AC 2007-706: EFFECTS OF SEX AND ETHNICITY ON PERFORMANCE ON THE MATERIALS CONCEPT INVENTORY

Elliot Douglas, University of Florida

Dr. Elliot P. Douglas is Associate Professor of Materials Science and Engineering at the University of Florida. His educational research interests are in the areas of active learning techniques and critical thinking. He has been involved in faculty development activities since 1998, through the ExCEEd Teaching Workshops of ASCE, the Essential Teaching Seminars of ASME, and the NSF-sponsored SUCCEED Coalition. He is a member of the American Chemical Society, American Society for Engineering Education, and the American Educational Research Association. He received the Presidential Early Career Award for Scientists and Engineers, the Ralph Teetor Education Award from the Society of Automotive Engineers, and was named University of Florida Teacher of the Year for 2003-04.

Effects of Sex and Ethnicity on Performance on the Materials Concept Inventory

Abstract

This paper describes results on using the Materials Concept Inventory in an introductory materials course. The validity of the MCI is confirmed by correlation with student course grades and student self-assessment of understanding. However, the reliability of the instrument is low, and content analysis suggests that the low reliability is related to the inclusion of a large number of concepts within the instrument. Results by sex[i] show that women score lower than men, despite no difference in academic ability. Results for differences by ethnicity are ambiguous due to the low numbers of students in some of the groups. Overall, the results highlight the importance of developing concept inventories by maintaining a narrow focus on a specific area of conceptual understanding within a particular field. The results also point to the potential role that the context of the items has on performance, although considerably more work is needed in this area.

Introduction

There continues to be considerable interest within engineering education to develop innovative modes of teaching that will improve student outcomes across a wide range of learning objectives.[1] In order to appropriately assess the efficacy of these techniques a variety of assessment techniques are needed. For assessment of content knowledge, concept inventories provide a means to assess knowledge in specific content domains. The first concept inventory developed was the Force Concept Inventory, and since then concept inventories have been developed for statics,[2] strength of materials,[3] engineering mechanics,[4] electrical circuits,[4] thermal and transport sciences,[5] and materials.[6,7] As the name implies, concept inventories are designed to test for fundamental concepts within a domain, as opposed to memorized facts. Typically concept inventories are developed by identifying misconceptions held by students, and constructing distracter items based upon these misconceptions.

Although concept inventories are generally considered to be neutral towards sub-populations (e.g. men vs. women), there is some evidence that this is not true. A persistent bias by sex has been found for the Force Concept Inventory, with men scoring higher than women even when controlling for educational background.[8,9] McCullough has investigated this effect further by creating a modified version of the Force Concept Inventory in which items with stereotypically male-oriented contexts (sports, rockets, etc.) were replaced with stereotypically female-oriented contexts (babies, kitchens, etc.).[10] Although there were differences in both men's and women's responses on the revised instrument, the pattern of responses did not provide a clear indication of what the reason for the changes might be. Thus, while it is clear that the Force Concept Inventory has a sex bias, the exact nature of that bias has not been identified.

This paper provides some preliminary data from a larger study on the use of different pedagogies in the introductory materials course within the engineering curriculum. This paper focuses on the reliability and validity of the Materials Concept Inventory, and especially differences in performance by sex and ethnicity. The data comes from two sections of the course, taught by the

author in an "active lecture" format. That is, the predominant teaching style was in lecture mode, but with considerable use of active teaching methods. The results described here, while specific to the MCI, suggest some general guidelines for the development and use of concept inventories.

Methodology

The particular experimental approach used was a quasi-experimental study using a convenience sample consisting of the author's sections of the introductory materials course in fall 2005. Students in the course were offered the opportunity to obtain extra credit towards their course grade by participating in the study. An alternative extra credit assignment was offered to students who did not wish to participate in the study. A total of 154 students participated to some extent, out of 157 students enrolled in the course. The IRB-approved research process specified that the author would not know which students were participating until the very end of the semester in order to avoid any appearance of bias in grading of the students' work. Thus, all data collection was done by the teaching assistant for the course. The teaching assistant did grade homework, but was not involved with grading of exams or assigning final course grades.

Student understanding was measured using the Materials Concept Inventory (MCI) developed by Krause et al.[6,7] This is a 30 item multiple choice instrument designed to assess students' level of conceptual knowledge in an introductory materials science class. The validity of this instrument has been established through its construction, which was done by expert evaluation of topics and writing of questions, use of student open-ended quizzes to develop distracters, and use of student focus groups to further refine ambiguous questions and answers. Reliability of this instrument has not been discussed in the literature. For this study Cronbach's α was calculated as a measure of reliability.[11]

In order to assess students' beliefs about their learning, the instrument used was the Student Assessment of Learning Gains,[12] an online instrument designed to focus student assessment on how the pedagogy of the class affected their learning gains, as opposed to issues of teacher performance or the extent to which students "liked" the class. The validity of this instrument has been established by comparison of the SALG instrument to written comments taken from both the SALG and other instruments.[13] Reliability has also been discussed, although a quantitative estimate of reliability has not been reported. Cronbach's α was calculated as a measure of reliability. For this study, a few of the questions were modified to ask about specific content associated with the class. For this study, only those questions related to course content were analyzed. A full analysis of the SALG will be described in a future publication.

Additional demographic and performance data was also collected from university records, course grading, and student self-reporting. Data collected was the students' sex, ethnicity, current grade point average, SAT scores, major, and final average in the introductory materials course. The IRB-approved informed consent form signed by the students specified that they were giving permission for this data to be collected.

In order to analyze the MCI, a pre-test/post-test design was used. Calculating the gain in the MCI score has an advantage over a simple post-test design in that it controls for prior knowledge. For the SALG, a post-test only design was used since this instrument asks about characteristics of the course. A t-test was used to test for differences between various groups. The data was also analyzed for the effect of student characteristics on various measures by conducting ANOVA on sub-groups of the populations. All significance testing was conducted at $p<.05$ using 2-tailed tests unless otherwise indicated.

Results

Table 1 provides descriptive statistics for the entire student sample as well as sub-groups. Two-tailed t-tests showed no differences between males and females on any of these variables (GPA: $t(99) = -0.254$, $p>.05$; SAT verbal: $t(97) = 0.168$, $p>.05$; SAT quantitative: $t(97) = -1.069$, $p>.05$; course grade: $t(118) = 0.588$, $p>.05$). Similarly, one way ANOVA shows no differences by ethnicity (GPA: $F(4,96) = 1.661$, $p>.05$; SAT verbal: $F(4,94) = 0.679$, $p>.05$; SAT quantitative: $F(4,94) = 2.322$, $p>.05$; course grade: $F(4,115) = 2.379$, $p>.05$).

Table 1: Descriptive statistics for the entire student sample and various sub-groups. Numbers in parentheses are number of students (N) followed by the standard deviations. The values of N reflect those students for whom data was available.

	GPA	SAT verbal	SAT quantitative	Course grade
All students	3.20 (120,0.53)	593 (117,82)	645 (117,75)	81.0 (133,10.4)
Male	3.27 (77,0.53)	593 (77,85)	649 (77,80)	81.5 (95,10.6)
Female	3.24 (24,0.51)	596 (22,66)	630 (22,50)	82.8 (25,7.8)
White	3.32 (59,0.48)	602 (62,80)	653 (62,74)	83.3 (75,8.8)
Hispanic	3.27 (20,0.55)	573 (17,72)	625 (17,73)	79.6 (22,9.9)
African American	2.79 (7,0.48)	584 (7,83)	591 (7,68)	72.3 (7,16.2)
Asian	3.22 (8,0.51)	603 (7,68)	691 (7,49)	80.5 (9,11.8)
Other	3.27 (7,0.74)	563 (6,131)	623 (6,80)	83.4 (7,10.8)

The MCI means (standard deviations in parentheses) for the entire sample of students were 11.5 (0.3) for the pre-test, 15.2 (0.3) for the post-test, and 3.7 (0.3) for the gain. Thus, on average, students showed a 32% improvement from the pre-test to the post-test. A two-tailed t-test shows that the difference between the pre-test and post-test scores is significant ($t(116) = -11.92$, $p<.05$) with a large effect size ($r = 0.74$). The reliability of the MCI was low. Cronbach's α was .54 for the pre-test and .58 for the post-test. In general, a value of .7 is considered the minimum acceptable. A further discussion of the reliability of the MCI is given in the next section.

Table 2 show the significant correlations obtained between MCI scores and various demographic variables. MCI scores are positively correlated with general academic ability, as well as the grade obtained in the introductory materials course. The positive correlations between the MCI scores and the course grade, and the increase in that correlation from the pre-test to the post-test, attests to the validity of the MCI. Additional measurement of validity comes from the student self-assessment of content knowledge as measured by the SALG. The reliability for the entire SALG instrument was $\alpha = .92$. Table 3 shows the significant correlations between objective

measures of content knowledge and student self-assessment. All questions on the SALG resulted in non-parametric distributions of scores, and so correlations are reported as Spearman's rho. Positive correlations are found for some, but not all, of the content areas. Correlations are lower for the MCI post-test than for the course grade, which may reflect the fact that the MCI was not created specifically for this course.

Table 2: Correlations between MCI scores and demographic variables. Pearson's correlation coefficient is used when both variables are parametric. Spearman's rho is used when at least one variable is non-parametric. Empty cells in the table indicate that the correlation was non-significant (one-tailed, $p>.05$).

	GPA	SAT verbal	SAT quantitative	Course grade
MCI pre-test	------	.43*	.35*	.17†
MCI post-test	.33*	.37*	.39*	.37†

*Pearson's correlation coefficient.
†Spearman's rho.

Table 3: Correlations between objective measures of course content knowledge (rows) and student self-assessment of knowledge in specific content at the end of the course (columns). Correlations are given as Spearman's rho. Empty cells in the table indicate that the correlation was non-significant (one-tailed, $p>.05$).

	Phase diagrams	Mechanical properties	Crystal structures	Diffusion	Kinetics	Corrosion
Course grade	.30	.26	.22	.21	------	------
MCI post-test	.25	.16	.17	------	------	------

Table 4: MCI results by sex. Numbers in parentheses are number of students (N) followed by the standard deviations. The values of N reflect those students for whom data was available. MCI gain may not equal the difference between the pre-test and post-test scores due to round-off error.

	MCI pre-test*	MCI post-test*	MCI gain
Male	11.9 (87, 3.6)	15.7 (89, 4.0)	3.8 (82, 3.4)
Female	10.3 (25, 2.7)	13.8 (25, 3.7)	3.6 (24, 3.0)

* Significant at $p<.05$ using a 2-tailed t-test.

Discussion

Validity of the MCI has been previously assumed based on its manner of construction.[6,7] General concepts for the instrument were obtained through faculty input. Specific misconceptions and potential distracters were developed through student interviews and quizzes (both open-ended and multiple choice). This study provides further evidence for its validity. Results show positive correlations between the MCI post-test and the grade obtained by the students in the course, and a significant increase in score from the pre-test to the post-test. It is assumed that this increase is due to the students participating in the course, although it is possible,

Table 5 MCI results by ethnicity. Numbers in parentheses are number of students (N) followed by the standard deviations. The values of N reflect those students for whom data was available. MCI gain may not equal the difference between the pre-test and post-test scores due to round-off error.

	MCI pre-test	MCI post-test*	MCI gain
White	12.0 (72, 3.5)	16.1 (73, 3.8)	4.0 (71, 3.3)
Hispanic	10.7 (18, 3.1)	14.7 (21, 3.3)	4.2 (17, 3.0)
African American	10.0 (6, 2.5)	10.7 (6, 3.0)	0.4 (5, 4.2)
Asian	9.8 (8, 1.8)	12.4 (9, 3.0)	2.6 (8, 3.4)
Other	11.6 (7, 5.3)	17.0 (5, 5.3)	4.6 (5, 2.2)

*Significant at $p<.05$ by one way ANOVA.

but not likely, that other factors may have contributed to that gain. The overall gain between pre-test and post-test in this study (32%) is comparable to those seen in other studies: the statics concept inventory showed a 92% gain from pre-test to post-test,[2] the force concept inventory a 20% gain,[9] and previous administrations of the MCI showed gains of 15-20% in lecture classes and 38% in an active learning class.[6] Given that this course was taught in an active lecture format, the gain observed is not surprising and is consistent with most other studies (excepting the statics concept inventory). An additional test of validity comes from the students' self-assessment of their knowledge in different content areas associated with the course. Positive correlations are found between the post-test and three of the six content areas asked about on the SALG. A detailed discussion of the content of the MCI is given below, but for here we note that of the three content areas for which there was no correlation with the post-test score, one (kinetics) had no items on the MCI and one (corrosion) had only one item on the MCI. Thus, lack of correlation for these two SALG items with the post-test would not be surprising. In comparison, the areas of phase diagrams, mechanical properties, and crystal structures all have at least two items on the SALG, which allows those areas to contribute more strongly to the overall score. In the case of diffusion, for which there are also two items on the SALG, it is somewhat surprising that the correlation with the post-test is non-significant. There are two possible reasons for this difference. One is that is may simply reflect students' lack of confidence with that particular topic. Also, at least in the context of this course, the interrelations among crystal structures, phase diagrams, and mechanical properties are made clear, while diffusion stands alone as a topic. Thus, there may be some interactive effects among those first three topics that strengthen the individual correlations.

Results for reliability of the MCI suggest that there are some issues with the instrument. In order to understand the low reliability, a content analysis was conducted by the author, in which the various items were categorized by the concept underlying that item. Fourteen different categories were created, with any category containing from one to five items. Table 6 provides a summary of this content analysis. Although other researchers would likely categorize the items in a slightly different manner than given in Table 6, it seems clear that the MCI covers a wide range of concepts, with only a few items (or sometimes only one item) for each concept. Thus, the low reliability is not surprising. In contrast, other concept inventories maintain a tighter focus on the concepts being assessed. For example, the statics concept inventory focuses on four concepts, all of which are related to forces acting on bodies.[2] This focus on only one or a few concepts within these concept inventories leads to their higher reliabilities.

Table 6: Content analysis of the Materials Concept Inventory

Concept	**MCI item numbers**
Diffusion	1, 11
Bonding	2, 5
Moles	3
States of Matter	4
Crystals/Glasses	6, 8
Thermal Properties	7
Crystal Structures	9, 10
Corrosion	12
Electrical Properties	13, 14, 15
Phase Diagrams	16, 17
Strengthening Mechanisms	18, 19, 20, 21, 24
Mechanical Properties	22, 23, 25, 26
Polymers	27, 28, 29
Composites	30

The low reliability of the MCI suggests potential problems in using it as a measurement instrument. A low reliability corresponds to low internal consistency among items, which leads to a larger variance due to random error. Thus, larger differences between groups are needed to identify significant differences compared to an instrument with high reliability. Within this limitation, however, the MCI can provide a useful measure of student achievement in the introductory materials course.

Use of the MCI provides a means to compare achievement for different sub-groups. Results on differences by sex show that women consistently score lower, despite there being no difference in general academic ability or grade in the course between men and women. There is also no difference in the gain score between men and women. Thus, despite the fact that women score lower, they have the same abilities, do just as well in the course, and learn the same amount as men. These results all seem to point to some type of bias against women in the MCI.

Results in the literature suggest that the way items are worded can have an effect on the differences measured between men and women. McCullough reviews some of the literature on context effects.[10] As one extreme example, she relates a situation where cultural differences prevented students from answering a physics question because it depicted a situation they saw as extremely rude. Rennie and Parker have described differences by sex in preference for context on physics questions at the high school level,[14] with more girls than boys stating that questions with real-world context are easier to understand. However, the reasons for these differences are not always apparent. For example, in McCullough's study using the revised Force Concept Inventory, the reduction in the sex bias when stereotypically female questions are used is caused, at least in part, by men performing worse rather than women performing better.[10] As Rennie and Parker point out, it is difficult to specifically understand the effect of sex, given all the other factors (culture, language, familiarity with different contexts, students' preferences) that may also affect performance.[14] Nevertheless, there does appear to be a difference in performance on the MCI between men and women that can not be explained simply based on academic ability.

Results by ethnicity show no significant differences across the different groups for the MCI pre-test or gain score. However, we note that there are relatively large differences in the gain scores (e.g. 4.2 for Hispanic vs. 0.4 for African-American). The lack of any significant difference as determined by one way ANOVA may be due to the low numbers of students in some of the groups (e.g. 17 Hispanics vs. 5 African-Americans). Additional data is needed before any firm conclusions can be reached.

Conclusions

The results of this study confirm the validity of the MCI, although some concepts covered in this particular version of the course are not assessed on the MCI. Thus, some care should be taken when using the MCI as a diagnostic test to ensure that the MCI actually measures the desired concepts. Reliability of the MCI is low, which seems to be related to the range of concepts that are assessed. In comparison to other concept inventories, the MCI covers a larger number of topics. This may reflect the way in which the content domains are taught within the engineering curriculum. Each of the concept inventories is constructed roughly around a particular course (even if they were not explicitly constructed that way), e.g. statics, thermal sciences, etc. The wider range of concepts in the MCI may reflect the nature of the introductory materials course as more of a survey course than other courses. Thus, this study highlights the importance of developing concept inventories by maintaining a narrow focus on a specific area of conceptual understanding within a particular field.

Despite the low reliability, the MCI can still serve as a useful means of assessing performance across the wide range of concepts typically taught within the introductory materials course. The results presented here on sex are consistent with reports in the literature that describe a similar effect for the Force Concept Inventory. While the exact cause is not clear, it appears that it may be related to the context of the items, whether they are male-oriented questions, female-oriented, or context-neutral. Further work is needed to address this question. Results regarding ethnicity are more ambiguous due to the small size of some of the groups.

Bibliography

(1) Prince, M. J.; Felder, R. M. "Inductive teaching and learning methods: Definitions, comparisons, and research bases", *Journal of Engineering Education* **2006**, *95*, 123-138.

(2) Steif, P. S.; Dantzler, J. A. "A statics concept inventory: Development and psychometric analysis", *Journal of Engineering Education* **2005**, *94*, 363-371.

(3) Richardson, J.; Steif, P.; Morgan, J.; Dantzler, J. "Development of a concept inventory for strength of materials", *Proceedings - Frontiers in Education Conference* **2003**, *1*, T3D29-T23D33.

(4) Streveler, R.; Geist, M.; Ammerman, R.; Sulzbach, C.; Miller, R.; Olds, B.; Nelson, M. "Identifying and investigating difficult concepts in engineering mechanics and electrical circuits", *Proc. ASEE Ann. Conf.* **2006**.

(5) Miller, R.; Streveler, R.; Olds, B.; Chi, M.; Nelson, M.; Geist, M. "Misconceptions about rate processes: Preliminary evidence for the importance of emergent conceptual schemas in thermal and transport sciences", *Proc. ASEE Ann. Conf.* **2006**.

(6) Krause, S.; Decker, J. C.; Niska, J.; Alford, T.; Griffin, R. "Identifying student misconceptions in introductory materials engineering classes", *Proc. ASEE Ann. Conf.* **2003**.

(7) Krause, S.; Tasooji, A.; Griffin, R. "Origins of misconceptions in a materials concept inventory from student focus groups", *Proc. ASEE Ann. Conf.* **2004**.
(8) McCullough, L. "Gender differences on multiple-choice tests", paper presented at the national meeting of the American Association of Physics Teachers, 1996, http://physics.uwstout.edu/staff/mccullough/physicseduc.htm, accessed January, 2007.
(9) McCullough, L.; Crouch, C. H. "Gender, educational reform, and instructional assessment: Part 1", paper presented at the national meeting of the American Association of Physics Teachers, 2002, http://physics.uwstout.edu/staff/mccullough/physicseduc.htm, accessed January, 2007.
(10) McCullough, L. "Gender, context, and physics assessment", *J. Inter. Women's Studies* **2004**, *5*, 20-30.
(11) McMillan, J. H.; Schumacher, S. *Research in education*; Addison Wesley Longman, Inc.: New York, 2001.
(12) "Student assessment of learning gains"http://www.wcer.wisc.edu/salgains/instructor/, accessed Jnauary, 2007.
(13) Seymour, E.; Wiese, D. J.; Hunter, A.-B.; Daffinrud, S. M. "Creating a better mousetrap: On-line student assessment of their learning gains", *paper presented at American Chemical Society National Meeting* **2000**, http://www.wcer.wisc.edu/salgains/ftp/SALGPaperPresentationAtACS.pdf.
(14) Rennie, L. J.; Parker, L. H. "Equitable measurement of achievement in physics: High school students' responses to assessment tasks in different formats and contexts", *Journal of Women and Minorities in Science and Engineering* **1998**, *4*, 113-127.

[i] Although the terms "sex" and "gender" are often used interchangeably, the standard definitions are that "sex" refers to male and female defined by physical attributes, while "gender" refers to masculine and feminine defined by culturally derived roles. Since in this study we are simply considering male and female as defined by physical attributes and not considering particular roles, the term "sex" is the appropriate one to use.