
New Practices for Administering and Analyzing the Results of Concept Inventories

PAUL S. STEIF

*Department of Mechanical Engineering
Carnegie Mellon University*

MARY A. HANSEN

*Department of Secondary Education and Graduate Studies
Robert Morris University*

ABSTRACT

Concept inventories can be refined and honed into informative assessment tools to serve instruction. The present paper suggests new practices for administering and analyzing the results of concept inventories. Web-based administration enables broader participation across universities and colleges, and ensures the retention of the full set of data necessary to conduct other analyses. Issues related to the provision of meaningful concept level information are addressed, as are the benefits of making direct comparisons with other measures of performance. The value of administering inventories as pre-tests is examined, and the potential for offering misconception diagnosis based on inventories is explored.

Keywords: statics, concept, inventory

I. INTRODUCTION

Concept inventories (CIs), tests which gauge student understanding of basic concepts, have been under development in a number of engineering courses and other subjects [1]. Publications on concept inventories typically focus on the development of a particular inventory and on the presentation of test results. In the present paper, we address new approaches to administration of inventories and processing of their results. These approaches can increase the usefulness of inventories as tools in the learning/instructional process. While these approaches have been pursued and are illustrated for the case of the Statics Concept Inventory (SCI) [2, 3], they are applicable to many inventories.

A. Testing Online

Developers of inventories need to recognize that most instructors are busy with their classes, and have only so much time and expertise to devote to the assessment process. While it is common for inventory developers to provide instructors with paper-and-pencil tests to administer in class, enabling the inventory to be taken online at a centralized Web site offers significant benefits. In this way, developers always have access to the full set of test results, so they can use that information to refine and improve the test items.

Furthermore, as developers expand the information that can be extracted from the test, we can provide the instructor with the best practices in interpreting the test scores. Finally, Web-based testing also simplifies the process for the instructor and frees up class time.

B. Concept Subscores

It is traditional for inventory developers and instructors to focus on total scores. A total score that is low compared to other institutions might signal to an instructor that an overall improvement in teaching is warranted. Hake [4] indeed found that higher normalized gains on the Force Concept Inventory [5] were found in classes using more active engagement techniques. But, most inventories are developed to assess student knowledge of a set of concepts that are core to a subject; an instructor will have more actionable guidance if those specific concepts on which students performed most poorly are identified. Having concept-specific information from the inventory could, therefore, be of great benefit.

C. Compare with Other Measures

Scores on inventories should signal learning that is observable by other means. To this end, inventory developers should seek to collaborate with instructors and make arrangements to compare inventory results with other measures of class performance (with proper procedures to maintain student anonymity). One obvious comparison is with scores on class exams. Comparison of student performance on exams and the inventory can be one basis for validating the inventory results. In addition, if instructors use the same inventory over several years, it becomes possible to evaluate the benefits of new instructional approaches. By retaining some conceptually similar problems on exams from year to year, instructors can determine if improved student performance on an inventory is indeed associated with broader improvement in the classroom. In fact, improved performance on an inventory might be more than just a by-product of improved learning, but also a route for achieving it.

D. Relevance of Pre-test

The notion of testing at the start of a course (pre-testing) has been reflexively assumed to be necessary and valuable. This presumes that the same instrument is capable of offering useful information at both ends of the course. While this appears to be true for the Force Concept Inventory, this relevance may need to be re-evaluated for each inventory. Through analysis of pre-test, post-test, and classroom exam scores, we show that for most purposes, the SCI offers negligible information as a pre-test.

II. SAMPLE

Over the past three years, the SCI has been taken by over 4,000 students (post-statics) at 22 institutions. Results from previous

administrations have informed the evolution of the inventory. Data presented here are from the fall 2005 semester when the SCI was administered on the Web to 1,255 students from 16 classes, at 14 different institutions. Generally, students were sophomores and juniors enrolled in primarily Statics, but also in two follow-on courses: Dynamics and Mechanics of Materials. Some analyses of the Fall 2005 administration, broken down by class, have been presented recently [6, 7]; we refer to some of these classes by number (1 to 16).

III. COLLECTION OF DATA ONLINE

Concept inventories typically consist of a set of multiple choice items, where the incorrect answer choices (distracters) for each item reflect common student misconceptions regarding the concept underlying that item. Collecting the full set of data online—including whether each student answers each question correctly and the answer chosen for each question—enables nearly all the developments described below. An online version of the SCI has been available for two years [8]. Typically, students complete the test on their own within a time frame of several days to a week. Instructors can alternatively choose to set aside time in class for students to take the test online, but under proctored conditions, so that collaboration or extra time is not allowed. The authors suspect that little collaboration takes place with the online testing, since instructors do not generally give students credit based on their inventory scores. Negligible collaboration would also be consistent with the correlations discussed between exam scores and inventory scores. In addition, for other CIs, studies have found little difference in performance of students on paper-and-pencil and online versions. As reported by Cheng [9], no appreciable differences in results were found from the FCI administered online or in-class. The Statistics Concept Inventory is another CI that is currently available online [10].

When students take the inventory without the score affecting their class grade, some fraction will inevitably devote little effort to answering the questions. One could simply accept the fact that the data from a given class will always include some fraction of students who did not put in much effort. Offering the test online, however, provides a quantitative basis for excluding data that is unlikely to be representative of student ability. The Web-based testing system for the SCI captures the time elapsed from logging in to the exam site to completion. Statistics of performance were studied as a function of the time spent on the test; these are displayed in Table 1 for a limited time range.

Two aspects of the data are notable: the scores for the ranges of zero to five minutes and five to ten minutes are only slightly above

Time	0:5	5:10	10:15	15:20	20:25	25:30	30:35
%	4.2	3.0	4.9	7.9	10.0	10.2	12.0
Mean	6.42	7.18	9.70	10.72	12.64	13.02	14.90
Max	13	21	26	26	26	25	27

Table 1. Percentages of students who completed the SCI in ranges of zero to five minutes, five to ten minutes, etc., and the mean and maximum scores in each group (27 total items).

the random guessing mean (5.4), which is nearly equal to the mean at the beginning of Statics at most institutions. Furthermore, no student in these two groups received a score in excess of 21. By contrast, at least one student taking from 10 to 15 minutes was able to score as high as 26. Furthermore, this group's mean score of 9.70 is nearly equal to the overall mean for some classes after Statics. There is no noticeable pattern in the variation of mean scores for the remainder of examinees with time above 25 minutes; the mean score for all examinees is 12.71. Thus, the normally suggested testing time of one hour is plenty of time in which to complete this test. Limiting students to this amount of time, as might be done in an in-class setting, would produce few changes in the scores. From these findings, results based on tests completed in less than 10 minutes were eliminated from further analysis, reducing the data set from 1,255 to 1,164 examinees.

IV. CONCEPT SUBSCORES

In the development of all inventories, there is an attempt to identify those concepts which are core to the subject. Questions (items) are then devised to address those concepts. In contrast with classroom exams, which in principle require reasoning based on multiple concepts, one goal of CIs is to isolate the understanding of discrete concepts. Thus, by their design, many CIs should be able to associate each item to a concept.

In constructing the SCI, we have deliberately devised items that pertain to specific concepts. Currently there are three items for each of nine concepts which span much of the subject of Statics. (A conceptual framework for Statics is presented in [11].) Thus, in addition to the total score, we can report back to each student nine concept subscores (as well as whether each item was correct). By breaking down conceptual understanding in this way, we provide information to students and instructors enabling them to focus on specific concepts where there are apparent weaknesses. The concepts are listed in Table 2.

The development of a CI faces competing demands: information derived should be reliable, and should be reflective of as much of the subject as possible. Both of these goals must be accomplished with a limited number of items; it is not practical to expect students to spend more than approximately one hour on such a test. If there is only one item on a concept, then random error will make the information on

Concept
A: Drawing forces on separated bodies
B: Newton's 3 rd Law
C: Static equivalence
D: Roller joint
E: Pin-in-slot joint
F: Loads at surfaces with negligible friction
G: Representing loads at connections
H: Limits on friction force
I: Equilibrium

Table 2. Breakdown of the SCI into nine concepts tested.

that concept unreliable; with more questions the reliability increases. We must guard against another danger: if items are too similar, then correct answers only indicate proper use of the concept in a narrow context.

It is standard in evaluating tests to extract correlations between items; two items correlate highly if each student is likely to answer both correctly or both incorrectly. If two questions require the same knowledge, or are part of the same concept, they should be highly correlated; if they are too similar, their correlations may be too high. Two questions requiring very different conceptual knowledge may have a low correlation.

Commonly applied psychometric analyses of tests, such as reliability and factor analyses, are based on the correlations between items. Reliability, as measured by Cronbach's coefficient α , increases with the average correlation between *all* pairs of items and the number of items. Factor analysis also considers inter-item correlations and tries to identify factors, each of which combines several items. Even though the number of factors is much less than the total number of items, if the factor representation is valid, the factors still explain much of the variability in the data. As a simple example of factors that is relevant here, each factor could be the sum of the scores on the items corresponding to a single concept. The higher the correlation between items within a factor, and the lower the correlation between items in different factors, the better such a breakdown into factors would summarize the data. Thus, high reliability and a strong factor structure compete with each other to some extent. (See Carmines and Zeller [12] for other measures of reliability, which account for multidimensionality in the data.)

We reported previously [7] on reliability for this data set ($\alpha = 0.838$), as well as on factor analyses. Both exploratory and confirmatory factor analyses were conducted; the results supported an overall factor structure consistent with the concepts defined in Table 2. To display essential features of the correlation structure, we show in Table 3 the reliability of the concept subscores, the mean correlation between items within each concept (Intra), and the mean correlation with items from other concepts (Inter). The subscore reliability of each concept depends on its intra-concept correlations and on the number of items in the concept. Considering the relatively small number of items in each concept (3), the reliabilities of a number of the concepts are quite high. Furthermore, the correla-

tions with items in other concepts are relatively low. The correlations shown here differ from those presented in [6], where tetrachoric correlations, sometimes used for dichotomous data (each item only right or wrong), were displayed.

The factor structure would have been even more pronounced if there were higher Intra-Concept correlations. Lower than desired correlations between items within a concept have been one of several bases for improving items from year to year. For example, the items in concept B were found to be lacking in other respects as well; they were modified for the 2006–2007 administration, and improved results, including higher correlations, are expected. The items in concept G touch on a common conceptual issue: representing unknown loads at three distinct connections. In fact, the correct answer for one item is the wrong answer for the other two and so forth (this is the only concept for which this is the case). Thus, these three items may inherently have low correlations. The items in this concept have been altered with each version of the SCI, since the wording has also been known to be problematic. Additional studies of their quality are underway.

Analysis can shed light on the source of lower correlations for other concepts by viewing the individual inter-item correlations

Concept	α	Intra-Concept r	Inter-Concept r
A	0.72	0.46	0.18
B	0.52	0.27	0.13
C	0.57	0.30	0.15
D	0.71	0.44	0.11
E	0.71	0.45	0.15
F	0.48	0.24	0.16
G	0.37	0.16	0.11
H	0.68	0.41	0.14
I	0.43	0.19	0.15

Table 3. Concept-subscore reliability (α), mean correlation between items within a concept (Intra-Concept), and mean correlation between items in different concepts (Inter-Concept).

Concept	Item1-Item2 r	Item1-Item3 r	Item2-Item3 r
A	0.43	0.56	0.41
B	0.25	0.31	0.25
C	0.29	0.28	0.34
D	0.38	0.59	0.36
E	0.45	0.45	0.45
F	0.33	0.15	0.23
G	0.28	0.12	0.09
H	0.50	0.35	0.39
I	0.14	0.30	0.14

Table 4. Correlations r among items within each concept.

within each concept (Table 4). For example, notice that for concept I (equilibrium), two correlations are much lower than the third. The items with a higher correlation (1 and 3) are comparable in difficulty; the other item is much more difficult. (Item 26, that is item 2 in concept I, is the most difficult on the test.) The correlation between a very difficult item and one of average difficulty is likely to be much lower than the correlation between two items of average difficulty. As an alternative to correlation, the results for items 25, 26, and 27 that make up concept I were analyzed as follows (Table 5). Students were split into those who did and did not answer item 25 correctly. Of those, the fractions who answered items 27 and 26 correctly were found (rows 2 and 3). We similarly split students based on answers to 26, and the fractions answering 27 correctly were tabulated (rows 4 and 5). The results would be very similar if we switched items 25 and 27 in the analysis.

Consider items 25 and 27, which have the highest of the three correlations. Of those who answered 25 correctly, 68.2 percent answered 27 correctly; likewise, of those who answered 25 incorrectly, 62.3 percent answered 27 incorrectly. When two items have roughly equivalent difficulties, not being able to answer one correctly implied higher likelihood of not answering the other correctly; likewise for the implication of answering the first incorrectly. By contrast, 26 is much more difficult than 25 (and 27). Those who answered 26 incorrectly (the vast majority of students) performed nearly the same on 27 as the entire group. But, answering 26 correctly had significant positive implications for answering 27 correctly: 70.9 percent of those answered 26 correctly also answered 27 correctly. High correlations are associated with many incidences of wrong paired with wrong and right paired with right. However, with a very difficult item and an average or simple item, we can only hope that a right answer on the difficult item implies a right answer on the less difficult item. Likewise, we can hope a wrong answer on the simpler item implies a wrong answer on the difficult item; those who answered incorrectly on 25 were indeed more likely to answer incorrectly on 26, 90.0 percent to 80.2 percent. Thus, even though the correlation between an average and a difficult item is relatively low, which negatively affects reliability and factor analyses, there is evidence that these questions measure the same concept. Thus, the total score on questions in this concept may still be a reasonable measure of understanding this concept. A similar effect occurs with concept F, where item three is much simpler than items one and two.

Item	27 Right (54.6%)	27 Wrong (45.4%)	26 Right (15.4%)	26 Wrong (84.6%)
25 Right (55.2%)	68.2%	31.8%	19.8%	80.2%
25 Wrong (44.8%)	37.7%	62.3%	10.0%	90.0%
26 Right (15.4%)	70.9%	29.1%		
26 Wrong (84.6%)	51.6%	48.4%		

Table 5. Different performances on items 27 and 26 of students who gave right and wrong answers on item 25 (rows 2 and 3), and different performances on item 27 of students who gave right and wrong answers on item 26 (rows 4 and 5).

As an example of using subscores as a tool for formative assessment, we recently reported on the experience of one class [6]. The inventory was administered towards the end of the semester. As usual, the subscores were provided to the instructor. A review session prior to the final exam was conducted at which a number of items from the inventory were discussed. The instructor chose to focus the review session on items from those concepts in which the class performed relatively poorly. Students were surveyed as to whether the review session helped them to better understand the concepts underlying the test. On a 5-point Likert scale (5 most positive), the mean response and standard deviation for the 41 respondents were 4.15 and 0.7, respectively.

V. COMPARISON WITH OTHER MEASURES

A concept inventory is of benefit if it gauges levels of understanding that are pertinent to performance elsewhere. Ideally, one would like scores to indicate whether the tester is prepared to apply the concepts during authentic use of the subject in an engineering context. As an admittedly weaker test of this relation, one can seek to compare scores with performance in a Statics course. We have used class exams for comparison, since exams are a measure of an individual's abilities (more so than homework) as is the inventory.

Correlations have been computed between final exam scores, when available, and inventory scores. For the present data set, exam scores from six schools were available; the correlations between inventory and exam were $r = 0.387, 0.528, 0.536, 0.578, 0.596,$ and 0.614 . To judge the significance of these correlations, we have compared them with the correlations between class exams. As reported previously [3], correlations between the inventory and the final exam tend to be comparable to correlations between class exams. For example, in the case of Class #9 from the present data set, for which a correlation of 0.536 was found between the inventory and the final exam, correlations of 0.547, 0.518, and 0.329 were found between the three other class exams and the final. As a second example of the test validation by comparison with other measures, a detailed analysis of three types of errors (roller, pin-in-slot, and friction) committed in two exam problems was conducted previously. Students who made an error of a particular type on the exam were found to have significantly lower inventory subscores specifically on the concept pertaining to that error (see [3] for details).

Comparisons with exams can provide additional insight into aspects of the test that may be of interest. For example, we have sought to determine whether there are significant gender differences in performance on the inventory, as a possible signal of bias in the test. Students can identify their gender as part of the Web-based implementation. In some classes, differences in performance by gender were found. The statistical significance of these

Gender	N	Exam mean	Inventory mean	r
Male	100	79.0%	62.5%	0.526
Female	48	69.8%	55.4%	0.488

Table 6. Exam, inventory scores, and correlations for male and female students in Class #9.

n	0	1	2	3	4	5	6	7	8	9	10	11	12	>12
Actual.	1	4	7	15	23	26	15	13	12	7	6	3	1	0
Eqn.(1)	0	2	7	15	22	25	23	17	11	6	3	1	0	0

Table 7. Numbers of students having various pre-test scores n compared to the predictions based on random guessing given in Equation (1). The Class of 133 students took pre-test on paper and pencil in class.

differences were quantified for only the six schools where exams were available for comparison. In five cases, no statistically significant differences between males and females were found (for exams and for the inventory). In one case, however, significant differences were found (Table 6). From t-tests, the difference in the mean scores of males and females was found to be statistically significant for the final exam ($p = 0.001$) and for the inventory ($p = 0.03$). (The effect sizes, the difference in means relative to the standard deviation, were found to be 0.59 and 0.39, respectively.) Thus, the lower performance of females on the inventory, though statistically significant, appears to mirror the performance in class exams.

As an additional indicator, correlations between final exam scores and inventory scores for the genders are also given in Table 6; the overall correlation between the final exam and the inventory is $r = 0.536$. Since the inventory scores of males and females are each individually correlated with their exam scores at a comparable level, we can assert that the lower scores of females on the inventory are unlikely to be an artifact solely of the inventory itself, but at least partially reflective of their state of knowledge in this subject, at least as indicated by the class exam. This is also supported by the finding that no statistically significant differences between genders were found for any other class.

These findings suggest that one needs to be careful in looking at gender effects of a new exam, such as a concept inventory, in the absence of other measures. The numbers of students in other groups were too small to perform comparable analyses.

VI. RELEVANCE OF PRE-TEST

Based on the successful application of the Force Concept Inventory, it has become an accepted practice to interpret post-test scores in light of the pre-test scores [4]. This makes sense when students have seen the concepts prior to the course in which the test is administered. However, for many subjects in engineering, while there are certainly concepts in previous courses that are relevant, a test that measures conceptual development adequately by the end of the course may not be a valuable measure at the beginning of the course. Again through the use of comparisons with examinations, the relevance of pre-test scores is explored more fully here.

In most classes, the distribution of inventory scores at the start of Statics corresponds closely to the distribution expected with random guessing. Since each item has five answers, a probability $f = 0.2$ that any question will be answered correctly corresponds to random guessing. Hence the probability, p , that there will be exactly n correct answers in a test of N questions is given by

$$p = (f)^n (1 - f)^{N-n} \frac{N!}{(N!)(N - n)!} \quad (1)$$

This prediction of the distribution of test scores is compared with one class. The numbers of students with various pre-test scores (n) out of $N = 27$ items for a class of 133 examinees are shown in Table 7, along with the theoretical probabilities p from equation (1). The closeness of the predicted and observed numbers of scores is evident. This class had a mean of 5.49, while the theoretical mean for random guessing is $0.2(27) = 5.4$. Except for two classes discussed in more depth below, pre-test scores are in the range from 5 to 6.5. Thus, most students at the start of Statics have little or no ability to answer questions on the inventory correctly. As shown recently [11], however, such students often invoke relevant ideas and/or language when explaining their answers to questions from the SCI.

Further study of the effect of the pre-test was conducted with the aid of comparisons with performance at the end of Statics. As an example of a class with a low pre-test mean (5.58), the correlation between the pre-test and the final exam was $r = 0.220$ for class #13. By contrast, the correlation between the post-test and exam for this same class was much higher ($r = 0.500$). Thus, the pre-test has minimal predictive validity for subsequent performance in the course.

For two classes, the pre-test scores are noticeably above the level of random guessing. These classes were studied in greater depth. There are moderate correlations between pre-test performance and exams for these two classes, but we now demonstrate that higher correlations can largely be associated with those students who achieve scores markedly above random on the pre-test. To do so, we split each of these two classes into two groups, those scoring 9 (33 percent) or less on the pre-test and those scoring 10 or more. Note that the probability of scoring above 9 with random guessing is less than 3 percent.

When the classes are split, one does find that the performances of the students in the two groups are different. The scores are shown in Table 8: those who scored higher on the pre-test did perform at a higher level on the exam. T-tests were run to compare the exam means of the two groups, and the differences were found to be statistically significant.

Correlations between the inventory and exams were used to further investigate the role of the pre-test. As shown in Table 9, the

Class	Pre-test ≤ 9		Pre-test > 9		p
	N	Exam	N	Exam	
#8	65	70.4	31	80.6	<0.0005
#16	64	88.5	40	92.7	0.003

Table 8. Mean exam scores for students scoring above and below 9 (33 percent) on pre-test, and result of t-tests showing statistically significant differences.

