are available for five courses. These are all introductory statistics courses, with four in engineering departments and the same Math course as above with a different professor. Three courses are at four-year universities outside OU. The data are presented in Table 2.

Table 2: Correlation of SCI Scores with Overall Course Grade(%), Fall 2003 (p-values in parenthesis)

Course	**SCI Pre**	**SCI Post**	**SCI Gain**	**SCI Norm.Gain**
Engr (n=47)	r = 0.360* (p = 0.012)	r = 0.406** (p = 0.005)	r = 0.114 (p = 0.444)	r = 0.139 (p = 0.352)
Math #2 (n=14)	r = -0.066 (p = 0.823)	r = -0.054 (p = 0.854)	r = 0.011 (p = 0.970)	r = 0.200 (p = 0.492)
External #1 (n=43)	N/A	r = 0.343* (p = 0.024)	N/A	N/A
External #2a (n=51)	r = 0.224 (p = 0.113)	r = 0.296* (p = 0.035)	r = 0.094 (p = 0.514)	r = 0.052 (p = 0.716)
External #2b (n=48)	r = 0.400** (p = 0.005)	r = 0.425** (p = 0.003)	r = -0.034 (p = 0.818)	r = -0.041 (p = 0.780)

** significant at 0.01, * significant at 0.05

Note: Math #1 for Fall is not listed because the Post-Test was not given due to a scheduling conflict. Gains are not given for External #1 because the Pre-Test scores are not available.

The SCI consistently has significant correlations on the Post-Test except the Math course, which has near-zero correlations on all measures. However, unlike Summer 2003, the Gain and Normalized Gain do not provide significant correlations for any courses. In fact, they are close to zero. These results imply that the SCI Post-Test measures the same basic material as the courses cover, which is the most important evidence for concurrent validity.

Predictive Validity

For Summer 2003, the SCI Pre-Test lacks predictive validity with respect to final course grade. This implies that pre-knowledge is not playing a significant role in how much a student learns in the course. The conclusion is similar to that reached in the pilot study[5], which stated that statistics experience does not play a vital role in the SCI score for students in an introductory statistics course. However, data from Fall 2003 yield significant correlations of SCI Pre-Test with overall course grade for two courses. The magnitudes of the correlations are similar to the Summer 2003 Engr course, but the larger sample sizes make the Fall 2003 correlations significant.

Reliability

A reliable instrument is one in which measurement error is small, which can also be stated as the extent that the instrument is repeatable[7]. There are several types of reliability: test-retest measures answer stability on repeated administrations; alternative forms requires subjects to take two similar tests on the same subject; and internal consistency is based on inter-item correlations and describes the extent to which the test measures a single attribute (e.g., statistical knowledge).

Internal consistency is the most common measure because it requires only one administration of the test. This reduces administration costs and eliminates the issue of students gaining knowledge between test administrations. Internal consistency is measured using Cronbach's alpha[18], which is a generalized form of Kuder-Richardson Formula 20 [19]. Typically, a test is considered to be reliable if alpha is above 0.80 [7]. Other sources consider a value of 0.60 to 0.80 to be acceptable for classroom tests[20]. The Force Concept Inventory has been reported to have an alpha of 0.86 on the pre-test and 0.89 on the post-test[1]. The SCI is expected to have a lower alpha than the FCI because the FCI covers a narrower range of material. The data are presented in Table 3.

Table 3: Coefficient Alpha, Summer 2003

Course	Pre-Test Alpha	Post-Test Alpha
Engr	0.6805	0.8100
Math #1	0.6765	0.8587
Math #2	0.6902	--
REU	0.5983	--
External #1	--	0.5781

Note: Math #1 is the Math course discussed in the Validity section. REU is a summer research program; the subjects' statistics experience varied from none to several semesters. REU is listed under Pre-Test because the SCI was given only at the beginning of the summer. External #1 is an Engineering Statistics course at a four-year university. It is the first administration outside OU.

The instrument is generally reliable on the Post-Test but is slightly lacking on the Pre-Test. It is interesting to note that alpha increases from Pre-Test to Post-Test. Students are encouraged to answer all questions, which leads to a large amount of guessing on the Pre-Test and tends to lower alpha. Research on the mathematical behavior of alpha has shown this to be plausible from a theoretical viewpoint[21].

The low alpha at External #1 is a concern because it possibly indicates that the instrument is unreliable at other universities. The instrument was written based on the researchers' knowledge of the Engineering Statistics course as it is taught at OU. Data from the Fall 2003 Post-Test show that External #1 yields an alpha very similar to the OU courses. However, a new test site (External #2a and #2b) has alphas that are somewhat lower. More data are needed to determine if the test is reliable at universities outside OU. The Fall 2003 data are presented in Table 4.

Table 4: Coefficient Alpha, Fall 2003, Pre-Test and Post-Test

Course	Pre-Test Alpha	Post-Test Alpha
Engr	0.6863	0.7496
Math #1	0.7122	--
Math #2	0.6715	0.7232
External #1	0.7025	0.7314
External #2a	0.5709	0.6452
External #2b	0.6648	0.5843

Discriminatory Power

Discrimination refers to a test's ability to produce a wide range of scores. It is a desirable property of a test because tests are designed to look for differences between subjects. The discriminatory power depends on the shape of the score distribution. For example, if scores are normally distributed, it is easiest to differentiate between scores at the tails because there are few extreme scores; the middle scores are hard to differentiate because they are clustered.[22]

Discriminatory power can be measured by Ferguson's delta, which ranges from 0 (all scores the same) to 1 (each person has a unique score). A test is considered discriminating if delta is above 0.90 [8]. The SCI is found to be discriminating for all groups tested, shown in Table 5.

Table 5: Discriminatory Power, Fall 2002, Summer 2003, and Fall 2003, Post-Tests

Course	Ferguson's Delta
Fall 02 All courses	0.944
Summer 03 Engr	0.941
Summer 03 Math #1	0.936
Summer 03 REU	0.938
Summer 03 Ext. #1	0.926
Fall 03 Engr	0.964
Fall 03 Math #2	0.918
Fall 03 External #1	0.944
Fall 03 External #2a	0.943
Fall 03 External #2b	0.931

It is also important for each question to be discriminating. Item discrimination is measured by the discriminatory index, which compares the top-scoring students to the low-scoring students. For example, if 75% of the top students and 30% of the bottom students get a question correct, the

item has a discriminatory index of 0.45. To determine the top and bottom students, the optimal split is considered to be 27% at each end[23]. For simplicity, the researchers looked at the bottom and top 25% of students with all ties included. An item is considered poor if the discrimination index is below 0.20, while above 0.40 is considered high[24]. A broad view of item discrimination for the instrument is presented in Table 6. The application of these results to specific questions is discussed in the Item Analysis section.

Table 6: Item Discriminatory Index, Number of Questions in Each Range, all semesters

Course	**Poor (< 0.20)**	**Moderate (0.20 to 0.40)**	**Good (≥ 0.40)**
Fall 02 All courses	9	17	6
Summer 03 Engr	8	11	14
Summer 03 Math #1	7	5	21
Summer 03 REU	16	7	10
Summer 03 Ext. #1	15	9	9
Fall 03 Engr	11	8	15
Fall 03 Math #2	11	13	10
Fall 03 External #1	10	15	9
Fall 03 External #2a	13	12	9
Fall 03 External #2b	15	11	8

The discriminatory index improved from Fall 2002 to Summer 2003 in Engr and Math courses at OU, with more questions rating "good." The REU and External #1 each had nearly half the items rated poor for Summer 2003. For Fall 2003, results are relatively constant across the different universities. The best result is the Engr course (15 items "good"), but even that is a slight decline from Summer 2003 due to more questions being rated "poor." The other groups are similar to the Summer 2003 External #1.

Item Analysis

This section documents how specific questions have evolved as a result of answer distribution analysis, reliability, discrimination, and focus group comments. This section includes four sample questions: 1) one that was thrown out, 2) one that underwent major revisions, 3) one that required only minor changes, and 4) one that required no changes.

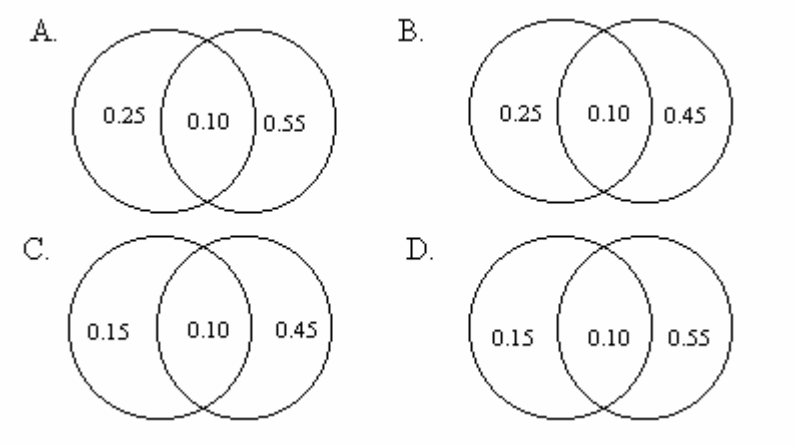
Each question's reliability is measured by alpha-if-deleted, which is calculated by SPSS ™ or SAS ™ and describes how the overall alpha would change if that one question were deleted from the test. A question which contributes favorably to reliability will have an alpha-if-deleted below the overall alpha because deleting that one item would lower the overall alpha.

For the pilot study (Fall 2002), statistics were calculated for the entire population of six statistics courses (Engr, two Math, Communications, Regression, Design of Experiments) to give a general idea of how the question behaved. For later work, the data were divided by class to examine differences across courses.

Thrown Out – Deleted completely

Fall 2002 (Question #9)

- The union of A and B = 0.80. The intersection of A and B = 0.10. A = 0.25. Which diagram correctly illustrates these conditions?



Note: D is the correct answer

This question had a negative effect on alpha (overall alpha 0.6114, alpha-if-deleted 0.6174). Although the question's alpha-if-deleted is only slightly above the overall alpha (i.e., only slightly bad), it is one of just 8 of 32 questions to be unreliable by this measure. The discriminatory index was 0.30, which falls between the low and high ranges.

Aside from choice C (chosen by 2% of respondents), the answer distribution is indicative of guessing (36% A, 28% B, 35% D). The question fits only loosely into the AP Statistics category "Exploring Data" and the faculty survey category "Methods of Displaying Data." This implies that the question does not conform to content validity. This item was not discussed in focus groups because the above considerations had already rendered it inappropriate.

Major Revisions – Totally new question under the same concept

Fall 2002 (Question #13)

- If P (A|B) = 0.70, what is P (B|A)?

  a) 0.70
  b) 0.30
  c) 1.00
  d) 0
  e) Not enough information (** correct **)
  f) Other: _____

This question had a negative effect on alpha (overall alpha 0.6114, alpha-if-deleted 0.6153). It is one of just 8 of 32 questions to be unreliable by this measure. The discriminatory index was 0.16, which was the 8th worst for the 32-item Fall 2002 SCI and in the low range. The researchers also felt the question was too symbol-oriented, which could confuse some students. The correct answer also may be an option which students would naturally want to disregard (Gibb's Categorical Exclusive), therefore making the problem unfair.

The topic (conditional probability) was considered too important to omit. It is listed explicitly in the AP Statistics outline. Conditional probability scored 2.85 out of 4 on the faculty survey (mean for all topics 2.63, median 2.62). Therefore, a new question was devised which was less formulaic and more focused on the concept.

Summer 2003 (Question #14) and Fall 2003 (Question #13)

- In a manufacturing process, the error rate is 1 in 1000. However, errors often occur in bursts. Given that the previous output contained an error, what is the probability that the next unit will also contain an error?

  a) Less than 1 in 1000
  b) Greater than 1 in 1000 (** correct **)
  c) Equal to 1 in 1000
  d) Insufficient information

From the Summer administration, the question generally had a positive effect on alpha. The exception is the Math course, which has been cited for possible differences in teaching method and topics covered. The answer distribution for the Math course also indicates possible guessing (33% B, 42% C, 25% D). The discriminatory index displays the same basic pattern as alpha-if-deleted and is at an acceptable level. The information is summarized in Table 7.

Table 7: Reliability Statistics for a Question with Major Revisions, Summer 2003, Post-Test

Course	**Alpha-if-deleted**	**Overall Alpha**	**Difference (+ is good)**	**Discriminatory Index**
Engr	0.7977	0.8100	+0.0123	0.45
Math #1	0.8626	0.8587	-0.0039	0
REU	0.5899	0.5983	+0.0084	0.34
External #1	0.5297	0.5781	+0.0484	0.55

Focus group comments reveal that students have an understanding of the problem's purpose (i.e., conditional probability or a "non-memoryless" property). One student correctly chose B because the problem did not say memoryless, while another made the connection that the bursts would "throw off the odds" (direct quote). Several students felt the notion of a burst was not well-defined, which led to choice D. This potential problem must be monitored on future administrations. The question was not further revised for the Fall 2003 SCI.

Minor Revisions – Basic question remains the same but choices change somewhat

Fall 2002 (Question #20)

- Which of the following could never be considered a population?

  a) The students in your statistics class
  b) The football teams in the Big 12
  c) The players on a football team
  d) Three randomly selected Wal-Mart stores (** correct **)

Focus group comments provided useful insights with this question. Specifically, incorrect answers were eliminated by logic that did not match the researchers' goal for this question. For choice A, the concept of bias was mentioned by a student who felt that you would not want to conduct an experiment on a group that you are closely associated with. The choice was changed to "a physics class." Several students felt that the number of items in the choice was important, and this led to at least one student correctly choosing D because it is the smallest number.

The researchers felt that choice D was too obvious because it was the only choice that contains the word "random." Choice C was modified to help eliminate this problem.

This question was kept in a similar format for three reasons. First, it was a reliable question in terms of alpha-if-deleted (0.5970, overall alpha 0.6114). Secondly, the question's discriminatory index is 0.36, which ranks 9th best out of 32 items. Finally, the question fit into the AP Statistics category "Populations, samples, and random selection."

Summer 2003 (Question #24)

- Which of the following could never be considered a population?

  a) The students in a physics class
  b) The football teams in the Big 12
  c) The players on a randomly selected football team
  d) 100 randomly selected Wal-Mart stores (** correct **)

This new version of the item has potential problems with reliability and the discriminatory index. The results are summarized in Table 8.

Table 8: Item Analysis Statistics for a Question after Minor Revisions, Summer 2003, Post-Test

Course	**Alpha-if-deleted**	**Overall alpha**	**Difference (+ is good)**	**Discriminatory Index**
Engr	0.8128	0.8100	-0.0028	0.09
Math #1	0.8611	0.8587	-0.0024	0.25
REU	0.6050	0.5983	-0.0067	0.16
External #1	0.5799	0.5781	-0.0018	0.25

Focus group comments revealed minor problems that required revision. One student chose D because it is the only option that is not people. Choice A, the least-selected option, was changed so that it did not relate to people.

Upon further inspection, it became clear that choice D looks different from the incorrect options because it is the only choice that does not begin with "The" (similar to Gibb's Precise criteria). Further modification was made so that the answers looked more similar to each other.

Fall 2003 (Question #14)

- Which of the following could never be considered a population?

  a) Four-door cars produced in a factory in Detroit
  b) Football teams in the Big 12
  c) Players on a randomly selected football team
  d) One hundred randomly selected Wal-Mart stores (** correct **)

Results from the Fall 2003 Post-Test indicate that the item's reliability has improved due to the changes discussed above. The results are summarized for the Post-Test in Table 9.

Table 9: Item Analysis Statistics for a Question after Minor Revisions, Fall 2003, Post-Test

Course	**Alpha-if-deleted**	**Overall alpha**	**Difference (+ is good)**	**Discriminatory Index**
Engr	0.7545	0.7496	-0.0049	0.19
Math #2	0.7114	0.7232	+0.0118	0.40
External #1	0.7322	0.7314	-0.0008	0.20
External #2a	0.6401	0.6452	+0.0051	0.34
External #2b	0.5703	0.5843	+0.0140	0.36

These five groups match the target audience of Introductory Statistics courses (although one is in the Mathematics department). While the results are mixed, it is an improvement over the Summer administration. With a maximum discriminatory index of 0.40 and two courses with a negative difference, the question may still need improvement.

No Changes – Question remains in its original form

Fall 2002 (Question #1), Summer 2003 (Question #1), Fall 2003 (Question #16)

- 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.

The SCI pilot test and focus groups show that students understand the question and utilize logic that is consistent with what the researchers anticipated. Choice D was the most common incorrect choice, and focus groups commented that remembering to order the number before finding the median is essential. The placement of the largest number (101) in the middle of the original list makes this crucial.

The question is generally reliable as measured by alpha-if-deleted. The discriminatory index is more than adequate for the Fall 2002 groups and the Summer 2002 Engr class. The most notable short-coming is with the Summer 2003 External #1, previously cited as having reliability problems. The Fall 2003 data are even more promising, with all groups exhibiting reliability and

all discriminatory indices above 0.20. However, the question is still under-performing at the External sites compared to OU. The results are presented in Table 10.

Table 10: Item Analysis Statistics for a Question with No Changes, Fall 2002, Summer 2003, and Fall 2003, Post-Tests

Course	Alpha-if-deleted	Overall alpha	Difference (+ is good)	Discriminatory Index
Fall 02 All courses	0.5929	0.6114	+0.0185	0.56
Summer 03 Engr	0.7968	0.8100	+0.0132	0.75
Summer 03 Math #1	0.8605	0.8587	-0.0018	0.25
Summer 03 REU	0.5955	0.5983	+0.0028	0.16
Summer 03 Ext. #1	0.5746	0.5781	+0.0035	0.14
Fall 03 Engr	0.7327	0.7496	+0.0169	0.63
Fall 03 Math #2	0.7045	0.7232	+0.0187	0.67
Fall 03 External #1	0.7179	0.7314	+0.0140	0.33
Fall 03 External #2a	0.6311	0.6452	+0.0141	0.28
Fall 03 External #2b	0.5729	0.5843	+0.0114	0.21

The topic (data summary) is of extreme importance in the faculty survey, with a mean score of 3.65 out of 4 (2nd highest overall). The question also satisfies the AP Statistics topic "Measuring center: median and mean." Results from the Summer 2003 suggest that some students are gaining this knowledge during their Introductory Statistics course. However, results from Fall 2003 show very little gain. In fact, the Pre-Test scores from Fall are most similar to the Post-Test scores from Summer. The results are summarized in Table 11.

Table 11: Knowledge Gain on a Question, Summer 2003 and Fall 2003

Course	Pre-Test % Correct	Post-Test % Correct	Gain
Summer 03 Engr	41%	68%	27%
Summer 03 Math #1	58%	83%	25%
Fall 03 Engr	74%	72%	-2%
Fall 03 Math #2	71%	86%	15%
Fall 03 External #2a	81%	84%	3%
Fall 03 External #2b	94%	88%	-6%

Note: This only includes students who took Pre- and Post-Tests. External #1 is not available because the Pre-Test was given anonymously, which ruled out matching up students' Pre- and Post-Test scores.

Analysis of the answer distributions from Summer 2003 show that very few students change from being correct on the Pre-Test to incorrect on the Post-Test. For Engr, 6 of 9 students who were correct on the Pre-Test were also correct on the Post-Test, while 9 of 13 students who were incorrect on the Pre-Test responded correctly on the Post-Test. For the Math course, all 7 students who were correct on the Pre-Test were also correct on the Post-Test, while 3 of the 5 students incorrect on the Pre-Test responded correctly on the Post-Test. This is summarized in the matrices on the next page. Matrices for Fall 2003 are not as instructive due to the small gains.

	Engr Pre-Test					Math Pre-Test				
	A	B	**C**	D	Post		A	B	**C**	D
A	0	0	**0**	1		A	0	0	**0**	1
B	0	1	**1**	0		B	0	0	**0**	0
C	**2**	**2**	**6**	**5**		**C**	**0**	**0**	**7**	**3**
D	0	1	**2**	1		D	0	0	**0**	1

Note: This only includes students who took both Pre- and Post-tests. C is the correct answer.

For the SCI, it is important that students show the ability to gain knowledge on questions which are typically covered in an Introductory Statistics course. While many students will be familiar with the concept of median from previous experience, it is a topic which will almost assuredly be covered by the instructor in the course and it is therefore expected that students should have an increased knowledge at the end of the course.

These factors illustrate that the question as originally written is meeting its intended purpose. The question will continue to be monitored along these lines for future administrations, with special attention to the Pre-Test scores and gains to see if the question may be too easy.

Conclusions

The SCI has been administered in three semesters (Fall 2002, Summer 2003, Fall 2003). The results indicate that the instrument has improved in terms of validity, reliability, and discriminatory power. The improvements were brought about through careful editing of questions, drawing on objective statistics (e.g., alpha-if-deleted, discriminatory index, answer distributions) and subjective analysis (e.g., focus groups, researchers' experience).

The next stage in the development process is to perform a more sophisticated statistical analysis of the results. The current focus is on structural equation modeling, which is expected to provide meaningful groupings for the questions. By grouping the questions in a statistically meaningful manner, the test can provide meaningful sub-scores. Subsequently, these sub-scores can be used to calculate Cronbach's alpha and correlations with overall course grades. This will provide additional information on reliability and validity.

As with the Force Concept Inventory, the ultimate goal is to produce an instrument which is nationally recognized as a useful tool for improving student learning in statistics, both by identifying students' problem areas and providing feedback which professors can use to improve teaching. The progress in this project's first year provides a solid background for this to happen.

References

1. Halloun, I. and Hestenes, D., "The initial knowledge state of college physics students", American Journal of Physics, 1985, **53** (11): pp. 1043-1055.

2. Hestenes, D., Wells, M., and Swackhamer, G., "Force Concept Inventory", The Physics Teacher, 1992, **30** (March): pp. 141-158.

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

4. Evans, D.L., Gray, G.L., Krause, S., Martin, J., Midkiff, C., Notaros, B.M., Pavelich, M., Rancour, D., Rhoads, T.R., Steif, P., Streveler, R.A. and Wage, K., "Progress On Concept Inventory Assessment Tools", Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference, 2003, Session T4G-8.

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

6. Schau, C., Dauphinee, T. L., Del Vecchio, A., and Stevens, J. J., "Surveys of Attitudes toward Statistics", retrieved October 2, 2002 from http://www.unm.edu/~cshau/downloadsats.pdf.

7. Nunnally, J., Psychometric Theory. McGraw-Hill: New York, 1978.

8. Kline, P., A Handbook of Test Construction. Methuen & Co. Ltd: New York, 1986.

9. College Board, "Course Description: Statistics", retrieved December 18, 2003 from http://www.collegeboard.com/prod_downloads/ap/students/statistics/ap03_statistics.pdf.

10. Garfield, J. and Ahlgren, A., "Difficulties in Learning Basic Concepts in Probability and Statistics: Implications for Research", Journal for Research in Mathematics Education, 1988, **19** (1): pp. 44-63.

11. Kahneman, D. and Tversky, A., "Subjective Probability: A Judgment of Representativeness", Cognitive Psychology, 1972, **3** (3): pp. 430-453.

12. Konold, C., "Issues in Assessing Conceptual Understanding in Probability and Statistics", Journal of Statistics Education, 1995.

13. Konold, C., Pollatsek, A., Well, A., Lohmeier, J., and Lipson, A., "Inconsistencies in Students' Reasoning About Probability", Journal for Research in Mathematics Education, 1993, **24** (5): pp. 392-414.

14. Pollatsek, A., Lima, S., and Well, A. D., "Concept or Computation: Students' Understanding of the Mean", Educational Studies in Mathematics, 1981, **12**: pp. 191-204.

15. Montgomery, D. and Runger, G., Applied Statistics and Probability for Engineers. Wiley: New York, 1994.

16. Moore, D., The Active Practice of Statistics. W. H. Freeman and Company: New York, 1997.

17. Gibb, B., Test-Wiseness as Secondary Cue Response. Dissertation, Stanford University, 1964.

18. Cronbach, L., "Coefficient Alpha and the Internal Structure of Tests", Psychometrika, 1951, **16** (3): p. 297-334.

19. Kuder, G.F. and Richardson, M.W., "The Theory of the Estimation of Test Reliability", Psychometrika, 1937, **2** (3): pp. 151-160.

20. Oosterhof, A., Developing and Using Classroom Assessment. Merrill / Prentice Hall: Englewood Cliffs, New Jersey, 1996.

21. Allen, K., "Explaining Cronbach's Alpha", available at http://coecs.ou.edu/sci under Publications. A paper is also being prepared with the findings.

22. Ausubel, D., Educational Psychology: A Cognitive View. Holt, Reinhart, and Winston: New York, 1968.

23. Kelley, T., "The Selection of Upper and Lower Groups for the Validation of Test Items", Journal of Educational Psychology, 1939, **30**: pp. 17-24.

24. Ebel, R., "Procedures for the Analysis of Classroom Tests", Educational & Psychological Measurement, 1954, **14**: pp. 352-364.

Biographical Information

KIRK ALLEN
Kirk Allen is a graduate student in the Industrial Engineering Department at the University of Oklahoma. He earned a B.S. in Chemical Engineering in May 2000. He expects to earn his M.S. in Industrial Engineering in 2004. He is currently working on the Statistics Concepts Inventory project.

ANDREA STONE
Andrea Stone is a graduate student in the Department of Mathematics at the University of Oklahoma. She earned an M.S. in applied mathematics (1998) and a B.S. in mathematics (1997) from the University of Notre Dame. She currently is pursuing a Ph.D. in Mathematics with a specialization in research in undergraduate curriculum and pedagogy.

TERI REED RHOADS
Teri Reed Rhoads is the Director of Engineering Education of the College of Engineering, the University of Oklahoma and an Assistant Professor of Industrial Engineering. Dr. Rhoads is actively involved in research with industry as well as with the National Science Foundation, the U. S. Department of Education, and the local school district Foundation.

TERI J. MURPHY
Teri J. Murphy is an associate professor in the Department of Mathematics at the University of Oklahoma. She earned an M.S. in mathematics (1994), an M.S. in applied mathematics (1990), and a Ph.D. in undergraduate mathematics education from the University of Illinois at Urbana-Champaign (1995).