



## Multiple Choice Learning Assessments for Intermediate Mechanical Engineering Courses: Insights from Think-Aloud Interviews

**Dr. Matthew J Ford, Cornell University**

Matthew Ford is currently a Postdoctoral Teaching Specialist working with the Cornell Active Learning Initiative. His background is in solid mechanics.

**Dr. Hadas Ritz, Cornell University**

Hadas Ritz is a senior lecturer in Mechanical and Aerospace Engineering, and a Faculty Teaching Fellow at the James McCormick Family Teaching Excellence Institute (MTEI) at Cornell University, where she received her PhD in Mechanical Engineering in 2008. Since then she has taught required and elective courses covering a wide range of topics in the undergraduate Mechanical Engineering curriculum. In her work with MTEI she co-leads teaching workshops for new faculty and assists with other teaching excellence initiatives. Her main teaching interests include solid mechanics and engineering mathematics.

**Dr. Benjamin Finio, Cornell University**

**Prof. Elizabeth M. Fisher, Cornell University**

Elizabeth M. Fisher is an Associate Professor in the Sibley School of Mechanical and Aerospace Engineering at Cornell. She received her PhD from U.C. Berkeley.

# **Multiple Choice Learning Assessments for Intermediate Mechanical Engineering Courses: Insights from Think-Aloud Interviews**

## **Abstract**

Concept inventories (CIs)—validated tests based on carefully-articulated models of conceptual knowledge in a field—have been developed for many introductory STEM courses such as Physics / Mechanics, Statics, Chemistry, and Electricity and Magnetism. CIs can be powerful research tools for measuring students’ progression towards expert-level thinking, but can be difficult to develop for intermediate courses where domain-specific knowledge, problem-solving strategies, and technical fluency are important learning goals alongside conceptual frameworks. For such intermediate courses, it is still valuable to develop high-quality, multiple-choice tests to measure students’ progress towards course learning objectives or to assess the efficacy of instructional interventions.

We describe the development process and early results for multiple-choice learning assessments (MCLAs), drawn from a mix of pre-existing instruments and original content, for four 300-level mechanical engineering courses taken in the junior year: Fluid Mechanics, Mechanics of Materials, System Dynamics, and Mechatronics. We conducted a series of 16 “think-aloud” interviews with undergraduate students with a range of prior experience with each subject. Think-aloud interviews provide rich, qualitative data about student thought processes which is not available from multiple-choice or even free-response data. Students use a variety of problem-solving strategies including application of memorized formulas, identification with personal experience, mental simulation, strategic elimination, and reverse psychology of the presumed test author. We highlight some challenges in developing MCLAs, including naively-designed questions which admit correct answers with incorrect reasoning, and schematic diagrams which may unintentionally cue irrelevant concepts.

Three of the assessments were delivered as low-stakes quizzes in large-enrollment, lecture-based courses at a private R1 university. Results from the large sample show that many of the student-held misconceptions identified during the interviews are widely held even after instruction. Assessment scores are moderately correlated with final exam scores ( $0.30 < r < 0.62$ ) and course grades ( $0.33 < r < 0.49$ ). Data from the pilot tests suggest possible further enhancements to the assessments.

## Introduction

The practice of teaching is intimately tied to the problem of educational measurement. Effective teaching requires continuous and timely feedback about what students are learning and what concepts they are struggling with. The type of assessment used depends on what is being measured, and how and by whom the data will be used. A multiple-choice learning assessment (MCLA) is an easily-graded test instrument designed to measure a student's mastery of important key concepts or skills. MCLAs can be used to probe students' prior knowledge, assess the efficacy of instructional changes, help students gauge their progress towards course goals, and prompt student questions about common misconceptions.

We draw a distinction between an MCLA and a concept inventory (CI). A CI is based on a clearly-articulated set of concepts (often narrowly-defined) or expert modes of thinking in a field [1]. Some notable examples include the Force Concept Inventory [2], the Statics Concept Inventory [3], and the Brief Electricity and Magnetism Assessment [4]. In contrast, an MCLA is based on a set of learning objectives specific to a particular course. Concept Inventories are frequently used as pre- and post-tests to measure learning gain. Items in a CI should be understandable to someone with no exposure to the relevant coursework so that an incorrect answer can be attributed to misunderstanding or misapplication of the underlying concept rather than confusion about the language or imagery. Therefore CI creators are very careful to avoid unnecessary jargon or schematic representations. In most intermediate and advanced subjects, familiarity with specific jargon, components, and visual representations are important learning outcomes and should therefore appear on an MCLA, even though these elements may make the assessment less applicable as a pre-test.

For a learning assessment to be considered valid, it must be established that it actually measures what it purports to measure. The ability to interpret and draw conclusions from testing results depends on the validity of the instrument. For a learning assessment specifically, a high score should indicate a high level of mastery of the course learning outcomes. Direct observation of student thinking is an important step in developing and validating a learning assessment. After gathering detailed evidence from a small sample, the learning assessment should be tested on a larger audience to gather statistical evidence.

In this study we focus on the insights and evidence that can be drawn from observation of student reasoning using "think-aloud" interviews. We begin by briefly describing the development process for learning assessments in four courses: Fluid Mechanics, Mechanics of Materials, Mechatronics, and System Dynamics. Next, we give specific examples of insights drawn from think-aloud interviews and how they relate to the validity of the assessment. Finally, we present statistical results from three of the assessments administered as online quizzes in Fall 2019.

## Methods

### *Development process for learning assessments*

We began by articulating a set of between 20 and 30 learning objectives for each course covering a broad range of appropriate levels of Bloom's taxonomy [5]. Next, we developed conceptual questions which demonstrate the abilities listed in our learning objectives. (Not all learning

objectives can be appropriately measured with an MCLA. Skills requiring synthesis or evaluation were not addressed in our instrument.) We drew on several existing sources [6, 7, 8, 9, 10] for questions and developed others from scratch.

After receiving feedback from multiple faculty members on the content of the assessments, we conducted a series of think-aloud interviews with students. Interview subjects were recruited through email and an advertisement posted in an academic building. Subjects were given a \$20 Amazon gift card for each interview. The data collection and methods were approved by the Cornell Institutional Review Board under protocol #1708007347.

After another round of revision based on the results of the think-aloud interviews, the assessments were released to the entire department faculty for comment. After further revisions, three of the assessments (Fluid Mechanics, Mechanics of Materials, and Mechatronics) were administered as online quizzes, graded for participation only.

### *Think-aloud interviews*

Think-aloud interviews were conducted by the first author, who is well-versed in the subject matter but had no prior contact with the interviewees. After explaining the purpose of the think-aloud, the interviewer explained the parameters around data privacy and obtained the subject's consent to audio-record the interview.

The subject was asked to think aloud while solving each problem, starting by reading the problem statement out loud. If the subject fell silent for several seconds, the interviewer prompted the subject by repeating the subject's last few words with an upwards inflection, giving a short vocalization such as "mm-hmm?" or "OK?", or by requesting the subject to "please think aloud." Occasionally the interviewer probed further by asking questions such as "can you clarify that?" or "why did you rule out the other options?" after the subject had selected an answer.

Most items were given to the subject as multiple-choice problems. Some items were given as open-ended questions without answer choices. In a few items, subjects were asked to sketch their answer. The open-ended questions were used to identify student misconceptions to generate tempting distractors for multiple-choice questions.

### **Insights from think-aloud interviews**

To establish the validity of a learning assessment question, it is important to verify that a correct answer is associated with accurate or desirable reasoning ("true positive"), and that an incorrect answer is associated with inaccurate or undesirable reasoning ("true negative"). Think-aloud interviews can help identify questions with a risk of false positives or false negatives by revealing students' thought processes.

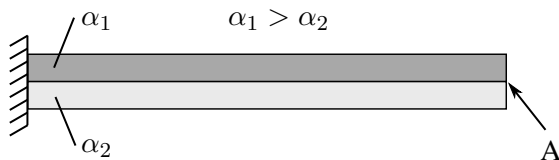
### *False positives*

A false positive results when incorrect reasoning nevertheless leads to the correct answer. One obvious possibility is that a student without knowledge guesses correctly. On a test which is not graded for correctness, the risk of guessing can be mitigated by providing an option to say "I'm

not sure.” We find that students are willing to select this option under low-stakes testing conditions.

A second type of false positive may result when a students’ flawed reasoning leads to an answer which is not one of the available choices, causing them to re-evaluate. The question shown in Box 1 is meant to test whether students correctly identify curvature caused by mismatched thermal strains. One subject initially failed to predict bending, but noted that they would have chosen the (incorrect) answer “Directly to the right” if it had been an option. They ultimately selected “Down and to the right” (correct) based on the reasoning that the greater transverse expansion in the top strip would push down on the bottom strip. The options were replaced with “Mostly to the left,” “Mostly to the right,” etc. A more sophisticated analysis would show that the vertical displacement due to bending is much larger than the horizontal displacement due to stretching and therefore “Mostly down” is a more appropriate answer than “Down and to the right” anyway.

The composite beam below is made of two bars with different thermal expansion coefficients bonded together. The material on the top has a larger coefficient of thermal expansion than the material on the bottom.



The beam is slowly heated up. In which direction does point A move *initially*?

Original options:	Modified options:
(a) Up and to the left.	(a) Mostly to the left.
(b) Up and to the right.	(b) Mostly to the right.
(c) Down and to the left.	(c) Mostly up.
(d) Down and to the right.	(d) Mostly down.
(e) There is not enough information to decide.	(e) There is not enough information to decide.

Box 1: Mechanics of Materials concept question, before and after revision. The absence of the student’s desired option “To the right only” led them to give the correct answer despite flawed reasoning.

A third type of false positive may result when the correct answer choice can be arrived at through incorrect reasoning. As a simple example, a problem which asks students to calculate the area of a square with side length of 4 units will not distinguish between correct reasoning and flawed reasoning by students who confuse area and perimeter. While it is sometimes possible to predict the types of incorrect reasoning strategies students may use, think-aloud interviews can give valuable and surprising insights.

The question shown in Box 2 is meant to test whether students understand that fracture strength varies inversely with the square root of the crack length (the course includes a module on introductory linear-elastic fracture mechanics). Initially, the answer options were divided into

equal intervals including and spanning 50% to 200%. Options C and D, while apparently non-intuitive, were included in case students remembered that there is an equation relating stress (at the crack tip) with crack length, but incorrectly interpreting this stress as the strength, rather than the far-field applied stress. During the interview, one student reasoned that the fracture strength of the plate should depend on the remaining cross-sectional area of the plate rather than the crack length.

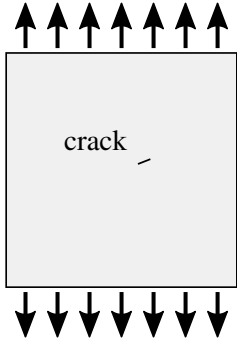
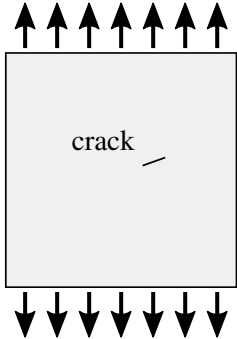
Student: So because the crack is small, it's going to be fairly close to 100 percent, but it's not going to be 0 percent because it's not a full-blown crack, it's only double the size. And because of that, it needs to be relatively close to 100 percent of the fracture strength of A, probably like 90-something.

Interviewer: Can you clarify what you mean by a full-blown crack?

Student: A full blown- crack as in... 0 percent of the fracture strength of A would be a crack that's fully across the [plate].

The answer option “Between 50% and 100%” is consistent with the desired answer (71%) and also with the student’s incorrect reasoning based on remaining cross-sectional area (90-something percent). The answer options were updated so that specific misconceptions map to different options.

A large plate with a small crack is loaded in tension.

If the crack length increases by a factor of 2, what is the new fracture strength of the plate?

Original options:	Modified options:
(a) About 50% of the original strength.	(a) About 50% of the original strength.
(b) Between 50% and 100% of the original strength.	(b) Between 50% and 90% of the original strength.
(c) Between 100% and 200% of the original strength.	(c) Between 90% and 100% of the original strength.
(d) About 200% of the original strength.	(d) About the same.
(e) There is not enough information to decide.	(e) There is not enough information to decide.

Box 2: Mechanics of Materials concept question with original and revised answer options.

### *False negatives*

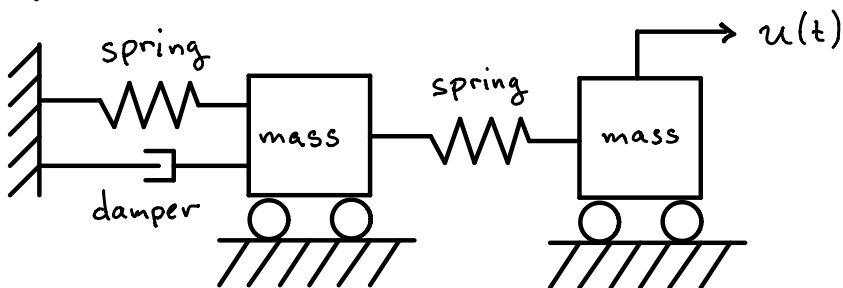
A false negative occurs when a student chooses an incorrect or undesirable answer choice despite demonstrating correct or desirable reasoning. A false negative may occur when the problem statement is not clear to the student. Care must be taken when developing an assessment meant to be used as a pre-test that the interpretation of the questions does not rely on an understanding of specialized terminology or visual representations. However, such questions may be appropriate for post-tests, as a working knowledge of common visual schematics may be an important learning outcome for the course.

Box 3 shows a problem from System Dynamics that relies on fluency in visual representations. An expert immediately recognizes the bent arrow attached to the right cart as an indication of the direction of the applied displacement history, which is visually distinct from an applied force. One student who had not yet taken the course was led astray by the shape of the arrow:

Student: Oh, it's interesting that the... that it's like an arrow that's like up and to the right. So that makes me think that we have two degrees of freedom because it can go up and only back down or to the right and then like back to the left. So I'm going to go with two... just because of the direction of the position function is in.

This type of error is generally not a concern if given as a post-test, provided that students have had sufficient exposure to the visual representations used and the instructor is independently confident that errors cannot be ascribed to unfamiliar notation. If given as a pre-test, however, the assessment should either independently test fluency with visual representations, or avoid them altogether.

The system below has a time varying position  $u(t)$  imposed on one cart as shown. How many degrees of freedom does the system have?



(a) 1   (b) 2   (c) 3   (d) 4

Box 3: System Dynamics concept question relying heavily on canonical schematic representations.

### *Probing student misconceptions with open-ended questions*

One of the most important uses of think-aloud interviews is to identify, and give specific language to, student misconceptions. Experienced teachers can generally predict what misconceptions are commonly held amongst students entering their course, but it can be challenging to identify the precise language or description that best captures students' thinking. Posing candidate questions as open-ended during interviews allows us to capture students' misconceptions in their own words.

For example, it is well known that the atomic mechanisms of work-hardening and annealing of metals are poorly understood by students [11]. Many students believe that the change in yield strength is due to changes in the strength of atomic bonds or the density of the lattice. But the particular nature of students' misconceptions depend on their familiarity with field-specific vocabulary and exposure to other topics. Even robust distractors developed for an introductory course may be inappropriate for a junior- or senior-level course covering the same concept. If the distractor options targeting each misconception are not carefully worded, students may reject them out of suspicion.

The problem shown in Box 4 was posed to interview subjects without answer options with the goal of generating tempting distractor options based on subject responses. Three out of four subjects failed to mention the role of dislocations or other defects. These three subjects reasoned that the wire would undergo deformation and that the change in strength is related to the deformation, and two students specifically mentioned compaction.

When a metal wire is pulled through a hole smaller than its initial diameter, its strength increases. This is primarily because:

- (a) the material has fewer dislocations.
- (b) the material has more dislocations.
- (c) the compaction of the atoms increases the strength of the interatomic bonds.
- (d) the compaction of the crystals increases the strength of the grain boundaries.
- (e) the wire has been heated by friction through the die.

Box 4: Final version of a Mechanics of Materials concept question, posed as an open-ended question to interview participants. The options shown were generated based on subjects' responses.

Student 1: So as you pull it through hole you're going to have to compact the molecules a little more, even though it's going to be probably pretty small because metal is incompressible, in order for it to be pulled through there has to be some sort of compression a little bit. And I think that as the molecules become a little closer to each other, I think that increases the strength of the material. I would say it has to do with a decrease in molecular distance.

Student 2: This is primarily because of the compaction of the crystals within the metal wire causing it to... it's not more dense, but it's... the crystals of the metal are more compacted, causing it to be stronger. [...] So yeah originally this would be the crystals... [student draws equiaxed grains schematically inside a wire] but then you decrease the diameter and all of them become thinner, and by becoming thinner the bonds may become stronger between them.

The specific word "compaction" was used in the two distractor answers (relating to interatomic bonds and intercrystalline bonds, specifically). Compaction provides the important (though incorrect) link between the process described (wire drawing) and the outcome students erroneously expect (stronger bonds), which makes both options sound more realistic. Based on comments made by students, it seems likely that if the distractors used the word "density" instead, students would reject the answer based on their intuition that metals are (relatively) incompressible.



## Insights from large student sample

Three of the learning assessments were administered to students at a private R1 university during the Fall 2019 semester. The assessments were given as online quizzes. (In Fluid Mechanics, students first took the assessment in class using digital classroom response devices. However, the data from the in-class quiz was lost and the assessment was administered again as an online quiz at the beginning of the following semester.) Students received participation credit only and were instructed to work alone. Details are summarized in Table 1.

Table 1: Summary of classroom-scale pilot tests.

Course	Test format	Timing	Number of students	Participation rate
Fluid Mechanics	Online	Post-test <sup>a</sup>	101	86%
Mechanics of Materials	Online	Post-test <sup>b</sup>	93 <sup>c</sup>	80%
Mechatronics	Online	Pre-test	129	93%
Mechatronics	Online	Post-test	105	76%

<sup>a</sup> Due to loss of data, the MCLA was administered during the first week of the following semester.

<sup>b</sup> The MCLA was divided into three parts and administered before each major exam.

<sup>c</sup> Number of students who completed all three sections.

These assessments will be used to evaluate student learning in courses undergoing instructional change. While grading policies and exams will change year to year, the assessments will remain the same. It is important to establish that the assessments measure student proficiency. The relationship between assessment scores and course outcomes (final exam score and final grade) was estimated from the correlation coefficient (Pearson's  $r$ ), tabulated in Table 2. In all cases, course outcomes are moderately ( $0.30 < r < 0.62$ ) and statistically significantly correlated with MCLA post-test scores.

For all three courses, the MCLA post-test score is a significant predictor of final exam score, even when controlling for grade in a prerequisite course (Statics for Mechanics Materials, Introductory Thermodynamics for Fluid Mechanics, and Physics: E&M for Mechatronics). We predict the final exam score with a linear model of the form

$$\text{Final Exam} \sim \beta_P(\text{Prereq grade}) + \beta_{MCLA}(\text{MCLA score}) \quad (1)$$

The standardized coefficients are given in Table 2. (Model variables are scaled to have zero mean and unit standard deviation.)

### *Test reliability*

Test-retest reliability refers to the consistency between repeat test results from the same student after some time period with no relevant instruction in the subject matter. Assessing the test-retest reliability of an instrument is time-consuming and requires access to a sample of students who can be retested without exposure to relevant material during the gap. This is difficult for intermediate and advanced courses because of the possible overlap of material with other courses. We did not measure test-retest reliability of our assessments.

Table 2: Correlation (Pearson's  $r$ ) between the MCLA score and course outcomes, and standardized coefficients for the linear model described by Eqn. (1).

Course	$r(\text{MCLA, Final Exam})$	$r(\text{MCLA, Final Grade})$	$\beta_P$	$\beta_{MCLA}$
Fluid Mechanics	0.30**	0.33***	0.77***	0.48***
Mechanics of Materials	0.44***	0.40***	0.84***	0.55***
Mechatronics <sup>a</sup> (pre-test)	0.33*	0.43**		
Mechatronics <sup>a</sup> (post-test)	0.62***	0.49***	0.40***	0.57***

\*  $p < 0.05$     \*\*  $p < 0.01$     \*\*\*  $p < 0.001$

<sup>a</sup> The final exam was optional and was only taken by 36% of students.

Internal reliability refers to the consistency of items within the assessment. Cronbach's  $\alpha$  is a commonly-reported statistic measuring internal reliability of an instrument. It is defined as

$$\alpha = \frac{k}{k-1} \left( 1 - \frac{\sum_i \sigma_i^2}{\sigma_T^2} \right) \quad (2)$$

where  $k$  is the number of items,  $\sigma_i^2$  is the variance of item  $i$  (where correct is coded as 1 and incorrect is coded as 0), and  $\sigma_T^2$  is the variance of total test scores. Cronbach's  $\alpha$  for each assessment is reported in Table 3. A value of  $\alpha \geq 0.7$  is commonly accepted as an indication of adequate reliability [12]. However, a high  $\alpha$  only indicates that items in the test generally correlate with one another, which may not be the case for MCLAs which cover disparate learning objectives [13].

### Item analysis

We evaluated the psychometric properties of the learning assessments using several metrics from Classical Test Theory: the difficulty index, the discrimination index, and the point-biserial correlation [4, 13]. The results are summarized in Table 3, along with the number of test items falling into a certain desirable range of each metric [4, 12].

The difficulty index,  $P$ , of a test item is the proportion of students who answered the item correctly.

$$P_i = \frac{N_{\text{correct},i}}{N_{\text{total}}} \quad (3)$$

A higher  $P$ -value indicates an easy question, while a lower  $P$ -value indicates a difficult question. The range of desirable  $P$ -values depends on the intended use of the test. In a test designed to efficiently discriminate student ability (a norm-referenced test), A  $P$ -value close to 0.5 is desirable in order to give opportunities to observe both high- and low-performing students. In a test designed to measure mastery of content, a wider range of  $P$ -values is acceptable. Very easy or very difficult items may not give useful information about a given student, but they can give valuable information about group mastery of a given concept.

Table 3: Statistical properties of learning assessments.

Course	Items	$\alpha$	Score		$D$		$r_{pbs}$	
			Avg	SD	Avg	Num $\geq 0.3$	Avg	Num $\geq 0.2$
Introductory Fluid Mechanics	20	0.80	61%	20%	0.44	14	0.46	18
Mechanical Behavior of Materials	36	0.63	51%	12%	0.29	17	0.28	25
Mechatronics (pre-test)	32	0.83	43%	16%	0.42	21	0.39	30
Mechatronics (post-test)	32	0.89	68%	21%	0.58	31	0.49	32

The item discrimination index,  $D$ , is the difference in the item difficulty index between high- and low-performing students, as determined by their performance on the entire test.

$$D_i = P_{high,i} - P_{low,i} \quad (4)$$

The precise definition of high- and low-performing students is arbitrary, but  $D_i$  is commonly calculated by comparing the top and bottom 25% of the sample. A low  $D$ -value is an indication that the question does not clearly discriminate between high and low performers. A value of  $D$  above 0.3 is generally considered acceptable, while some authors recommend 0.2 or 0.4 [12].

The point-biserial correlation,  $r_{pbs}$ , measures the degree to which performance on a single item (a binary variable) correlates with performance on the entire test (a pseudo-continuous variable). Its interpretation is similar to that of the discrimination index.

$$r_{pbs,i} = \frac{\bar{T}_{c,i} - \bar{T}}{\sigma_T} \sqrt{\frac{P_i}{1 - P_i}} \quad (5)$$

where  $\bar{T}_{c,i}$  is the average test score for students answering item  $i$  correctly,  $\bar{T}$  is the average test score for all students,  $\sigma_T$  is the standard deviation of the test scores, and  $P_i$  is the difficulty index for item  $i$ .

Both the discrimination index and point-biserial correlation are sensitive to the difficulty of a question. For example, even a question which perfectly discriminates between, say, the top 10% and bottom 90% of students would only have a discrimination index of  $D = 0.4$  (using the top and bottom 25% calculation method).

For very easy or very difficult questions, the discrimination index has an upper bound which depends on the difficulty index and the fraction of students  $f$  chosen to define the high- and low-performing groups. For an easy question ( $P < f$ ), the maximum  $D$  is obtained when the few students answering incorrectly are all in the low-performing group. For a difficult question ( $P > f$ ), the maximum  $D$  is obtained when the few students answering correctly are all in the high-performing group. The upper bound on  $D$  is therefore

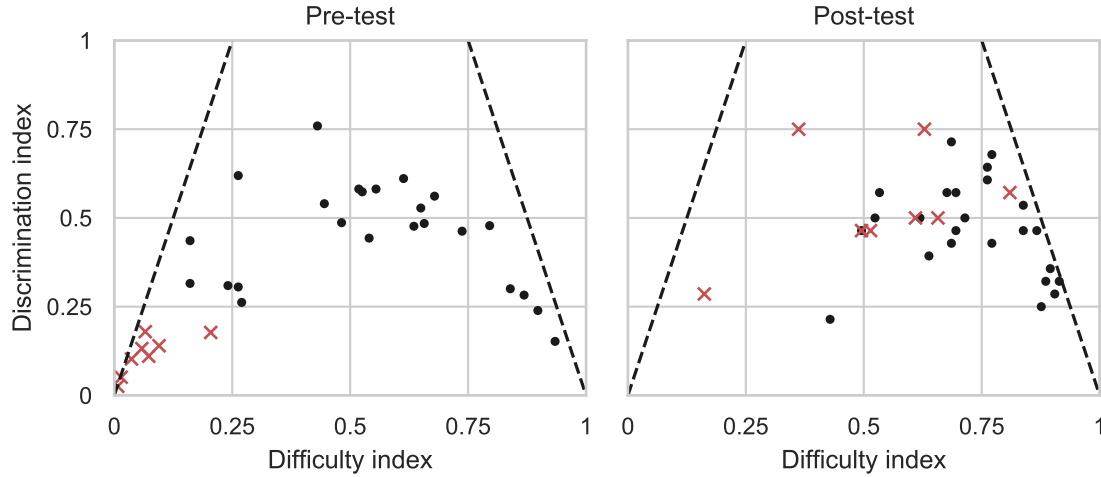


Figure 1: Relationship between difficulty index and discrimination index for Mechatronics concept questions. Problems which had a low discrimination index on the pre-test are indicated in both figures with red 'x's. The dotted line indicates the maximum discrimination index for difficult questions.

$$\max [D_i] = \begin{cases} P_i/f & P_i < f \\ 1 & f \leq P_i < 1 - f \\ (1 - P_i)/f & P_i \geq 1 - f \end{cases} \quad (6)$$

These metrics should not be interpreted as intrinsic properties of the assessment. They depend on the circumstances of administration and the population tested. Pre-test questions which rely on material for which students have no prior intuition are likely to have an extremely low difficulty index and discrimination index and therefore give almost no useful information. One such question on the Mechatronics assessment relies on knowledge of MOSFETs, a subject taught in Mechatronics which students are unlikely to have been exposed to previously. Only one student in 137 got the correct answer, resulting in both a low difficulty index ( $<0.01$ ) and a low discrimination index (0.03). In an assessment designed to rank or sort students (such as a placement exam), this question would likely be removed. In an assessment designed to measure progress towards learning objectives, this question gives valuable post-instruction information (50% of students answered the question correctly on the post-test, with a discrimination index of 0.76).

Figure 1 shows the discrimination index vs. difficulty index of items on the Mechatronics assessment before and after instruction. Questions which had extremely low difficulty and discrimination indices prior to instruction (red 'x's) showed very satisfactory properties when given after instruction.

Only 17 of 36 questions on the Mechanics of Materials assessment have  $D \geq 0.3$ . While low  $D$  can indicate problems with a question, there may be other reasons for poor correlation with the rest of the assessment. For example, only one question on the assessment requires numerical

evaluation (finding the maximum principal stress at a point). This question has an adequate difficulty (55% answered correctly), but a low discrimination index ( $D = 0.15$ ), possibly because the skill required is sufficiently different from the conceptual questions comprising the bulk of the assessment. Additional considerations must be weighed before removing an item due to poor psychometric properties.

## Conclusion

We developed multiple-choice learning assessments for four junior-level mechanical engineering courses. The insights we drew from observing students working on the assessments while vocalizing their thoughts were critical to the development process. The interviews helped us (1) identify problematic test items, (2) identify student misconceptions and use them to create distractors couched in language that sounds natural to students, and (3) build evidence for the validity of the assessment.

Due to our small sample sizes (four or five interviews and one round of class-scale testing per assessment), as well as the specificity of the subject matter tested, it is not possible to make general validity claims about our assessments. However we hope that other researchers and practitioners can learn from the specific examples of the types of insights which may be drawn from think-aloud interviews and how they supplement statistical measures.

## References

- [1] D. Sands, M. Parker, H. Hedgeland, S. Jordan, and R. Galloway, "Using concept inventories to measure understanding," *Higher Education Pedagogies*, vol. 3, no. 1, pp. 173–182, 2018.
- [2] D. Hestenes, M. Wells, and G. Swackhamer, "Force concept inventory," *The Physics Teacher*, vol. 30, no. 3, pp. 141–158, 1992.
- [3] P. S. Steif and J. A. Dantzler, "A Statics Concept Inventory: Development and Psychometric Analysis," *Journal of Engineering Education*, vol. 94, no. 4, pp. 363–371, 2005.
- [4] L. Ding, R. Chabay, B. Sherwood, and R. Beichner, "Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment," *Physical Review Special Topics - Physics Education Research*, vol. 2, no. 1, pp. 1–7, 2006.
- [5] S. A. Ambrose, M. W. Bridges, M. DiPietro, M. C. Lovett, and M. K. Norman, *How Learning Works*. Jossey-Bass, 2010.
- [6] J. Martin, J. Mitchell, and T. Newell, "Development of a concept inventory for fluid mechanics," *Proceedings - Frontiers in Education Conference, FIE*, vol. 1, pp. T3D23–T3D28, 2003.
- [7] M. Van Dyke, *An Album of Fluid Motion*. Stanford, CA: Parabolic Press, Inc., 1982.
- [8] M. A. Nelson, M. R. Geist, R. L. Miller, R. A. Streveler, and B. M. Olds, "How to Create a Concept Inventory: The Thermal and Transport Concept Inventory," in *Annual Conference of the American Education Research Association*, 2007.
- [9] P. V. Engelhardt and R. J. Beichner, "Students' understanding of direct current resistive electrical circuits," *American Journal of Physics*, vol. 72, no. 1, pp. 98–115, 2004.

- [10] H. Peşman and A. Eryilmaz, "Development of a three-tier test to assess misconceptions about simple electric circuits," *Journal of Educational Research*, vol. 103, no. 3, pp. 208–222, 2010.
- [11] S. Krause, A. Tasooji, and R. Griffin, "Origins of Misconceptions in a Materials Concept Inventory From Student Focus Groups," in *2004 American Society for Engineering Education Annual Conference & Exposition*, 2004.
- [12] R. L. Doran, "Basic Measurement and Evaluation of Science Instruction," tech. rep., National Science Teachers Association, Washington, D.C., 1980.
- [13] W. K. Adams and C. E. Wieman, "Development and validation of instruments to measure learning of expert-like thinking," *International Journal of Science Education*, vol. 33, no. 9, pp. 1289–1312, 2011.