PSYCHOMETRIC ANALYSIS OF A STATICS CONCEPT INVENTORY AND ITS USE AS A FORMATIVE ASSESSMENT TOOL

Paul S. Steif

Department of Mechanical Engineering
Carnegie Mellon University, Pittsburgh, PA 15213

Abstract

A multiple choice test, the Statics Concept Inventory, is used to measure conceptual progress of students in Statics. This paper reports on detailed comparisons of the results of this test to student performance on class examinations. Reasonably strong correlations are found between concept-specific sub-scores in the inventory and scores on related types of examination problems or solution errors involving similar concepts. With these findings we set the stage for using the inventory for formative assessment; that is, feeding back to students their scores on the inventory so as to point them to specific areas where improvement is necessary. A pilot effort to have a session that provides such remedial instruction is also described.

Introduction

Effective assessment is known to be key to improving learning outcomes[1,2]. For many engineering subjects, one hopes students will learn to *transfer* their newly gained knowledge to new situations, which then requires a deep understanding of the material[3]. This has been taken to mean conceptual understanding. One approach to assessing conceptual understanding, with its origins in the science education community, is the Force Concept Inventory[4]. The approach of concept inventories has been extended by the engineering education community to a variety of engineering subjects[5].

The present paper extends previous work by the author to articulate concepts in Engineering Statics[6] and then to develop an assessment instrument to measure conceptual understanding in this subject[7-8]. Here we begin to address the question of how such an inventory could be used to improve learning. In particular, we explore whether performance on a concept inventory can be correlated with other measures of performance in the respective course. We also begin to look at how one might judge the impact of feeding back to students details of their individual performance on the inventory.

Background on Statics Concept Inventory and Current Status

We explain very briefly here the underpinnings of the Statics Concept Inventory. More details on the conceptual background, typical errors, and the inventory itself are given elsewhere[6-8]. The conceptual framework, devised to be particularly relevant to problems with multiple, interconnected bodies, consists of the following four concept clusters:

C1. Forces are always in equal and opposite pairs acting *between* bodies, which are usually in contact.

C2. Distinctions must be drawn between a force, a moment due to a force about a point, and a couple. Two combinations of forces and couples are statically equivalent to one another if they have the same net force and moment.

C3. The possibilities of forces between bodies that are connected to, or contact, one another can be reduced by virtue of the bodies themselves, the geometry of the connection and/or assumptions on friction.

C4. Equilibrium conditions always pertain to the external forces acting directly on a chosen body, and a body is in equilibrium if the summation of forces on it is zero and the summation of moments on it is zero.

Accompanying these principles was a categorization of standard errors made by students, most of which involve errors associated with free body diagrams[6]. A Statics concept inventory, a twenty-seven question multiple-choice test was devised[7-8], based on this conceptual framework and categorization of errors. In the spring of 2004, the test was administered to 245 students at five universities, all of whom had completed or nearly completed a Statics course. Based on this set of data, we found the test to be reliable (Cronbach $\alpha = 0.89$) and valid. Validity was established in several ways, including comparisons with regular examinations in Statics and a confirmatory factor analysis. The confirmatory factor analysis yielded several statistics including the Goodness of Fit Index (GFI), the Comparative Fit Index (CFI), and the root mean square approximation (RMSEA), all of which were found to be in the acceptable range. This confirmatory analysis also pointed to several questions which did not fit well into the overall factor structure, questions which were also viewed as unsatisfactory for other reasons as well. In the summer of 2004, the test was revised; 8 of the 27 questions were changed completely and many others were improved in various ways. Many of the improvements were arrived at based on a distracter analysis, with the goal of having all wrong answers chosen with some non-negligible frequency. The new test is being administered at a number of universities during the 2004-2005 academic year, with the results to be reported in the near future.

Data Set of Focus

In this paper we analyze data from the 2004-2005 version of the inventory administered to the CMU Fall 2004 mechanical engineering class. We compare pre- and post-test performance on the inventory with details of the performance on various exam questions during the semester. The mechanical engineering Statics course at CMU is given over a semester of 14 weeks, with 3 hours of lecture and 1 hour of recitation per week. Nearly all students in the class are sophomores in mechanical engineering. They have taken a 3 credit introduction to mechanical engineering class in the freshman year, which touches on many of the basic subjects in mechanical engineering (and gives them exposure to design and CAE as well). Thus, they had approximately 3 weeks of Statics in the freshman course, along with physics courses, which provide some prior experience with engineering mechanics.

Students in the Statics course take four examinations during the semester, each 80 minutes in length; there is no final examination. Each exam has 3 to 5 problems. The inventory was administered to students in recitation the day before the first lecture (pre-test). The post-test was administered in recitation between the third and fourth exams, but after all concepts covered by the inventory had been addressed in lecture, in homework and in examinations.

Basic Statistics for Inventory Performance

The means, medians and standard deviations for pre- and post-tests are shown in Table 1. The t-test was performed to determine the significance of differences between the means of males and female students. As can be seen in Table 2, the differences were significant (p = 0.001 and p= 0.029) for the pre- and post-tests, respectively. As pointed out earlier, the current test involved some modifications to the previous year's test. Still, the tests are sufficiently similar (the pre- and post-test means of CMU mechanical engineering students taking last year's test were 10.60 and 20.34 respectively) that levels of reliability and validity similar to those reported in [8] can be assumed. (More detailed analyses of reliability and validity of the current test will be conducted in the future when data from a much larger group of students taking the test this year are pooled.).

Table 1. Data on the pre- and post tests (maximum score).

	N	Mean	Median	St. Dev.
Pre-test	105	10.96 (41%)	11	4.64
Post-test	97	20.50 (76%)	21	4.24

Table 2. Data on the pre- and post tests with gender differences

	N	Mean	St. Dev.	N	Mean	St. Dev.	p
	Male			female			
Pre-test	78	11.76	4.69	27	8.67	3.71	0.001
Post-test	70	21.14	3.92	27	18.85	4.65	0.029

Comparison Of Pre- And Post-Test Results With Examination Scores

One measure of the validity of the inventory is its correlation with examination scores. Scatter plots of average exam scores (for the four exams) vs. inventory are shown in Figures 1 and 2 for the pre- and post tests. (The maximum average exam score is 10.) For the post-test, the Pearson correlation coefficient $P = 0.642$. (In the previous year, with a slightly different inventory, the correlation was 0.547.) The correlation between pre-test scores and exam scores was 0.539. However, the pre-test correlation is weaker than this would suggest. While nearly all students who performed quite well on the pre-test inventory surely did well in exams, it can be seen that for students with pre-test scores less than 15, there is less correlation with exam scores. Indeed, if one only includes students with pre-test scores < 14, the correlation becomes 0.311. So all students, regardless of where they start as measured by the inventory, can progress significantly in the course. Also, while the inventory may signal students who are likely to do well, it should not be used to weed out students at the beginning, or even to signal concern. In fact, the inventory was never designed as a measure of Statics readiness, but as a measure of progress over the span of the course.

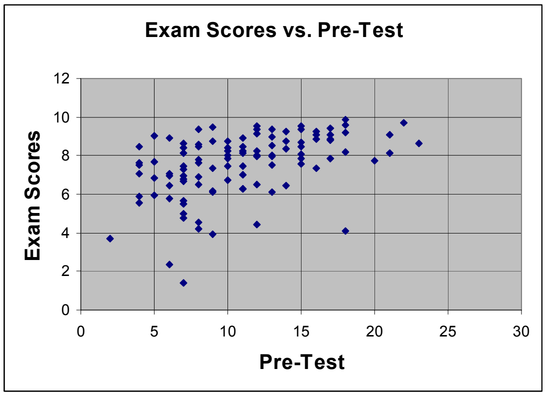

Figure 1. Exam Scores vs. Pre-test

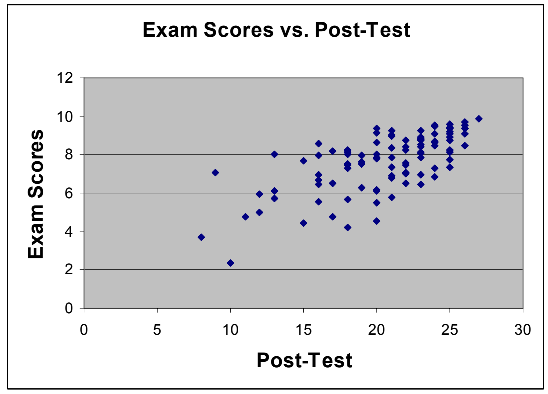

Figure 2. Exam Scores vs. Post-test

The correlation between the post-test and each of the 16 exam problems (offered in the four exams during the semester) are shown in Table 3. All remaining data in the paper are for the post-test.) Correlations ranged from 0.04 to 0.62, with the correlation being generally in the range of 0.3 and 0.4. Hence, with only a few exceptions, there is some correlation between the inventory and the exam problems. Interestingly, the strongest correlation of 0.62 was between the inventory and problem 10, which is shown in Figure 3. This is referred to as a multifaceted problem, as it draws upon many concepts in Statics. Indeed, the earlier study of concepts in Statics[6], which lead to this concept inventory, explicitly cited problems involving multiple interconnected bodies as being the prime target of study. A second exam problem involving multiple interconnected bodies had the second highest correlation of 0.57.

Table 3. Correlation coefficients between the 16 exam problems and inventory scores.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0.04	0.39	0.41	0.49	0.41	0.33	0.37	0.24	0.07	0.62	0.57	0.45	0.47	0.41	0.48	0.44

The motor shaft drives the driven link with a clockwise torque of 20 N-m. Determine the force crushing the object. In doing so, determine the forces acting between the various bodies and draw free body diagrams of each of the driven link, the slotted link, and roller alone (with no pins or other bodies attached). On each diagram draw all the forces and couples acting in correct directions with magnitudes labeled.

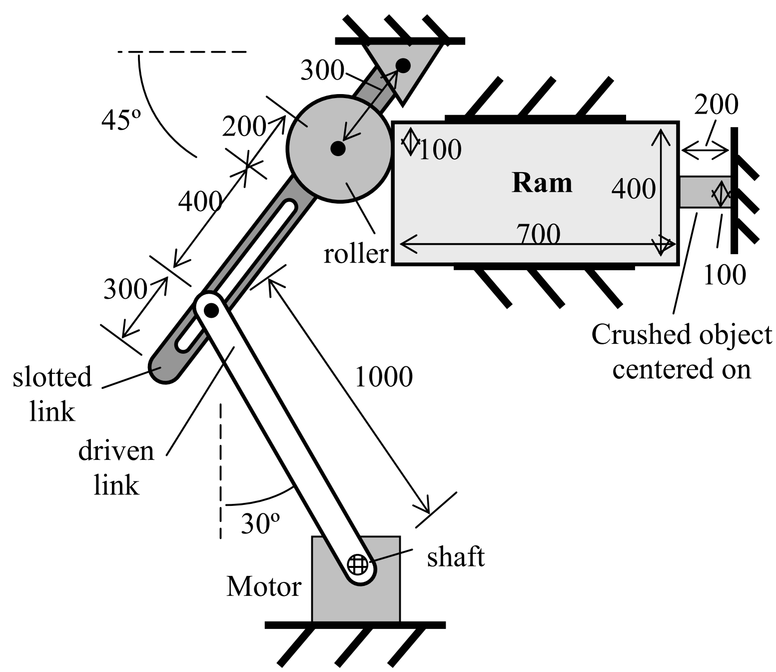

Figure 3. Exam problem #10, featuring multiple connected bodies, which had the highest correlation (0.62) with overall inventory scores. (Problem description is abbreviated.)

Comparison of sub-scores on specific concepts and examination scores

It is also of great interest to see if one can gain insight into student understanding of *specific concepts* from the inventory. Indeed, one can envision the inventory being used as a formative assessment tool if students could be fed back information on specific areas of weakness. Instructors could then, in principle, provide students with supplemental instruction which addressed the identified weaknesses.

Sub-scores on individual concepts are arrived at as follows. Each of 8 concepts has associated with it 3 to 5 questions. (Confirmatory factor analysis based on data from Spring 2004 suggested that grouping questions according to the anticipated concepts explained most of the variability[8].) Hence there is significant support for the idea that the test measures 8 distinct concepts.) The 8 concepts in the inventory are: Free body Diagrams (FBD), Conditions of Equilibrium (Equil.), Friction force and relation to Coulomb friction law (Friction), Static Equivalency (St.Eq.), Forces exerted by a Roller (Roller), Implications of Negligible Friction between Contact Surfaces (Neg.Fr.), Forces exerted by a Slot on a Pin (Slot), and Interpretation of Force Representations involving Variables and Vectors (Rep). All involved 3 questions, except Free body Diagrams has 5 questions, and Conditions of Equilibrium has 4 questions. For each student we determined the fraction of correctly answered questions in each concept.

In Table 4, we tabulate the concept sub-scores (average fraction correct) for each of the concepts, and the correlation coefficient between the sub-score and overall average score on the inventory. The high correlation for FBD may be because they represent 5 of the 27 questions. One notices that the Roller has a somewhat lower correlation. Failing to answer the roller questions seems to be more prevalent among both high and low scorers overall, compared with other concepts. As will be seen later, when we look at performance on examinations in more detail, however, the roller questions do seem to be representative of how well students solve problems involving rollers.

Table 4. Fraction of questions in each concept answered correctly and Pearson correlation coefficient (*P*) between overall score on inventory and individual concept sub-scores.

	FBD	Equil	Friction	St. Eq.	Roller	Neg.Fr.	Slot	Rep.
Fraction	0.841	0.745	0.708	0.649	0.759	0.605	0.900	0.818
Correl.	0.683	0.539	0.588	0.544	0.487	0.644	0.630	0.617

In Table 5, we show the correlation between the sub-scores on different concepts and average exam scores. Recall that the correlation between the overall inventory score and average exam score was 0.642, which is higher than in the correlations for individual concepts (no single one is higher than 0.475, and some are much lower). Thus, performance on examinations does seem to require a combination of concepts, at least as judged by the inventory.

Table 5. Correlation coefficients between the sub-scores on concepts addressed by inventory and average exam scores.

FBD	Equil	Friction	St. Eq.	Roller	Neg.Fr.	Slot	Rep.
0.475	0.216	0.372	0.449	0.162	0.344	0.378	0.435

Still, each concept should correlate to some extent with the exam scores, though this will depend on how well the inventory question captures the concept and on how frequently this concept is drawn upon in the solving exam problems. Several comments are worthwhile. There is a surprisingly poor correlation with equilibrium, which should be a major concept in Statics. Closer examination of results for individual questions in this concept reveals that 3 were answered correctly by nearly all students, and 1 by very few students, yielding the 0.75 fraction of correct responses. (For no other concept is there such a marked difference between the scores for questions addressing the same concept.) Thus, the set of questions on equilibrium fails to discriminate among this group of students. (Data from five universities in the Spring of 2004 showed that this set of questions did discriminate better among students in the larger group.) The weak correlation with Roller found in Table 5 is probably consistent with roller questions not correlating well with the inventory overall. It is also the case that there are few examination problems dealing with rollers, but that is also true of the concepts of slot and friction.

A more detailed analysis was conducted of aspects of performance on problems in the third examination, including the problem shown in Figure 3 (which correlated very well with the inventory scores overall). Specifically, we noted whether each student erred in the free body diagram involving the roller or the slotted member. An error was defined as taking the roller force to be anything but perpendicular to the rolled surface or taking the force between the slot and the pin to be anything but perpendicular to the slot. In each case, we compared the inventory scores of students who made the error with students who did not. Typical inventory questions addressing the roller and the slot are shown in Figures 4 and 5.

The dump is kept in the position shown by a roller, link and hydraulic cylinder. The coefficient of friction between the roller and the dump is 0.6.

What is the direction of the force exerted by the roller on the dump at the point of interest?

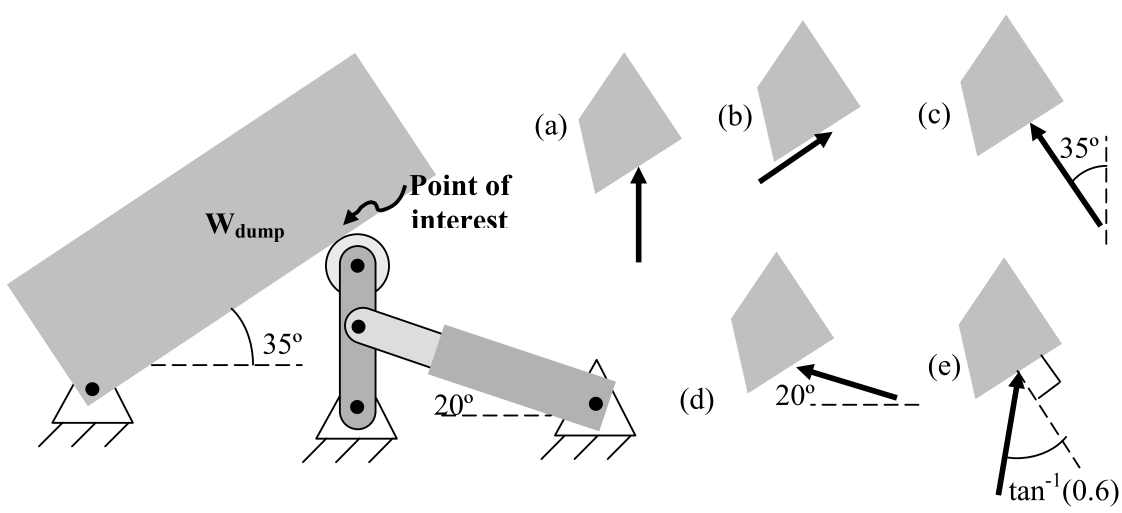

Figure 4. Inventory question addressing force between a roller and the rolled surface.

The mechanism is acted upon by the downward force shown. A spring acts on the slotted link.

What is the direction of the force exerted by the slot on the pin A?

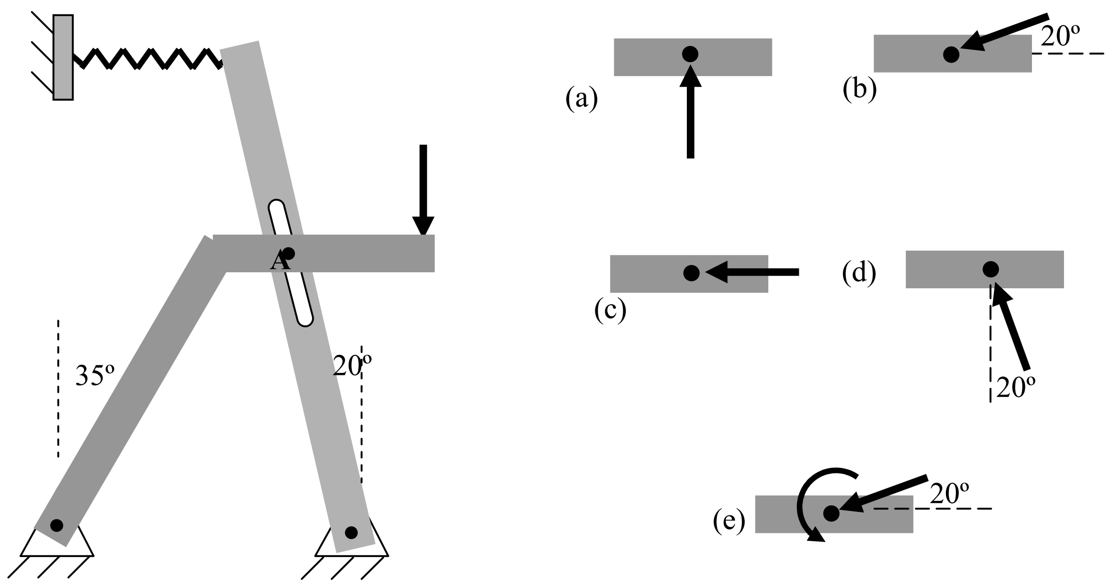

Figure 5. Inventory question addressing force between a pin and a slot.

In Table 6, we show the mean scores in each of the concepts and the inventory overall for students who erred (Err) and did not err (no Err) on the roller force. One can see that the mean scores of students who erred were in all cases less than those who did not. The significance of the difference in the means was quantified with a t-test, and the probability (p) that the two groups actually had the same means is also shown in the row labeled p. (A value of 0.000 represents a value less than 0.0005.) One can see that the probability is not below the 0.05 significance level for 5 of the concepts, but it is below for 3 concepts (including roller), as well as for the inventory overall. Thus, while we previously found the roller to be weakly correlated with exams scores overall, it is strongly correlated with performance on a problem drawing specifically on the concept of the roller.

Table 6. Comparison of inventory scores for students who erred and did not err in the direction of the force exerted by the roller in the exam question of Figure 3.

	N	FBD	Equil	Friction	St.Eq.	Roller	Neg.Fr.	Slot	Rep.	Overall
Err	47	0.813	0.739	0.652	0.532	0.645	0.546	0.872	0.709	0.700
no Err	50	0.868	0.750	0.760	0.760	0.867	0.660	0.927	0.920	0.816
p		0.192	0.767	0.127	0.001	0.002	0.075	0.199	0.000	0.000

A similar analysis was undertaken comparing students who erred and did not err in the force exerted by a pin on the slot. These results are shown in Table 7. Again, the differences in inventory sub-scores on the slot questions were significant for students who did and did not draw the force correctly in the exam question featuring the slot. However, it can be seen that those who erred and did not err had significantly different averages on most of the remaining concepts in the inventory as well.

Table 7. Comparison of inventory scores for students who erred and did not err in the direction of the force exerted by a pin on the slot in the exam question of Figure 3.

	N	FBD	Equil	Friction	St.Eq.	Roller	Neg.Fr.	Slot	Rep.	Overall
Err	33	0.739	0.712	0.667	0.495	0.646	0.455	0.778	0.667	0.654
no Err	64	0.894	0.762	0.729	0.729	0.818	0.682	0.964	0.896	0.814
p		0.002	0.258	0.42	0.003	0.023	0.002	0.001	0.000	0.000

We also investigated performance on a problem in exam 3 addressing friction between sliding blocks (Figure 6). This corresponds to problem number 12 in Table 3. A typical question on the inventory addressing friction is shown in Figure 7. One can see that the inventory question addresses the point that the friction force can be *less than* μN, (the friction coefficient times normal force), provided there is no relative motion between the bodies. The comparison with the inventory was done by grouping all students who scored 7.5 out of 10 on the exam problem of Figure 6 and those who scored below. (A score of 7.5 or above was typical when students recognized in at least one case that the friction force between the blocks *can be less than* μN.) The mean scores for these two groups on the friction sub-score of the inventory are shown in Table 8, as well as the probabilities that the two means are actually equal (using a t-test). Notice that the probabilities are above 0.05 for some of the concepts, but below for others. The performance of the two groups on the concept questions involving friction is clearly significantly different ($p = 0.001$), although it also different for some other concepts.

Blocks of three different weights have distinct friction coefficients between them and the table they rest on. The upper block is tied to a cord which is held fixed. Consider the force P to be increased just up to the level at which the lowest block slides to the right <u>at constant speed</u>.

(i) Take $\mu_2 = 0.4$. Determine the force P and whether or not the middle block (40 N) moves. Draw each block alone with <u>all</u> the normal and friction forces that act on it.

(ii) Repeat part (i), except now take $\mu_2 = 0.05$.

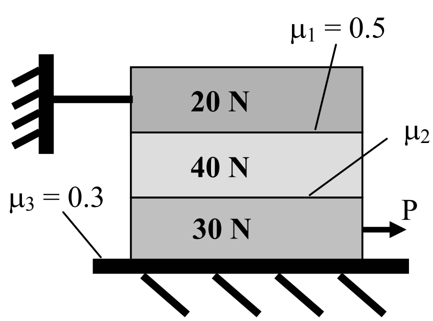

Figure 6. Exam problem #12 focusing on frictional slip. (Problem description is abbreviated.)

Table 8. Comparison of inventory scores for students scored above and below 7.5 out of 10 on the exam problem depicted in Figure 6.

	N	FBD	Equil	Friction	St.Eq.	Roller	Neg.Fr.	Slot	Rep.	Overall
<7.5	53	0.804	0.745	0.610	0.579	0.723	0.535	0.868	0.748	0.711
>=7.5	44	0.886	0.744	0.826	0.735	0.803	0.689	0.939	0.902	0.818
p		0.047	0.978	0.001	0.028	0.265	0.014	0.075	0.001	0.001

The blocks on the two sides are supported by two hands which also squeeze them together as shown. (The upward forces of the hands are represented with rollers.) The side blocks support the center block via friction. The friction coefficient between the blocks and the center block is 0.4. (Take this to be both the static and kinetic coefficient of friction). Gravity acts in the vertical direction.

What is the vertical component of force exerted by the left block on the center block?

(a) 2 N (b) 3 N (c) 5 N (d) 8 N (e) 16 N

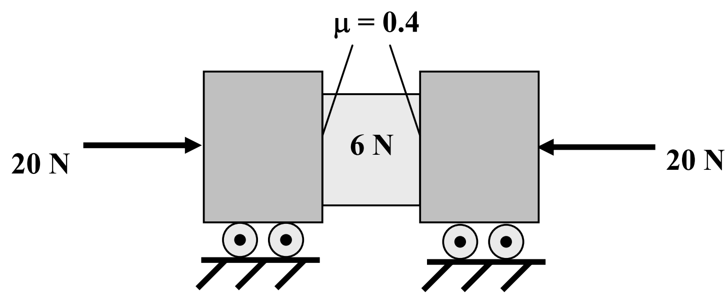

Figure 7. Inventory question addressing limits on friction force.

It is worth highlighting what was just observed. The means on inventory questions involving rollers were not significantly different among poor and good performers on exam problems involving friction, although these groups had very different means on inventory questions involving friction. Precisely the reverse was found for the exam problem demanding knowledge of the roller. Thus, there appears to be meaningful concept specific information that could be extracted from the inventory.

Potential For Using Inventory For Formative Assessment

As has been just reported, performance on individual concepts in the inventory does seem to have implications for the ability to solve different kinds of problems. Still, we have also seen that there are good correlations between scores on individual concepts and inventory scores overall. It would seem that, for many students, weaknesses tend to be distributed across concepts roughly equally. However, when a very pointed weakness in a single area is present, it might be helpful to have some means of identifying it. One way of doing this is looking for very low scores on individual concepts, particularly in students whose overall scores are high. Thus, rather than being sloppy or fuzzy on a concept, this would suggest a significant lack of understanding. One can easily identify such students from the results. In addition, however, studies involving in-depth interviews of students may be necessary to determine whether such students truly exhibit less understanding of the associated concept than other students. Some such studies are currently underway.

We conclude by describing one modest effort to take advantage of feeding back to students information on their performance on the inventory. Inventory scores overall and for each concept (and the correctness of each question) were emailed to all students. A special evening session was scheduled (one week before the end of the semester), at which the questions of the inventory were discussed concept by concept. Discussing questions by concept allowed students to single out for focus those questions and concepts which they answered incorrectly. Only 6 students attended the session.

The fourth examination was given one week after the optional session. This exam featured a problem that also involved multiple interconnected bodies, including both a slotted member and a roller. We tracked which students made errors with respect to the roller and slot forces, and compared with errors from the problem on the third examination. Students were divided into four categories related to erring and not erring with respect to the two exam questions. Of course, students had opportunities to improve their understanding for the fourth examination by consulting the solutions that were posted for the third examination. These results were tabulated in Table 9 for all 110 students taking the two examinations. In the case of the slot, few students made erred at all, and nearly all of the remaining ones improved. In the case of the roller, about half of the class erred to begin with, and of those about half improved. The performance of a non-negligible fraction of students worsened.

Table 9. Errors in two successive examination questions relating to forces on roller and slot.

	Never Erred	Erred on 1st, not 2nd	Erred on both	Erred on 2nd, not 1st
Slot	83 (75%)	23 (21%)	3 (3%)	1 (1%)
Roller	46 (42%)	28 (25%)	24 (22%)	12 (11%)

The performances of the 6 students who attended the optional session to review the inventory questions are depicted in Table 10. Their errors on the examination problems have been coded as follows: (1) never erred, (2) erred on 1st, not on 2nd, (3) erred on both, (4) erred on 2nd, not on 1st. Thus, of the 12 opportunities for learning, 6 had no need to learn, 3 learned, 2 did not improve, and 1 worsened. Though it is somewhat disappointing, it is hard to discern any patterns, and the sample is rather small in any event.

Table 10. Errors in two successive examination questions for students who participated in optional session reviewing inventory questions

	Student 1	Student 2	Student 3	Student 4	Student 5	Student 6
Post-test	24	22	16	16	16	11
Roller	1	3	1	2	3	4
Slot	1	1	1	1	2	2

Summary and Conclusions

We have reported on data from a multiple-choice Statics Concept Inventory which was administered to a sophomore mechanical engineering Statics class both before and towards the end of the course. The primary purpose has been to investigate the potential of using the inventory as a formative assessment tool to improve learning in the class. To this end, we have sought to determine whether scores on the inventory are consistent with other measures of performance in the class, in particular exam scores. Moreover, we have inquired as to whether the inventory sub-scores on specific concepts are consistent with aspects of exam scores that clearly draw upon those specific concepts.

We have found that the inventory scores overall correlate reasonably well with average exam scores (correlation of 0.642), and that the scores on exam problems that address multiple interconnected bodies correlate particularly well with the inventory. When the inventory was originally developed, this was identified as the class of problems that was most relevant.

Moreover, specific errors were tracked in this exam question for each student. Among students who made certain errors and those who did not, the inventory sub-scores on various concepts were found to differ. The concept most relevant to the error was always markedly different for the two groups; among the remaining concept sub-scores, some were not significantly different statistically and while others were. A similar comparison was made for performance on another exam problem, which was related to friction, with favorable findings. Thus, we conclude that performance on specific concept sub-scores of the concept inventory have implications for specific aspects of exam performance that apparently demand those concepts. A pilot study was also conducted in which all students were given concept-specific feedback on their performance in the inventory and then invited to an optional session to discuss questions in the inventory. Relatively few students participated in the session, and analysis of changes in their performance on the last exam yielded inconclusive results. Still, the correlations observed between inventory performance and exam performance suggest that the inventory scores ought to contain information about student understanding that could form the basis for identifying appropriate remedial instruction. This prospect continues to inspire future work in this area.

BIBLIOGRAPHIC INFORMATION

1. P. Black and D. William, "Assessment and Classroom Learning, "Assessment in Education, Vol. 5(1), 1998, pp. 7-73.
2. National Research Council, 2001, Knowing What Students Know: the Science and Design of Educational Assessment, J.W. Pellegrino, N. Chudowsky, and R. Glaser (Eds.), Washington, D.C., National Academy Press
3. National Research Council, 1999, How people learn: Brain, mind, experience and school, Committee on Developments in the Science of Learning, Bransford, J.D., Brown, A.L., Cocking, R.R. (Eds.), Washington, D.C., National Academy Press
4. Hestenes, D., Wells, M. and Swackhamer, G., "Force Concept Inventory", The Physics Teacher, Vol. 30, 1992, p. 141.
5. D. Evans, C. Midkiff, R. Miller, J. Morgan, S. Krause, J. Martin, B. Notaros, D. Rancor, and K. Wage, "Tools for Assessing Conceptual Understanding in the Engineering Sciences," Proceedings of the 2002 FIE Conference, Boston, MA.
6. Steif, P.S., "An Articulation of the Concepts and Skills which Underlie Engineering Statics," 34rd ASEE/IEEE Frontiers in Education Conference, Savannah, GA, October 20-23, 2004.
7. Steif, P.S., "Initial Data from a Statics Concepts Inventory," Proceedings of the 2004, American Society of Engineering Education Conference and Exposition, St. Lake City, UT, June, 2004.
8. Steif, P.S. and J.A. Dantzler, "A Statics Concept Inventory: Development and Psychometric Analysis" (under review by Journal of Engineering Education).

BIOGRAPHICAL INFORMATION

PAUL S. STEIF
Professor, Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, Pa

Degrees: Sc. B. 1979, Brown University; M.S. 1980, Ph.D. 1982, Harvard University.
Research area: solid mechanics and engineering education.