International Journal of Engineering Education Vol. 27, No. 5, pp. 968–984, 2011
Printed in Great Britain

0949-149X/91 $3.00+0.00
© 2011 TEMPUS Publications.

Rigorous Methodology for Concept Inventory Development: Using the 'Assessment Triangle' to Develop and Test the Thermal and Transport Science Concept Inventory (TTCI)*

RUTH A. STREVELER
School of Engineering Education, Purdue University, West Lafayette, Indiana 47907, USA

RONALD L. MILLER
Center for Engineering Education, Colorado School of Mines, Golden, Colorado 80401, USA. E-mail: rlmiller@mines.edu

AIDSA I. SANTIAGO-ROMÁN
Department of General Engineering, University of Puerto Rico, Mayagüez, Mayagüez, Puerto Rico 00681, USA

MARY A. NELSON
Department of Applied Mathematics, University of Colorado, Boulder, Colorado 80309, USA

MONICA R. GEIST
Department of Mathematics, Front Range Community College, Westminster, Colorado 80031, USA

BARBARA M. OLDS
United States National Science Foundation, Arlington, Virginia 22230, USA

This paper describes a methodology for creating concept inventories that can be used to validly and reliably measure student misconceptions in engineering and science domains. Following the successful impact of the Force Concept Inventory on undergraduate physics education, creating concept inventories in engineering subjects provides engineering faculty and researchers with tools for measuring the effect of new curricular and pedagogical tools that are designed to repair misconceptions. The methodology involved aligning the three corners of the assessment triangle: cognition, observation, and interpretation. In the cognition corner, engineering students' important misconceptions in thermal science were identified using a Delphi study and validated with current learning theory. In the observation corner, items for the TTCI were created and piloted. In the interpretation corner, classical test theory and item response theory were used to evaluate the performance of TTCI items and establish instrument reliability. Versions of the TTCI have been developed for heat transfer, thermodynamics, and fluid mechanics and piloted to over 1000 undergraduate engineering students. The heat transfer portion of the instrument consists of 12 items with an overall KR-20 reliability of 0.77. Item difficulty indices range from 0.25 to 0.75 and item discrimination index exceeds 0.20 for each item. These values are sufficient for using the TTCI as a tool to identify students' misconceptions in thermal and transport science in two ways: (1) as an informal classroom assessment or (2) to establish pre-test/post-test learning gains during a course of study.

Keywords: concept inventory; assessment triangle; misconceptions

1. Introduction

Most methods for assessing engineering student learning focus on either procedural knowledge (e.g. solving specified classes of problems, designing a process or artifact, using appropriate engineering tools, oral and written communication) or development of affective and behavioral characteristics (e.g. teamwork, life-long learning, professional and ethical responsibility). Beginning in the 1970s, education researchers and educators began to identify conceptual shortcomings in students and the propensity for students to carry with them strongly-held misconceptions as to how the world around them worked [1].

One of the first systematic methods for assessing students' conceptual understanding was reported for undergraduate physics education by David Hestenes and his colleagues [2]. The instrument they developed, known as the Force Concept Inventory (FCI), consists of` 29 multiple-choice items, each designed to probe the students' understanding of Newtonian force concepts [3]. Halloun and Hestenes wrote each question using language and objects familiar to students. Each FCI item consists of a question, often accompanied by a picture, a correct answer, and four carefully developed distractors based on commonly held beliefs or misconceptions [4].

The visibility and impact of the FCI were increased in the 1990s by physics educators Hake and Mazur. Hake published FCI results for approximately 6000 undergraduate students that clearly showed the positive effect of active-learning

* Accepted 29 April 2011.

and inquiry-based pedagogical techniques on students' understanding of the force concept as measured by FCI scores [5]. Mazur at Harvard used the FCI with his students and found that, much to his surprise, student gains were no better than results reported in Hake's study [6]. Along with other innovators, Mazur began the revolution in physics education in which a renewed focus on conceptual understanding replaced some of the emphasis on routine problem-solving.

As the positive effect of the FCI on physics education has become more widely known, concept inventories (CIs) have been developed for many science and engineering fields. In addition to the thermal and transport science concept inventory, or TTCI, which will be discussed extensively in this paper, CIs are now available or under development in electric circuits [7], electromagnetic waves [8], fluid mechanics [9], heat transfer [10], materials engineering [11], signals and systems [12], statics [13], statistics [14], strength of materials [15], and thermodynamics [16], among other fields. These CIs have been created using a variety of methodologies and have been subjected to varying degrees of validity, reliability, and bias testing [17]. Others are surely being developed as well.

With CIs becoming ever more important in engineering education, it is useful to propose a framework for developing reliable, valid instruments to measure students' conceptual understanding in engineering [18]. This will help assure that CIs can be used to provide formative or even summative feedback to students and programs. Following the lead of the Physics education community's use of the FCI as a catalyst for reform, strong engineering CIs can be used to assess the effectiveness of engineering pedagogies that strengthen conceptual understanding and repair misconceptions.

In our search for a suitable framework for concept inventory development, we turned to the National Research Council publication, *Knowing What Students Know: The Science and Design of Educational Assessment* [19]. This book, created by a panel of eminent assessment experts, was commissioned by the National Academies to describe state-of-the-art assessment practices to a general audience. This work recommends that assessment instruments be designed in accordance with a framework they call the 'assessment triangle', depicted in Fig. 1, and which is composed of three interrelated elements: cognition, observation, and interpretation, which are defined below:

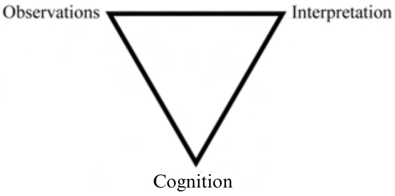


Fig. 1. Assessment triangle (adapted from [19, p. 44]).

- The cognition corner of the triangle refers to 'a theory or set of beliefs about how students represent knowledge and develop competence in a subject domain' [19, p. 44]. The domain of interest is referred to as the target domain. In other words, the cognition corner takes into account how the students learn about the target domain. When addressing the cognition corner one could consider the misconceptions that students might have about the target domain, developmental trajectories as students gain expertise, common errors that are made, etc.
- The observation corner 'represents a description or set of specifications for assessments tasks that will elicit illuminating responses from students' about the target domain to be measured' [19, p. 48]. Simply said, the observation corner represents the kinds of tasks that will make up the assessment itself. The assessment tasks that are chosen should make sense with respect to the cognition corner. For example, if one is interested in measuring higher-level thinking then the tasks on the assessment should require that higher-level thinking be exhibited. This sounds obvious. But there are numerous examples where the mismatch is dramatic.
- The interpretation corner 'encompasses all the methods and tools used to reason from fallible observations' that have been made in response to the tasks defined by the observation corner of the triangle [19, p. 48]. We might also say that the interpretation corner focuses on what we make of, or how we interpret, the results of the assessment tasks. Thus the interpretation corner guides us in choosing analysis methods appropriate for the tasks that have been created in the observation corner.
- A crucial element of the assessment triangle framework is the alignment of the three corners. Thus one's beliefs about how students learn the target domain must be consistent with the kinds of assessment tasks one creates, and with the methods one uses to analyze the results of the assessment. The interpretation of assessment results can then inform our knowledge about how students learn the target domain, and the cycle begins again.

The purpose of this paper is to describe a rigorous methodology for developing concept inventories informed by current assessment theory and methods. Specifically, we will answer the overarching research question 'How can the assessment triangle framework be used to guide development of a valid and reliable concept inventory?' To answer this overarching question, each corner of the assessment triangle (cognition, observation, and interpretation) will be addressed using the TTCI as an exemplar. Although our instrument focuses on three related domains in thermal and transport science (i.e. fluid mechanics, heat transfer, and thermodynamics), we will illustrate the use of the assessment triangle to guide concept inventory developing using only the heat transfer portion of the TTCI to simplify the discussion.

2. The cognition corner of the assessment triangle

Recall that the cognition corner of the assessment triangle asks those creating assessment instruments to consider the underlying theory or beliefs about how students develop knowledge in the target domain. In the realm of CIs, developers need to ask themselves what misconceptions or alternate conceptions students possess in the target domain and why those misconceptions might exist and persist.

To address this corner of the triangle, we ask the following research questions:

- What misconceptions do engineering students hold about heat transfer?
- Why do misconceptions about heat transfer exist and persist?

These two research questions will be addressed in Sections 2.1 and 2.2 below.

2.1 What misconceptions do engineering students hold about heat transfer?

The TTCI development team conducted a Delphi survey to elicit faculty opinion about important concepts that they felt their students did not understand and also interviewed students by posing open-ended conceptual questions. The results of the Delphi survey and student interviews were triangulated with literature on misconceptions about heat and heat transfer. The results of Delphi survey will be discussed below, followed by a section addressing the second research question. Results of student interviews will be addressed in the section on the observation corner of the assessment triangle.

Several techniques have been used to identify difficult engineering concepts. Many developers of engineering CIs have used expertise from textbook authors, course instructors, students' journals, and students' focus groups among others [17]. We chose to use Delphi methodology because it is a structured process for collecting and distilling knowledge from a group of experts by means of a series of questionnaires interspersed with controlled opinion feedback [20]. Selection of appropriate participants is crucial in the Delphi methodology and therefore well-respected engineering faculty experts and prominent thermal and transport science textbook authors were invited to participate. The participants were then asked to identify important concepts in thermal and transport science disciplines that are consistently difficult for students to understand and for which the students possess significant and robust misconceptions [21].

The distinguishing features of the Delphi technique are its use of experts and its methodology. Delphi proponents recognize human judgment as a legitimate and useful input in generating predictions and therefore believe that the use of experts, carefully selected, can lead to reliable and valid results. In addition, the Delphi technique attempts to overcome weaknesses implicit in other methods such as relying on a single expert, a group average, or a round table discussion. Using a single expert puts too much weight on one person's opinion; the group average method fails because, as Clayton notes, 'the individuals consulted have neither the opportunity to provide their most thoughtful input nor the benefit of hearing other responses that might encourage a refinement of the contributions' [22]; and the round-table approach is unreliable because some members of the group may unduly influence the decision. The Delphi method addresses the latter concern by soliciting input anonymously so that influences such as the professional reputation of a respondent or the forcefulness of a respondent's personality are neutralized. Thus all participants have equal stature in the process and their comments influence the other participants only through the logic of their argument, not their name recognition.

According to Linstone and Turoff [23, pp. 5–6], 'Usually Delphi [methodology] . . .undergoes four distinct phases. The first phase is characterized by exploration of the subject under discussion, wherein each individual contributes additional information he [sic] feels is pertinent to the issue. The second phase involves the process of reaching an understanding of how the group views the issue (i.e., where the members agree or disagree and what they mean by relative terms such as importance, desirability, or feasibility). If there is significant disagreement, then that disagreement is explored in the third phase to bring out the underlying reasons for the differences and possibly to evaluate them. The last phase, a final evaluation, occurs when all previously gathered information has been initially analyzed and the evaluations have been fed

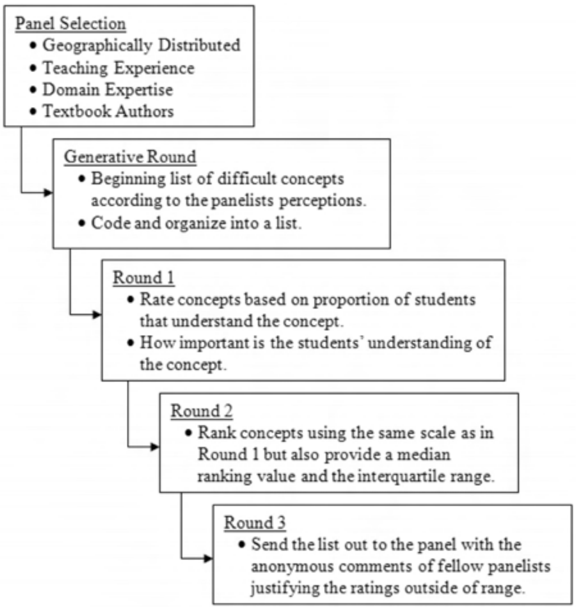

Panel Selection
• Geographically Distributed
• Teaching Experience
• Domain Expertise
• Textbook Authors
Generative Round
• Beginning list of difficult concepts according to the panelists perceptions.
• Code and organize into a list.
Round 1
• Rate concepts based on proportion of students that understand the concept.
• How important is the students' understanding of the concept.
Round 2
• Rank concepts using the same scale as in Round 1 but also provide a median ranking value and the interquartile range.
Round 3
• Send the list out to the panel with the anonymous comments of fellow panelists justifying the ratings outside of range.

Fig. 2. Graphical representation of the Delphi study applied to the TTCI.

back for consideration'. A graphical representation of this process is presented in Fig. 2.

2.1.1 Delphi study

Since the Delphi method relies on expert opinion, it was important to select the right experts. In some cases, Delphi participants are selected through a 'nomination' process in which recognized experts are solicited but they are also asked to provide the names of other experts [24]. Furthermore, selection criteria should be clearly articulated, e.g. number of years of experience, number of publications or other expert qualifications. For our panel, we started with a geographically distributed list of people with extensive expertise in the thermal and transport sciences and considerable undergraduate teaching experience. As anyone who has taught engineering students will attest, classroom teaching experience and student interactions provide a rich source of anecdotal data about lack of student understanding and the presence of misconceptions. We asked these experts to help us identify others including textbook authors in the relevant fields. Once we had identified approximately 35 experts, we sent each an email explaining the Delphi process and our project along with an invitation to join the group. Thirty-one experts agreed to participate, a number corresponding well with Clayton's rule-of-thumb that 15–30 people are an adequate panel size [22]. The group included tenured and tenure-track engineering professors from research universities and undergraduate institutions. Five of the participants had authored well-known texts in thermodynamics, fluid mechanics, heat transfer, or thermal science. Although we purposely included new faculty members in the study, the average number of years taught was 23 [25]. We guaranteed confidentiality for all participants during the process, another important element of the Delphi procedure.

Generative round. We wanted our expert panel to generate a beginning list of difficult concepts in thermal and transport science rather than generate our own. Therefore, we included a pre-Delphi generative round in which we asked panelists to describe concepts that their students found difficult. Table 1 summarizes example comments received for the heat transfer domain [25]. Of the 31 experts who agreed to participate in the study, 23 provided approximately 60 ideas, which were coded and organized into a list of 28 concepts; this list included all the concepts that had been submitted by at least two experts. These 28 concepts spanned the heat transfer, thermodynamics, and fluid mechanics domains and formed the basis for our subsequent rounds.

Round 1. We asked the experts to rate each concept based on two factors: (1) the proportion of his/her students that understand the concept and (2) how important it is for students to understand the

Table 1. Examples of generative round comments by Delphi participants related to heat transfer misconceptions

Comments of Delphi participants
• Confusion between temperature and heat transfer. Many students believe that if the path to a state involves a heat transfer input the temperature of the system will increase—even if the heat transfer is coupled with work leaving the system.
• Students do not appear to have precise understanding of heat in the sense that it is used in thermodynamics.
• Students have difficulty identifying heat and work interactions between the system and the surroundings.
• [Students] have trouble with work and energy relationships.
• There is always student confusion about heat vs. internal energy.
• There is always student confusion about internal energy and enthalpy.
• Misconception: Temperature is a measure of energy. Example: Students often believe that if you add energy, heat for example, to any system, the temperature must go up. A corollary to this is that students often believe that if the temperature goes up the energy (either internal energy or enthalpy) must have increased. A good example of a system that is very confusing is an evaporative cooling process in psychrometrics where the enthalpy of the moist air stays constant but the temperature decreases.
• Heat, like energy, is a familiar term but its common use differs from thermodynamic definition.
• Heat as transferred energy. No matter how often you make the point, some [students] insist on talking about the heat content of a system.
• Confusion about the difference between heat and temperature. How can a process occur where heat is added but the temperature drops?

concept. We used a scale of 0 to 10 for each question (0 = no one understands this concept to 10 = everyone understands this concept, and 0 = it is not at all important to understand this concept to 10 = it is extremely important to understand this concept). Thirty members of our expert panel ranked these 28 concepts. In all rounds, participants were told that 'you will not have to rate any concept for which you don't feel you have sufficient expertise or classroom experience.'

Round 2. Delphi methodology prescribes three rounds of the ranking exercise. For the second round of our study, we presented the panel of experts with the same 28 concepts and asked them to rank the concepts using the same scales as in round 1. However, in round 2 we also provided them with the median ranking value and the interquartile range (containing the middle 50% of rankings) for each concept. In this round, if participants rated a concept outside of the interquartile range established in round 1, they were asked to provide a justification for their rating. In this way, the median ranking of each concept approached a stable value and the interquartile range decreased in size, representing the consensus opinion of the participant group. Twenty-eight members of our expert panel ranked the 28 concepts in round 2.

Round 3. In the third round, we again asked the experts to rank all 28 concepts. We provided the median rating and interquartile range from round 2 and the anonymous comments that fellow panelists submitted justifying ratings outside of the specified range. Twenty-six panelists participated in round 3. Based on this final iteration, we identified 12 of the least understood but most important concepts in thermal and transport science; these formed the content domain for developing TTCI items.

Delphi results. Results from each Delphi round are summarized in Table 2. The non-parametric median and interquartile range are reported (rather than mean and standard deviation) because concepts were rated on an ordinal scale. Two important results are summarized in this table. First, the median and interquartile ranges for most concepts stabilized by round 2 (the median for 19 of the 28 concepts changed by a value of 0.5 or less and interquartile ranges became narrower); similar results have been reported by other Delphi studies [23].

Second, concepts can be identified that are of high importance (those that were given a high ranking in the 'importance' scale) and also conceptually difficult (those that were given a low ranking on the 'conceptual understanding' scale). As shown in Table 2, a total of 12 concepts were identified as meeting the criteria of high importance but low conceptual understanding (shown in italics). These items included key topics in thermal and transport science domains such as the 2nd law of thermodynamics including reversible vs. irreversible processes, conservation of fluid momentum, viscous momentum transfer, several energy-related topics (heat, temperature, enthalpy, internal energy), and steady-state vs. equilibrium processes. Two of the concepts (differential vs. integral analysis and system vs. control volume analysis) were deemed mathematical rather than physical concepts and were not included in the TTCI. At the request of several Delphi participants, we also included the ideal gas law and conservation of mass concepts in the TTCI, since both are fundamental concepts in fluid mechanics and thermodynamics. Finally, we temporarily set aside the thermal radiation concept, which will eventually be included in the instrument. This decision was made so that we could focus the development of early versions of the instrument on what were deemed more fundamental heat transfer topics. Based on these adjustments, the heat transfer portion of the TTCI focuses on three key concepts as indicated in Table 2: heat vs. energy, heat vs. temperature, and steady-state vs. equilibrium processes [26].

2.1.2 Literature review to validate Delphi results

To validate the results of the Delphi process independently, we consulted the available misconception literature, especially a comprehensive bibliography of approximately 8400 studies reported by Duit [1]. The bibliography contains over 500 references to work on heat transfer misconceptions and slightly fewer citations for thermodynamics and fluid mechanics.

Confusion about thermal processes and heat transfer has been identified in students of all ages and focuses on the following five conceptual themes [27–29]:

- heat and temperature are equivalent (related to concept 13 from Table 2);
- temperature determines how 'cool' or 'warm' a body feels (related to concept 12 from Table 2);
- heat is a substance transferred between bodies (related to concept 12 from Table 2);
- addition of energy as heat always increases the temperature in a body (related to concept 15 from Table 2);
- temperature should change in a phase transition (e.g. boiling) since energy is being added or removed (related to concept 15 from Table 2).

Thus, we found that 'important but poorly understood' heat transfer misconceptions identified in the

Table 2. Results of thermal and transport concepts Delphi study (*Italicized* concepts are those that the Delphi study identified as poorly understood but highly important (i.e. low scores on the 'understanding' scale but high scores on the 'importance' scale)

	'Understanding' data median (interquartile range)			**'Importance' data median (interquartile range)**		
Concept	**Round 1**	**Round 2**	**Round 3**	**Round 1**	**Round 2**	**Round 3**
Adiabatic vs. Isothermal processes	7.5 (6–8)	8 (6–8)	8 (6.75–8.25)	9 (8–10)	9 (9–10)	9 (9–10)
Bernoulli equation	7 (4–8)	6 (5–7)	6 (5–7)	9 (7–10)	9 (8–9)	9 (8–9)
Compressible vs. Incompressible flow	5 (3–7)	6 (4–6.5)	6 (5–7)	7.5 (6–8)	7 (7–8)	7.5 (7–8)
Conservation of linear momentum	5 (3–6)	5 (4–6)	5.5 (5–6)	9 (8–10)	9 (8–10)	9 (8–9.25)
Differential vs. Integral analysis	4.5 (3–6)	4 (3–5.25)	4 (4–5)	7 (6–9)	8 (6–8)	8 (7–9)
Dimensional analysis	6 (4–7)	5.5 (4.25–7)	6 (5–6.25)	7 (5–7)	6 (5–8)	7 (5–8)
Entropy & 2nd law of thermodynamics	4 (2–6)	4 (3–5)	5 (3–5.25)	8 (7–9)	9 (8–9)	9 (8–10)
Extensive and intensive properties	8 (6–9)	8 (7–8)	8 (7–9)	7 (6–9)	8 (7–9)	8 (7–9)
First law of thermodynamics	8 (7–9)	8 (7–9)	8 (8–9)	10 (10–10)	10 (10–10)	10 (10–10)
Fluid vs. Flow properties	7 (5–8)	6 (5–7)	6 (5–6)	7 (5–9)	7 (5–8)	7 (5–8)
Heat transfer modes	8 (6–9)	8 (6.25–8)	8 (7–9)	9 (8–10)	9 (9–10)	9 (9–10)
Heat vs. Energy	6 (5–8)	6 (5–7)	6.5 (5–7)	9 (8–10)	9 (8–10)	9 (8–10)
Heat vs. Temperature	6 (4–8)	6.5 (5–8)	7 (6–8)	9 (8–10)	10 (9–10)	10 (9–10)
Ideal gas law	8 (7–9)	8 (8–9)	8 (8–9)	9 (8–10)	9 (9–10)	9 (9–10)
Internal energy vs. Enthalpy	6 (3–7)	5 (4–6)	6 (5–6.25)	8 (7–9)	9 (8–9)	9 (8–9)
No-slip boundary conditions	8 (6–9)	8 (7–9)	8 (8–9)	8 (7–9)	9 (8–9)	9 (8–9)
Nozzles and diffusers	6 (5–8)	6 (6–7.5)	7 (6–7)	7 (5–9)	7 (6–8)	7 (6–8)
Pressure	8 (6–9)	8 (7–8)	8 (7.75–9)	9 (8–10)	10 (9–10)	10 (9.75–10)
Reversible vs. Irreversible processes	5 (4–7)	5 (4–6)	5 (5–6)	8 (8–9)	9 (8–9)	9 (8–9)
Spatial gradient of a function	4 (3–7)	5 (4–6)	5 (4–5)	7 (3–9)	7 (6–8)	7 (6–8)
Specific heat capacity	7 (6–8)	7 (6–7)	7 (6–8)	8 (7–10)	9 (8–9)	9 (8–9)
Steady-state vs. Equilibrium process	5 (3–8)	5 (3–6)	5 (4–5.25)	8 (5–10)	9 (7–9)	9 (8–9)
Steady-state vs. Unsteady-state process	8 (7–8)	8 (7–8)	8 (7–8)	9 (8–10)	9.5 (9–10)	9.5 (9–10)
System vs. Control volume	7 (4–8)	6 (5–7)	6 (6–7)	8 (6–10)	9 (8–10)	9 (8.5–10)
Temperature scales	7 (5–9)	8 (8–9)	9 (8–9)	8 (6–10)	9 (8–10)	9 (9–10)
Thermal radiation	6 (4–8)	5 (5–6)	5 (5–6)	7 (5–9)	8 (6.75–8)	8 (7–8.25)
Thermodynamic cycles	7 (5–8)	7 (6–7)	7 (7–8)	8 (8–10)	9 (8–10)	9 (8–9.25)
Viscous momentum flux	5 (3–7)	4 (3.75–5)	4 (3–4)	7.5 (6–9)	8 (7–8)	7 (7–8)

'Understanding' Scale
0 = no one understands the concept
10 = everyone understands the concept

'Importance' Scale
0 = no at all important to understand the concept
10 = extremely important to understand the concept

Delphi study were prominently mentioned in the misconception literature. Each of these themes has been discussed in more detail elsewhere [18].

2.2 Why do misconceptions about heat transfer exist and persist?

With misconceptions in heat transfer identified via the Delphi survey and validated through a literature search, we now turn our attention to the second research question that the cognition corner suggests. Why do these misconceptions about heat transfer exist? And why do these misconceptions persistent, even after repeated instruction? Although other engineering CIs have used Delphi methodology to identify the key concepts to be included in the respective inventory, the TTCI is unique among engineering CIs in addressing the question of why some misconceptions persist [17].

Why do misconceptions about heat persist? The cognition corner of the assessment triangle should take into account theories or beliefs about students' knowledge in the target domain. In the case of concept inventories, this means that CI developers should identify a theory that helps them explain why misconceptions in the target domain persist. In the parlance of learning scientists, misconceptions that persist even in the face of repeated instruction are called 'robust.' Another way to phrase the question that CI developers should ask is: 'Why are some misconceptions (in the target domain of the CI) robust?'

We assumed that the concepts rated as having low understanding and high importance in the Delphi survey represented robust misconceptions. Therefore, we were looking for a theory that would help us explain our Delphi results.

Research about misconceptions began in the 1980s with the work of Posner and colleagues [30] who posited that instructors could convince students that their own ideas about certain phenomenon were incorrect by showing them the scientifically accepted explanation. If students were shown why their thinking was incorrect, their concepts about a phenomenon would change. Although the work of Posner has been widely cited, this theory cannot explain why some concepts are still misunderstood or misconceived after years of instruction in the 'correct' mode of thinking. For example, our studies have demonstrated that even advanced engineering students will incorrectly explain basic concepts in their field [31]. Therefore, the work of Posner and colleagues was not consistent with our findings and thus was not chosen as a foundation for our work.

More recent work in misconceptions has been championed by three prominent researchers: Chi, diSessa, and Vosniadou, who have each posited their respective theories about why some misconceptions are robust [32]. As previous work has chronicled [33, 34] the works of diSessa and Chi seem to speak most directly to engineering educators. diSessa's theories focus on explaining misconceptions of force and other phenomena of mechanics [35], not the target domain of the TTCI. However, Chi's work prominently discusses misconceptions about diffusion and equilibrium (fundamental concepts in the thermal and transport sciences) and thus spoke quite directly to the target domain of the TTCI [36]. Therefore, Chi's theories about misconceptions were particularly salient for the development of the TTCI and became the foundation of our 'cognition corner'. Her theory predicted that the most widely and persistently held misconceptions are those around phenomena that arise as the emergent properties of systems. This prediction guided the creation of questions for the TTCI (the observation corner) and helped us explain our results (the analysis corner). The alignment of the cognition, observation, and interpretation corners of the assessment triangle will be expanded in Section 5.

3. The observation corner of the assessment triangle

In this section, we discuss in detail the development of specific items included in the TTCI. Specifically, we addressed the following research question in this phase of the work: 'How does one appropriately measure conceptual understanding by undergraduate engineering students in the domains of fluid mechanics, heat transfer, and thermodynamics?'

3.1 TTCI item development

We chose to create an instrument patterned after the Force Concept Inventory (FCI). As the developers of the FCI [3] found, the best questions are simple, do not involve mathematics, and have distractors that indicate the presence of common misconceptions. Once we identified concepts from the Delphi study that would be included in a multiple-choice misconception instrument, we began developing candidate items for each concept [25, 37–38]. Each item was developed in alignment with a process recommended by Downing [39] in the Handbook of Test Development:

1. Drafting open-ended questions about the concept
2. Collecting student response data orally (think-aloud problem solving sessions) and in written form
3. Using the responses to convert the open-ended questions to multiple choice items with distractors describing plausible but incorrect answers
4. Beta testing the drafted items on groups of engineering students
5. Collecting expert reviews of each item (which also provides evidence of content validity
6. Revising the items based on statistical performance and expert feedback, and
7. Collecting additional beta test data.

To illustrate this process, we will describe the genesis and development of one TTCI item we will term Hotplate (see Fig. 3). This item was developed very early in TTCI project work and as described below has evolved through several rounds of editing to become one of the best-performing items in the inventory.

Step 1—From concept to open-ended item
Hotplate was designed to assess the students' conceptual understanding of the relationship among energy (specifically internal energy), temperature and heat. The genesis for the item was a similar question included in the *Chemistry Concepts Inventory* authored by Melford [40]. The original open-ended version of Hotplate is shown in Fig. 4. In this item, each fluid is heated in an identical beaker between the same starting and ending temperature using the same rate of energy addition with identical hot plates. Based on our preliminary student think-aloud data and Mulford's work, we expected that students who were confused about the relationship between energy, temperature, and heat capacity

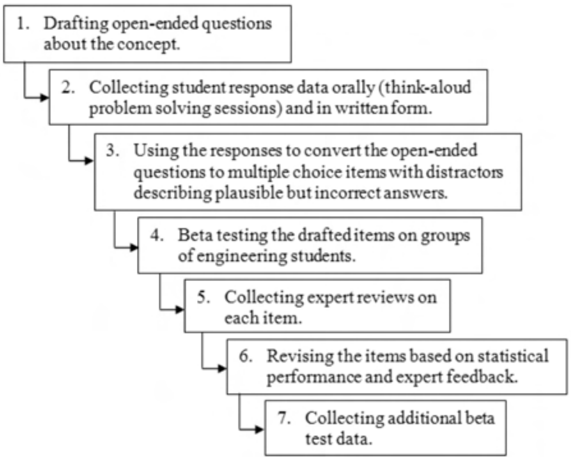
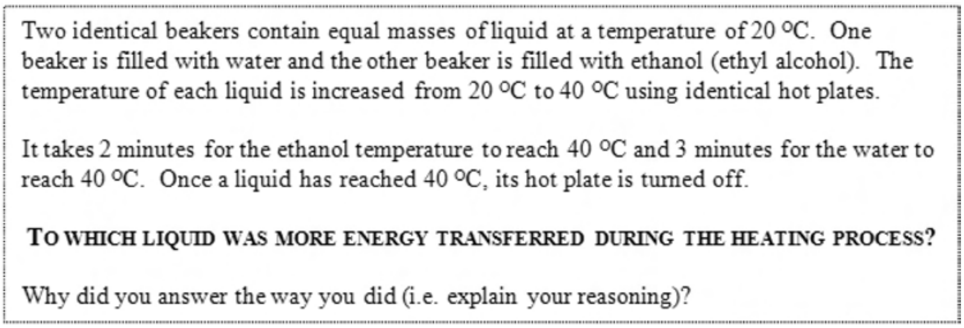

1. Drafting open-ended questions about the concept.
2. Collecting student response data orally (think-aloud problem solving sessions) and in written form.
3. Using the responses to convert the open-ended questions to multiple choice items with distractors describing plausible but incorrect answers.
4. Beta testing the drafted items on groups of engineering students.
5. Collecting expert reviews on each item.
6. Revising the items based on statistical performance and expert feedback.
7. Collecting additional beta test data.

Fig. 3. Graphical representation of the process used to develop TTCI items.



Two identical beakers contain equal masses of liquid at a temperature of 20 °C. One beaker is filled with water and the other beaker is filled with ethanol (ethyl alcohol). The temperature of each liquid is increased from 20 °C to 40 °C using identical hot plates.

It takes 2 minutes for the ethanol temperature to reach 40 °C and 3 minutes for the water to reach 40 °C. Once a liquid has reached 40 °C, its hot plate is turned off.

TO WHICH LIQUID WAS MORE ENERGY TRANSFERRED DURING THE HEATING PROCESS?

Why did you answer the way you did (i.e. explain your reasoning)?

Fig. 4. Open-ended version of hotplate item.

would not generally be able to answer this item correctly.

Step 2—Collecting open-ended student responses to the item

Six students (all juniors or seniors majoring in chemical or mechanical engineering) individually participated in think aloud sessions to discuss the hotplate item. The role of the interviewer in this session was to elicit more detailed student answers to *Hotplate* and to elicit explanations about why students answered as they did. Student responses were audiotaped, transcribed, and analyzed for evidence of conceptual understanding and prevalent misconceptions about energy and temperature. As predicted by the Delphi results, a majority of students participating in think-alouds demonstrated limited understanding of the concepts being addressed, though some provided reasonably correct answers. Examples included:

- Incorrect answers based on misconceptions:
  - 'They both received the same energy because the temperature change was the same.'
  - 'We can't tell because we don't know the [fluid] heat capacities.'
  - 'It has nothing to do with heat transfer, only temperature.'
  - 'The amount of time it takes to heat is based only on the heat transfer coefficient in the beakers.'
- Correct answers:
  - 'Just because ethanol gets hotter faster does not mean it gains more heat. Just that the ethanol has a lower heat capacity.'
  - 'Water because it was heated longer at the same rate of heating.'

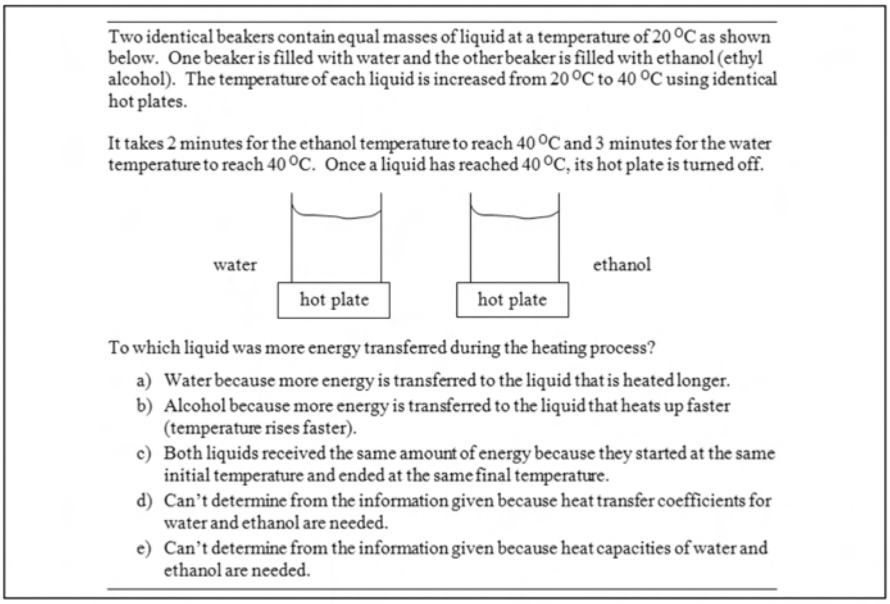
Step 3—From open-ended question to multiple choice test item

During coding of the *Hotplate* think-aloud data, we identified several misconception patterns in student responses including confusion about temperature vs. energy or heat, what heat capacity of a substance means, equating the rate of heat transfer with the amount of energy transferred, and incorrectly thinking that heat capacities and/or heat transfer coefficients were required to answer the item. These results confirmed the predictions of the Delphi experts (Table 2) and also were aligned with Chi's predictions about which misconceptions would be robust [36]. As mentioned earlier, these misconceptions have been prominently reported in the thermal science misconception literature [1] and gave us confidence that our think-aloud strategy was eliciting significant student thinking that was worthy of inclusion as plausible distractors for the multiple-choice version of *Hotplate*. Using student comments like those listed earlier combined with input from the misconception literature and Delphi participants, we drafted four Hotplate distractors (answers b–e) along with the correct answer (answer a). The original multiple choice version of Hotplate is shown in Fig. 5.

Step 4—Initial beta testing with students

The version of *Hotplate* shown in Fig. 5 was tested with 87 students at four engineering institutions. The distribution of student responses is shown in Fig. 6 and indicates that about 50% of the students answered the item correctly. Distractors 'c'(~15%), 'd' (~13%), and 'e' (~19%) were all selected by a significant number of students, suggesting that the item may have been eliciting the types of student misconceptions and incorrect thinking that we expected to see. Distractor 'b' was selected by only two students, which was a bit surprising given interview data that indicated student confusion between rate of heating and amount of energy transferred. Good test construction practice recommends removing or replacing distractor 'b' and that may happen in later versions of the instrument, but for now it remains part of the item until additional test data are collected to make a more reasoned decision.

Other statistical results for *Hotplate* were also computed including the item difficulty index (0.51) and item discrimination index (0.58). These statistical markers will be discussed in Section 4.2, but



Two identical beakers contain equal masses of liquid at a temperature of 20 °C as shown below. One beaker is filled with water and the other beaker is filled with ethanol (ethyl alcohol). The temperature of each liquid is increased from 20 °C to 40 °C using identical hot plates.

It takes 2 minutes for the ethanol temperature to reach 40 °C and 3 minutes for the water temperature to reach 40 °C. Once a liquid has reached 40 °C, its hot plate is turned off.

To which liquid was more energy transferred during the heating process?

a) Water because more energy is transferred to the liquid that is heated longer.
b) Alcohol because more energy is transferred to the liquid that heats up faster (temperature rises faster).
c) Both liquids received the same amount of energy because they started at the same initial temperature and ended at the same final temperature.
d) Can't determine from the information given because heat transfer coefficients for water and ethanol are needed.
e) Can't determine from the information given because heat capacities of water and ethanol are needed.

Fig. 5. Version 1 of *Hotplate* multiple choice test item.

both values were judged to be well within acceptable range for items in the TTCI.

Steps 5 and 6—Expert review and item revision
After we completed initial beta testing, each item in the TTCI was reviewed by two technical experts in the disciplines of fluid mechanics, heat transfer or thermodynamics. Expert feedback and comments about Hotplate focused on three issues:

- the effect of evaporation losses in the open beakers
- the effect of using hotplates where the effect of fluid properties might affect actual heat transfer rates for water and alcohol
- the wordiness in the item description and some of the distractors.

As a result of this valuable feedback, we replaced the use of hotplates and open beakers with immersion heaters in closed beakers to minimize the effects of evaporation and varying fluid properties on the heating rates. To improve clarity, we also rewrote the item and shortened the answers. Version 2 of *Hotplate* is shown in Fig. 7.

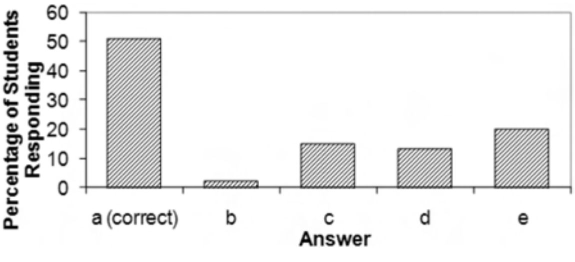
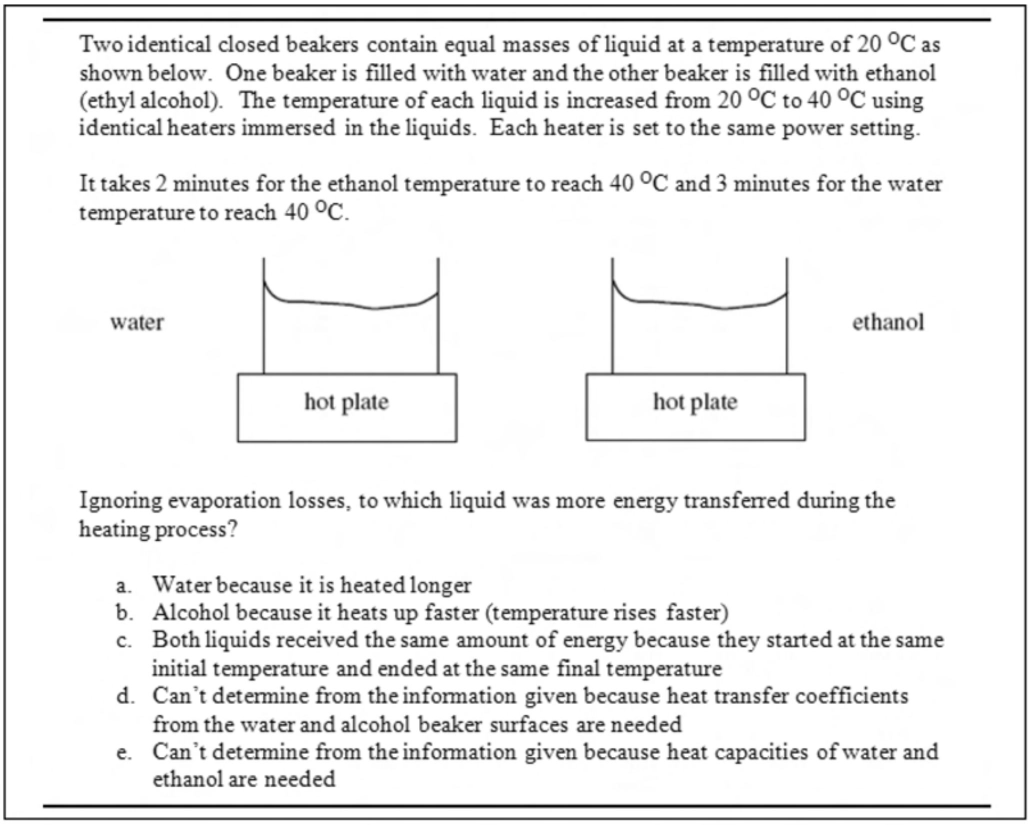


Fig. 6. Summary of student responses to *Hotplate* item.

Step 7—Additional beta testing
In addition to Hotplate, one item was developed for each concept identified in the Delphi study using the 7-step procedure described earlier. Collectively, these questions became version 1.0 of the TTCI, which was alpha-tested with ten engineering student volunteers. Initial results indicated that a range of distractors was selected for each item. We interpreted this result as an indication that significant misconceptions were represented by the distractors included in version 1.0 and that addition items should be written for each concept in the TTCI.

Version 1.0 was also used to test two types of item formats: single question items and items that consisted of an initial question ('what will happen?') with a follow-on question ('why will it happen?'). The two-question format was developed to allow for deeper probing of student reasoning for selecting the answers that they did and was designed to provide richer data sets for studying the existence and nature of robust misconceptions like those predicted by Chi *et al.* [36,41].

Based on the initial success of version 1.0, additional items were drafted for each TTCI concept using the 7-step process and it was tested on small groups of students. If student feedback suggested an item was written clearly and if response data were distributed among most or all of the distractors, the item was added to the inventory. Items that did not meet these criteria were rejected or rewritten. Once at least 2–3 acceptable items were developed for each concept, the items were collected into the next generation of the TTCI, designated as version 2.21. This version was used to collect sufficient data for psychometric analysis (see Sections 4.1 and 4.2 for a



Two identical closed beakers contain equal masses of liquid at a temperature of 20 °C as shown below. One beaker is filled with water and the other beaker is filled with ethanol (ethyl alcohol). The temperature of each liquid is increased from 20 °C to 40 °C using identical heaters immersed in the liquids. Each heater is set to the same power setting.

It takes 2 minutes for the ethanol temperature to reach 40 °C and 3 minutes for the water temperature to reach 40 °C.

Ignoring evaporation losses, to which liquid was more energy transferred during the heating process?

a. Water because it is heated longer
b. Alcohol because it heats up faster (temperature rises faster)
c. Both liquids received the same amount of energy because they started at the same initial temperature and ended at the same final temperature
d. Can't determine from the information given because heat transfer coefficients from the water and alcohol beaker surfaces are needed
e. Can't determine from the information given because heat capacities of water and ethanol are needed

Fig. 7. Version 2 of *Hotplate* multiple choice test item.

summary of these results). In some cases, items were edited to improve performance as measured by discrimination and difficulty indices (discussed in Section 4.2) while in other cases new items were developed to replace weak items and more new items were developed to increase the total number of items in the inventory. After editing weak items and drafting/piloting new items, version 3.04 of the TTCI was formed.

Table 3 shows the number of items and questions included in versions 2.21 and 3.04 for the heat transfer portion of the TTCI. Versions 2.21 and 3.04 represent major revisions in the TTCI; each version was used to collect sufficient student response data for statistical reliability measurements. Details of the psychometric performance of individual items and each TTCI version will be discussed in Section 4.2. Version 3.04 is the current version of the TTCI and is now available on-line.

4. The interpretation corner of the assessment triangle

The interpretation corner includes 'all the methods and tools used to reason from fallible observations' [19, p. 48]. It relates the observations collected from the assessment tasks with the cognitive knowledge and skills being assessed. The interpretation method is usually a statistical model that is directly related to the purpose of the instrument [19]. For example, if the purpose of an instrument is to assess short-term gains (such as those made after instruction in one course) or long-term gains (such as those made at the completion of a program) one would compute gain score (the difference between pre- and post-instruction scores). For that purpose Classical Test Theory (CTT) would be useful to determine traditional measures of reliability.

In the case of the TTCI the purpose of the

Table 3. Number of items/questions for each heat transfer concept included in the TTCI versions 2.21 and 3.04

TTCI version	Heat vs. Energy	Energy vs. Temperature	Steady-state vs. Thermal equilibrium
2.21	3/3	3/4	2/4
3.04	3/3	7/11	2/4

Version 3.04 is now available on-line at www.thermalinventory.com. A password can be obtained from Dr. Ron Miller at rlmiller@mines.edu.

instrument is to identify engineering misconceptions in the thermal and transport sciences (specifically heat transfer, thermodynamics and fluid mechanics). Therefore, it is important to ensure that TTCI items are both reliable and valid to meet the intended objective. This was achieved using various types of psychometric analysis techniques described in the following section.

Psychometrics is the field of study concerned with the theory and technique of educational and psychological measurement. It involves two major research tasks, namely: (1) the construction of instruments and procedures for measurement; and (2) the development and refinement of theoretical approaches to measurement [42]. Psychometric theory involves several distinct areas of study such as data analysis using CTT, Item Response Theory (IRT), and correlation and covariance techniques, which include factor analysis, multidimensional scaling, and data clustering.

We will discuss both CTT and IRT in the following sections and discuss how these theories were applied to TTCI evaluation.

4.1 Using classical test theory: reliability and validity

In this section we provide some background for CTT and discuss how it was applied to the TTCI. CTT is a body of related psychometric theory that predicts outcomes of psychological testing such as the difficulty of items or the ability of test-takers. In general, the aim of CTT is to understand and improve the reliability of a given test, so key traditional concepts in CTT include reliability and validity.

4.1.1 Reliability

Reliability is defined as the consistency of a set of measurements often used to describe a test. A reliable instrument is one in which measurement error is small, which can also be stated as the extent to which results using the instrument are repeatable.

One of the most common measure of reliability is internal consistency because it requires only one test administration, thereby reducing costs and eliminating the issue of students gaining knowledge between test administrations. Internal consistency is typically measured using Cronbach's alpha which is a generalized form of Kuder–Richardson formula 20 (KR-20). Typically, a test is considered reliable if alpha is above 0.80 [43]. Other sources consider a value of 0.60–0.80 to be acceptable for classroom tests [44]. Cronbach's alpha is a coefficient of consistency that measures how well a set of variables or items measures a single, unidimensional latent construct while KR-20 is a measure of internal consistency for measures with dichotomous (0 or 1) answer choices such as employed in the TTCI [45].

Table 4. KR-20 values for the heat transfer TTCI instrument, Version 3.04

TTCI version	KR-20 value	Sample size
3.04	0.77	749

KR-20 results for the TTCI heat transfer instrument are shown in Table 4 and suggest that the instrument reliability for version 3.04 is approaching the desired value of 0.8 as better items are added to the inventory. It should also be noted that by intentionally including very difficult questions (as opposed to achievement or diagnostic instruments in which a wider range of item difficulty is included), the overall reliability of the instrument can be expected to be lower since students will more likely guess on some questions or at least choose an answer in which they are not completely confident [46].

4.1.2 Validity

Validity refers to the extent that an instrument measures what it claims to measure [47]. Validation is an on-going process of accumulating evidence supporting the claim, thus the instrument must be constantly evaluated as its uses and needs evolve. Validity is considered the degree of correlation between the test and a criterion.

Although several types of validity are available to establish instrument performance, we focused on the two most relevant to a concept inventory: content and construct validity. The TTCI is most appropriately used as an assessment for identifying the existence of important misconceptions and for pre–post studies of misconception repair for a specific student population.

Content validity focuses on the question of whether items included in the instrument span the appropriate and desired technical domain. During TTCI development, we have worked in several ways to ensure content validity. The Delphi process discussed earlier represented an intensive exercise to identify important concepts that are poorly understood by students in thermal and transport sciences as defined by a panel of experts in these domains. This exercise achieved consensus about the key concepts and related misconceptions to be included in the instrument and informed our work as individual test items were developed. As discussed earlier, an extensive literature search also confirmed that many of the difficult concepts identified in the Delphi study were also mentioned by previous researchers working in the heat transfer, thermodynamics, and fluid mechanics domains.

Thus, we claim that the conceptual domain covered by the TTCI is the domain we intended to cover. It is important to note, however, that we do not claim that the TTCI covers all important concepts in thermal and transport sciences but rather focuses on concepts deemed important by experts, but often misunderstood by students, in these subject areas.

Construct validity attempts to answer the question of whether instrument items measure the concepts we think they are measuring. To address this issue, all items were drafted by domain experts and then reviewed by at least two additional experts who teach, conduct research, and in some cases write textbooks in the TTCI domains. Expert feedback was used to revise item wording and accompanying graphics as we strived for question correctness and clarity. Experts agreed that the intended concept and misconceptions were correctly targeted in each item included in the inventory.

We also used think-aloud sessions with engineering juniors and seniors to confirm that students could identify the concept associated with each draft item, that the item text and graphics were understandable, and that each item raised a conceptual difficulty with most of the students interviewed. We recorded and coded all think-alouds to ensure a complete and accurate picture of student responses. Student responses gave us important feedback that, in some cases, uncovered items that were misunderstood because of poor question construction or graphics, or because we inadvertently used unfamiliar vocabulary, symbols, or notation.

4.2 Using item response theory: item difficulty and item discrimination

In this section we will provide some background for Item Response Theory (IRT) and discuss how it was applied to the TTCI. IRT describes the application of mathematical models to data from questionnaires and tests as a basis for measuring abilities, attitudes, or other variables [48]. It is used for statistical analysis and development of assessment instruments. Furthermore, it is based on the idea that the probability of getting an item correct is a function of a latent trait or ability [43]. Specifically, a person with higher intelligence would be more likely to respond correctly to a given item on a given instrument.

IRT provides a basis for evaluating an assessment instrument or item. In psychometrics, IRT is applied to refine exams, maintain bank of items for exams, and equating difficulties of successive versions of exams [48]. Among other advantages, IRT provides a basis for obtaining an estimate of the location of a test-taker on a given latent trait as well as the standard error of measurement of that location. Scores derived by classical test theory do not have this characteristic, and assessment of actual ability (rather than ability relative to other test-takers) must be assessed by comparing scores with those of a 'norm group' randomly selected from the population [44]. Item parameters typically used by IRT specialists are item difficulty and item discrimination. In the case of the TTCI, item difficulty and item discrimination indices were used to determine which items in version 3.04 were performing satisfactorily.

4.2.1 Item difficulty

Item difficulty is the percentage of the total group that correctly answered the item [47]. Item difficulty is an important parameter because it reveals whether an item is too easy (too many people answering correctly) or too difficult (very few answering correctly). The optimal item difficulty depends on the question-type and on the number of possible distractors. Kline [46] suggested an item difficulty range of 0.25–0.75 for concept assessments. This means that between 25% and 75% of test takers answered an item correctly. Following Kline, we retained items in the TTCI that fell in the difficulty index range of approximately 0.25–0.75.

4.2.2 Item discrimination

Item discrimination refers to a test's ability to produce a wide range of scores by separating students who vary in their degree of knowledge of the material tested and their ability to use it [42]. For example, if one group of students has mastered the material and the other group had not, a larger portion of the former group should be expected to answer a test item correctly. Item discrimination is the difference between the percentages of correct answers for these two groups.

Item discrimination can be calculated by ranking the students according to total score and then selecting the top 33% and the lowest 33% in terms of total score [49]. For each item, the percentage of students in the upper and lower groups answering correctly is calculated. Therefore, the maximum item discrimination difference is 100%, which would occur if all those in the upper group answered correctly and all those in the lower group answered incorrectly. Zero discrimination occurs when equal numbers in both groups answer correctly. Negative discrimination, a highly undesirable condition, occurs when more students in the lower group than the upper group answer correctly. The levels presented on Table 5 may be used as a guideline for acceptable items. Thorndike [50] also recommends eliminating any item with a discrimination index of 0.24 or less.

Item discrimination was used as the second measure of item performance. For purposes of evaluating TTCI items, we followed the advice of

Table 5. Acceptable levels for item discrimination [50]

Item discrimination	Classification	Required action
Negative	Unacceptable	Check item for error
0–24%	Usually unacceptable	Item should be improved
25–39%	Good item	Keep item
40–100%	Excellent item	Keep item

Thorndike and only accepted questions with item discrimination indices greater than 0.25. All items and questions retained in version 3.04 met these discrimination criteria.

5. Alignment of the three corners of the assessment triangle

For an assessment to be effective, the three corners of the assessment triangle must be aligned [19]. In this section, we describe how each component of the assessment triangle work together in the TTCI. See Fig. 8 for a graphical illustration of this alignment.

5.1 Connections between cognition and observation corners

The assessment triangle framework dictates that the development of the assessment instrument (observation corner) should be linked to theories and beliefs about how students come to understand concepts in the target domain (cognition corner). The cognition and observation corners of the assessment triangle were connected in two ways during the development of the TTCI. First, we determined what concepts were most difficult by asking experienced engineering faculty to generate a list of concepts that were difficult for their students to understand. This list was used as the basis of a Delphi survey and the same faculty were asked to rate the concepts iteratively on their importance and difficulty until a stable rating (consensus) was reached. Concepts identified as difficult in the Delphi survey were also found in the research literature on misconceptions. Concepts that were rated as most difficult and most important in the Delphi survey were used to develop questions for the TTCI. This is a direct link between the cognition

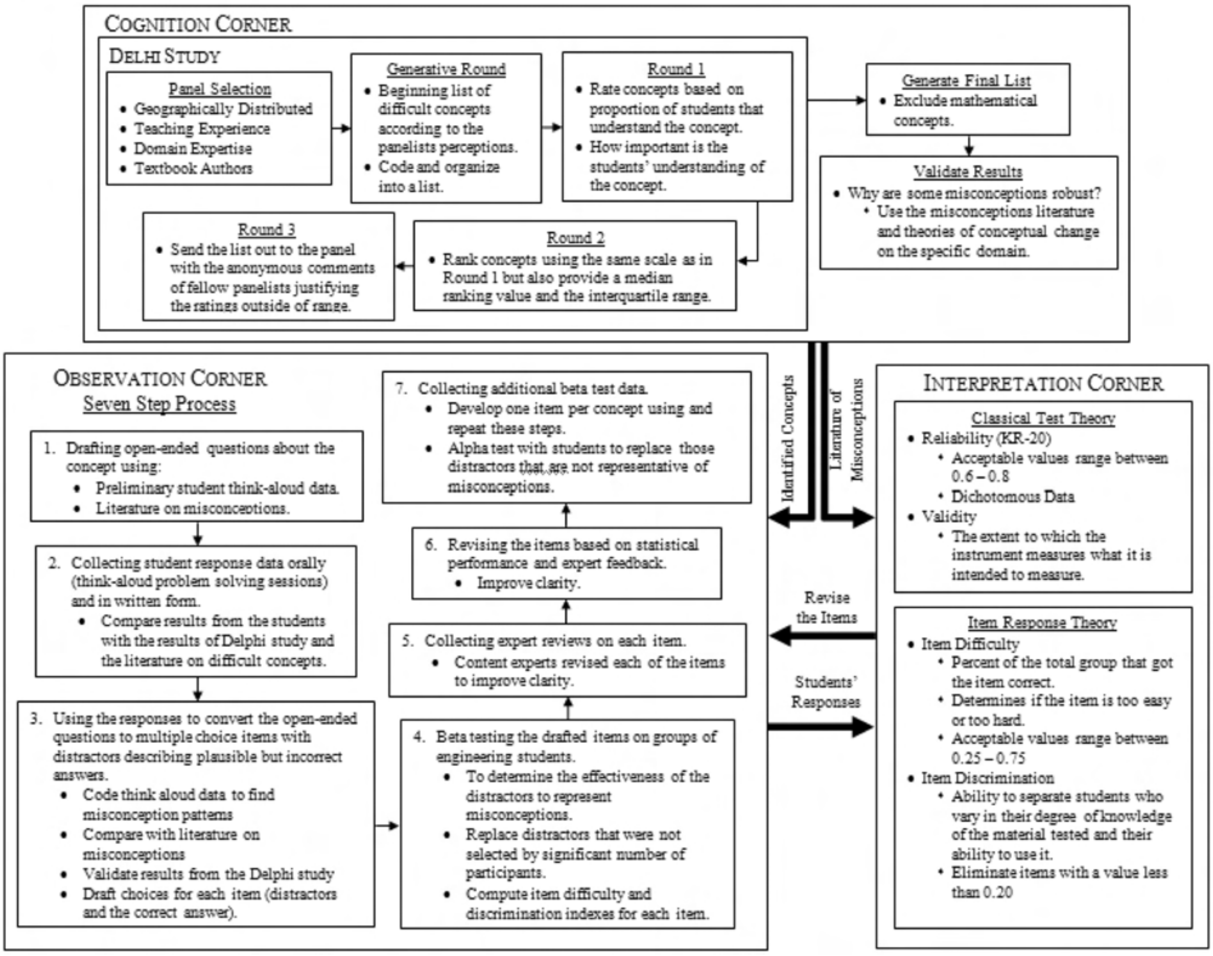

COGNITION CORNER
DELHI STUDY
Panel Selection
• Geographically Distributed
• Teaching Experience
• Domain Expertise
• Textbook Authors
Generative Round
• Beginning list of difficult concepts according to the panelists perceptions.
• Code and organize into a list.
Round 1
• Rate concepts based on proportion of students that understand the concept.
• How important is the students' understanding of the concept.
Generate Final List
• Exclude mathematical concepts.
Validate Results
• Why are some misconceptions robust?
• Use the misconceptions literature and theories of conceptual change on the specific domain.
Round 3
• Send the list out to the panel with the anonymous comments of fellow panelists justifying the ratings outside of range.
Round 2
• Rank concepts using the same scale as in Round 1 but also provide a median ranking value and the interquartile range.
OBSERVATION CORNER
Seven Step Process
1. Drafting open-ended questions about the concept using:
• Preliminary student think-aloud data.
• Literature on misconceptions.
2. Collecting student response data orally (think-aloud problem solving sessions) and in written form.
• Compare results from the students with the results of Delphi study and the literature on difficult concepts.
3. Using the responses to convert the open-ended questions to multiple choice items with distractors describing plausible but incorrect answers.
• Code think aloud data to find misconception patterns
• Compare with literature on misconceptions
• Validate results from the Delphi study
• Draft choices for each item (distractors and the correct answer).
4. Beta testing the drafted items on groups of engineering students.
• To determine the effectiveness of the distractors to represent misconceptions.
• Replace distractors that were not selected by significant number of participants.
• Compute item difficulty and discrimination indexes for each item.
5. Collecting expert reviews on each item.
• Content experts revised each of the items to improve clarity.
6. Revising the items based on statistical performance and expert feedback.
• Improve clarity.
7. Collecting additional beta test data.
• Develop one item per concept using and repeat these steps.
• Alpha test with students to replace those distractors that are not representative of misconceptions.
Identified Concepts
Literature of Misconceptions
Revise the Items
Students' Responses
INTERPRETATION CORNER
Classical Test Theory
• Reliability (KR-20)
• Acceptable values range between 0.6 – 0.8
• Dichotomous Data
• Validity
• The extent to which the instrument measures what it is intended to measure.
Item Response Theory
• Item Difficulty
• Percent of the total group that got the item correct.
• Determines if the item is too easy or too hard.
• Acceptable values range between 0.25 – 0.75
• Item Discrimination
• Ability to separate students who vary in their degree of knowledge of the material tested and their ability to use it.
• Eliminate items with a value less than 0.20

Fig. 8. Process used to develop the TTCI instrument.

and observation corners. Second a link between the cognition and observation corner was made during the creation of distractors. As discussed in Section 2.2, Chi's theory describing why some misconceptions are robust (cognition) helped guide the TTCI developers in creating distracters for the TTCI (observation). Distractors that captured common misconceptions based on Chi's theory were included whenever possible.

5.2 Connections between cognition and interpretation corners

Chi's theory as applied to robust misconceptions about heat transfer (cognition) was used to help explain the TTCI results analysis (interpretation). For example, Chi's theory can help explain why some concepts might be most resistant to instruction. Misconceptions about these concepts would be less likely to be 'repaired' after instruction and this would help to explain differences in conceptual gain for some concepts. One would also expect that the concepts rated most difficult in the Delphi survey (cognition) would also prove to be the concepts that were least likely to be answered correctly on the TTCI (interpretation corner). Thus the link between the cognition and observation corners was bidirectional.

5.3 Connection between observation and interpretation corners

The TTCI was constructed so that each item has only one correct answer and all distractors are equally incorrect. Thus the TTCI is currently scored dichotomously, with the only possible scores being 'correct' and 'incorrect.' There are no distractors that are assumed to be less incorrect than others. Because of the dichotomous construction of the TTCI (observation), CTT measures of reliability and validity (interpretation) are used. The dichotomous construction (observation) also makes the calculation of item discrimination and item difficulty (interpretation) appropriate.

Thus, all three corners of the assessment triangle were aligned during the development and testing of the TTCI. As will be discussed in the following section, expansion of our ideas about how students learn the target domain (cognition corner) is prompting the construction of revised TTCI items (observation corner), and this in turn will necessitate a different kind of analysis (interpretation corner). Thus the alignment between cognition, observation, and interpretation corners begins anew.

6. Implications and future research

It has always been our hope that the development of reliable, valid concept inventories around engineering topics would provide a yardstick to measure students' conceptual understanding of fundamental knowledge. Our intention with this paper is to assist the growing numbers of CI developers in creating sound instruments that can be widely applied for individual, program, and pedagogical assessment. The TTCI is now available online, through a password protected site, so that it can be widely used.

We have also learned that concept inventory development is a never-ending process. Items can always be refined and reliabilities increased. There is also room to adapt concept inventories for different populations. Although the fundamental phenomenon the concept the CI is testing will not vary, the language and examples necessary to measure students' understanding of those concepts accurately may well need to be adjusted for students from different cultures and with various mother tongues. As education becomes globalized, CIs will need to be modified. Since reliability is not a property of an instrument, but a statistic that speaks to the use of an instrument within a specific context, modified CIs will need to be re-evaluated and reliability re-established. We hope the assessment triangle will be used for this continuing improvement.

We have argued that the assessment triangle provides a framework to guide a rigorous methodology for concept inventory development and have proposed the TTCI as an exemplar. We want to bring special attention to the cognition corner of the assessment triangle and would urge concept inventory developers to make explicit their theories of what concepts are difficult and, especially, *why* those concepts are difficult. Insight into student thinking about these concepts will be needed to develop learning environments that help students learn these concepts, which is the ultimate goal of this research.

With our emphasis on the cognition corner, it is perhaps not surprising that our ideas about how students come to understand the fundamental concepts in our target domain (thermal and transport sciences) are evolving. We are now wondering if there is a developmental trajectory to how students learn concepts in this domain. Is conceptual understanding an all-or-nothing phenomenon? Or, as Minstrel and colleagues proposed [51], can students have varying levels of understanding? If this is the case, the TTCI must change from being a dichotomous instrument (with one correct answer and all other choices being equally incorrect) to one that can distinguish among different levels of understanding. We have begun a new project that moves the TTCI in this direction. This project employs diagnostic cognitive assessment (CDA), a method can be used to determine examinees level of performance [52]. CDA has been successfully tested with the Concept Assessment Tool for Statics [17] and is a promising method for the TTCI and other concepts inventories. Therefore, having a more nuanced view of how students come to understand concepts in our target domain (a change in the cognition corner) will bring about changes in the construction of the TTCI and the method of analysis (observation and interpretation corners).

Our long-term goal is to develop the TTCI as a diagnostic instrument that would identify areas of student difficulty. This would aid faculty in targeting their instruction to repair widely held misconceptions. This diagnostic use could be coupled with our related effort to develop materials whose aim is to repair misconceptions [53]. Thus, we could provide an individualized tool that would diagnose and then repair misconceptions. Such a tool would be a boon to instructors and students alike and address one of the National Academy of Engineering's grand challenges: to advance personalized learning [54].

7. Conclusions

The use of concept inventories has proliferated in engineering education. Our aim is to provide a model for concept inventory development that is both rigorous and aligned with current assessment theory. We use the thermal and transport concept inventory (TTCI) as an exemplar for concept inventory development as summarized in Fig. 8 in which the steps within each phase of TTCI development are depicted, as well as how the phases were connected.

We recommend the use of the 'assessment triangle' as a framework for ensuring alignment between how students learn in a target discipline (cognition corner) with appropriate assessment tasks (observation corner) and how those tasks are analyzed (interpretation corner). We would like to emphasize the importance of the cognition corner in concept inventory development. It is crucial to begin the development of any assessment by examining the literature on how students learn in the target discipline. In the case of concept inventories, describing why misconceptions exist and persist is an especially important concern. This aspect of concept inventory development has often been lacking. Use of the assessment triangle framework can lead to the creation of concept inventories that are not only valid and reliable but can also help begin to understand why misconceptions persist. This knowledge could become the foundation for creating instructional materials that help students understand some of the most conceptually difficult, and most important, content in engineering.

Acknowledgements—We wish to thank the United States National Science Foundation for supporting this work through grants DUE 0127806 'Developing an outcomes assessment instrument for identifying engineering student misconceptions in thermal and transport sciences'; ESI 0227558—'Center for the advancement of engineering education'; EEC 0550169—'Developing ontological schema training methods to help students develop scientifically accurate mental models of engineering concepts'; and DRL 0918531—'Integrating cognition and measurement with conceptual knowledge: Establishing the validity and diagnostic capacity of concept inventories.'

We gratefully acknowledge discussions with Dr. Michelene Chi at Arizona State University and the participation of Delphi survey participants. We also thank Dr. Holly Matusovich for formatting earlier drafts of this paper and Drs. Dazhi Yang, Alejandra Magaña de León, and Tameka Clarke Douglas for joining us in thinking together about revisions.

References

1. R. Duit, *Students' and Teachers' Conceptions and Science Education*, Kiel, Germany: Institute for Science Education, http://www.ipn.uni-kiel.de/aktuell/stcse/, accessed May 2009.
2. I. A. Halloun and D. Hestenes, The initial knowledge state of college physics students, *American Journal of Physics*, **53**(11), 1985, pp. 1043–1055.
3. D. Hestenes, M. Wells and G. Swackhamer, Force concept inventory, *The Physics Teacher*, **30**(3), 1992, pp. 159–166.
4. I. A. Halloun and D. Hestenes, Common sense concepts about motion, *American Journal of Physics*, **53**(11), 1985, pp. 1056–1065.
5. R. R. Hake, Interactive-engagement versus traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses, *American Journal of Physics*, **66**(1), 1998, pp. 64–74.
6. E. Mazur, Qualitative vs. quantitative thinking: are we teaching the right thing?, *Optics and Photonics News*, **3**, 1992, pp. 38.
7. D. L. Evans, G. L. Gray, S. Krause, J. Martin, C. Midkiff and B. M. Notaros, Progress on concept inventory assessment tools, *Proceedings of the 33rd Annual Frontiers in Education* (electronic), Boulder, CO, 2003.
8. T. R. Rhoads and R. J. Roedel, The wave concept inventory—a cognitive instrument based on Bloom's taxonomy, *Proceedings of the 29th Annual Frontiers in Education* (electronic), San Juan, PR, 1999.
9. J. K. Martin, J. Mitchell and T. Newell, Work in progress: analysis of reliability of the fluid mechanics concept inventory, *Proceedings of the 34th Annual Frontiers in Education* (electronic), Savannah, GA, 2004.
10. A. Jacobi, J. Martin, J. Mitchell and T. Newell, A concept inventory for heat transfer, *Proceedings of the 33rd Annual Frontiers in Education* (electronic), Boulder, CO, 2003.
11. S. Krause, J. C Decker and R. Griffin, Using a materials concept inventory to assess conceptual gain in introductory materials engineering courses, *Proceedings of the 33rd Annual Frontiers in Education* (electronic), Boulder, CO, 2003.
12. K. E. Wage, J. R. Buck, C. H. G. Wright and T. B. Welch, The signals and systems concept inventory, *IEEE Transactions on Education*, **48**(3), 2005, pp. 448–461.
13. P. S. Steif and J. A. Dantzler, A statics concept inventory: development and psychometric analysis, *Journal of Engineering Education*, **94**(4), 2005, pp. 363–371.
14. A. D. Stone, A psychometric analysis of the statistics concept inventory, Ph.D. dissertation, University of Oklahoma, Norman, OK, 2006.
15. J. Richardson, P. Steif, J. Morgan and J. Dantzler, Development of a concept inventory for strength of materials, *Proceedings of the 33rd Annual Frontiers in Education* (electronic), Boulder, CO, 2003.
16. K. C. Midkiff, T. A. Litzinger and D. L. Evans, Development of engineering thermodynamics concept inventory instruments, *Proceedings of the 31st Annual Frontiers in Education Conference* (electronic), Reno, NV, 2001.

17. A. I. Santiago Román, Fitting cognitive diagnostic assessment to the content assessment tool for statics, Ph.D. dissertation, Purdue University, West Lafayette, IN, 2009.
18. R. A. Streveler, T. A. Litzinger, R. L. Miller and P. S. Steif, Learning conceptual knowledge in engineering: overview and future research directions, *Journal of Engineering Education*, **97**(3), 2008, pp. 279–294.
19. J. Pellegrino, N. Chudowsky and R. Glaser, *Knowing What Students Know: The Science and Design of Educational Assessment,* National Academy Press, Washington, DC, 2001.
20. M. Adler and E. Ziglio, *Gazing Into the Oracle: The Delphi Method and its Application to Social Policy and Public Health*, Jessica Kingsley Publishers, London, 1996.
21. R. A. Streveler, B. M. Olds, R. L. Miller and M. A. Nelson, Using a Delphi study to identify the most difficult concepts for students to master in thermal and transport science, *Proceedings of the 110th Annual Conference of the American Society for Engineering Education* (electronic), Nashville, TN, June 2003.
22. M. J. Clayton, Delphi: A technique to harness expert opinion for critical decision-making tasks in education, *Educational Psychology*, **17**, 1997, pp. 373–386.
23. H. A. Linstone and M. Turoff, *The Delphi Method: Techniques and Applications,* Addison-Wesley Publishing Company, Reading, MA, 1996.
24. L. S. Fish and D. M. Busby, The Delphi method. In D. H. Sprenkle and S. M. Moon (eds) *Research Methods in Family Therapy*, Guilford Press, New York, 1996, pp. 469–482.
25. M. A. Nelson, M. R. Geist, R. A. Streveler, R.L. Miller, B. M. Olds, C. S. Ammerman and R.F. Ammerman, From practice to research: using professional expertise to inform research about engineering students' conceptual understanding, paper presented at the Annual Conference of the American Educational Research Association, Montreal, Quebec, Canada, April 2005.
26. B. M. Olds, R. A. Streveler, R. L. Miller and M. A. Nelson, Preliminary results from the development of a concept inventory in thermal and transport science, *Proceedings of the 111th American Society for Engineering Education Annual Conference and Exposition* (electronic), Salt Lake City, UT, June, 2004.
27. K. Carlton, Teaching about heat and temperature, *Physics Education*, **35**(2), 2000, pp. 101–105.
28. P. L. Thomas and R. W. Schwenz, College physical chemistry students' conceptions of equilibrium and fundamental thermodynamics, *Journal of Research in Science Teaching*, **35**(10), 1998, pp. 1151–1160.
29. M. F. Thomaz, I. M. Malaquias, M. C. Vanente and M. J. Antunes, An attempt to overcome alternative conceptions related to heat and temperature, *Physics Education*, **30**, 1995, pp. 19–26.
30. G. J. Posner, K. A Strike, P. W. Hewson and W. A. Gertzog, Accommodation of a scientific conception: toward a theory of conceptual change, *Science Education*, **66**(2), 1982, pp. 211–227.
31. R.L. Miller, R.A. Streveler, B. M. Olds, M. T. H. Chi, M. A. Nelson and M. R. Geist, Misconceptions about rate processes: preliminary evidence for the importance of emergent schemas in thermal and transport sciences, *Proceedings of the 113th American Society for Engineering Education Annual Conference and Exposition* (electronic), Chicago, IL, 2006.
32. S. Vosniadou, Conceptual change research: an introduction. In *International Handbook of Research on Conceptual Change*, Routledge, New York and London, 2008.
33. R. A. Streveler, R. L. Miller and B. M. Olds, Threshold concepts and troublesome knowledge in engineering: evidence from misconceptions research, *Proceedings of the Threshold Concepts Conference* (electronic), Kingston, Ontario, Canada, June 2008.
34. R. A. Streveler, R. L. Miller and J. D. Slotta, Using ontology training to investigate why some engineering science concepts are so difficult to learn, *Proceedings of on Being an Engineer: Cognitive Underpinnings of Engineering Education Conference* (electronic), Lubbock, TX, February 2008.
35. A. A. diSessa, Knowledge in pieces. In G. Forman and P. Pufall (eds.) *Constructivism in the Computer Age*, Erlbaum, Hillsdale, NJ, 1988, pp. 49–70.
36. M. T. H. Chi, Commonsense Conceptions of emergent processes: why some misconceptions are robust. *Journal of the Learning Sciences*, **14**(2), 2005, pp. 161–199.
37. 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, paper presented at the Annual Conference of the American Educational Research Association, Chicago, IL, 2007.
38. R. L. Miller, R. A. Streveler, B. M. Olds, M. A. Nelson and M. R. Geist, Concept inventories meet cognitive psychology: using beta testing as a mechanism for identifying engineering student misconceptions, *Proceedings of the American Society for Engineering Education Annual Conference* (electronic), Portland, OR, June, 2005.
39. S. M. Downing, Twelve Steps for Effective Test Development, in *Handbook of Test Development*, edited by S.M. Downing and T. M. Haladyna, Erlbaum, Mahwah, NJ, 2006, pp. 3–26.
40. D. R. Mulford and W. R. Robinson, An inventory for alternate conceptions among first-semester general chemistry students, *Journal of Chemical Education*, **79**(6), 2002, pp. 739–744.
41. M. Reiner, J. D. Slotta, M. T. H. Chi and L. B. Resnick, Naive physics reasoning: a commitment to substance-based conceptions, *Cognition and Instruction*, **18**(1), 2000, pp. 1–35.
42. R. J. Mislevy, M. R. Wilson, K. Ercikan and N. Chudowsky, *Psychometric Principles in Student Assessment*, Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles, 2002.
43. T. M. Bechger, G. Maris, H. Verstralen and A. A. Beguin, Using classical test theory in combination with item response theory, *Applied Psychological Measurement*, **27**(5), 2003, pp. 319.
44. X. Fan, Item Response theory and classical test theory: an empirical comparison of their item/person statistics, *Educational and Psychological Measurement*, **58**(3), 1998, pp. 357.
45. G. F. Kuder and M. W. Richardson, The theory of the estimation of test reliability, *Psychometrika*, **2**(3), 1937, pp. 151–160.
46. T. J. B. Kline, *Psychological Testing: A Practical Approach to Design and evaluation*, Sage, Thousand Oaks, CA, 2005.
47. C. Stage, *Classical Test Theory or Item Response Theory: The Swedish Experience*, Umeå University, Department of Educational Measurement, 2003.
48. J. L. Ndalichako and W. T. Rogers, Comparison of finite state score theory, classical test theory, and item response theory in scoring multiple-choice items, *Educational and Psychological Measurement*, **57**(4), 1997, pp. 580.
49. F. B. Baker, *Item Response Theory: Parameter Estimation Techniques*, CRC Press, Cleveland, OH, 2004.
50. R. M. Thorndike, IRT and intelligence testing: past, present, and future. In S. E. Embretson and S. L. Hershberger (eds.) *The New Rules of Measurement: What Every Psychologist and Educator Should Know*, Erlbaum, Mahwah, NJ, 1999.
51. J. Minstrell, R. Duit, F. Goldberg and H. Niedderer, Facets of students' knowledge and relevant instruction. In *Proceedings, International Workshop—Research in Physics Learning: Theoretical Issues and Empirical Studies* (electronic)*,* The Institute for Science Education at the University of Kiel, Kiel, Germany, 1992
52. J. P. Leighton and M. J. Gierl, Why cognitive diagnostic assessment? In *Cognitive Diagnostic Assessment for Education: Theory and Applications*, Cambridge University Press*,* Cambridge UK, 2007, pp. 3–18.
53. D. Yang, R. A. Streveler, R. L. Miller and A. I. Santiago Román, Repairing misconceptions: a pilot study with advanced engineering students on their use of schema training modules, *Proceedings of the 116th ASEE Annual Conference and Exposition* (electronic)*,* Austin, TX, 2009.
54. National Academy of Engineering, *Grand challenges for engineering*, http://www.grandchallenges.org, accessed July 2009.

Ruth A. Streveler is Assistant Professor in the School of Engineering Education at Purdue University. Before coming to Purdue she spent 12 years at Colorado School of Mines, where she was the founding Director of the Center for Engineering Education. Dr. Streveler earned a BA in Biology from Indiana University-Bloomington, MS in Zoology from the Ohio State University, and Ph.D. in Educational Psychology from the University of Hawaii at Manoa. Her primary research interests are investigating students' understanding of difficult concepts in engineering science and helping engineering faculty conduct rigorous research in engineering education.

Ronald L. Miller is Professor of Chemical Engineering and Director of the Center for Engineering Education at the Colorado School of Mines where he has taught chemical engineering and interdisciplinary courses and conducted research in engineering education for over 25 years. He has received three university-wide teaching awards and has 12 times been chosen as the best teacher in the Chemical Engineering department by students. He has also held a Jenni teaching fellowship at CSM. He has received the Corcoran and Wickenden awards (best papers) and the Helen Plants award (best workshop) from the American Society for Engineering Education and in 2011 received the Lifetime Achievement Award in Pedagogical Science from the Chemical Engineering Division of the American Society for Engineering Education. His current research interests focus on assessing and repairing robust engineering student misconceptions in thermal and transport sciences.

Aidsa I. Santiago-Román is Assistant Professor in the Department of Engineering Science and Materials and the Director of the Strategic Engineering Education Development (SEED) Office at the University of Puerto Rico, Mayaguez Campus (UPRM). Dr. Santiago earned a BA (1996) and MS (2000) in Industrial Engineering from UPRM, and Ph.D. (2009) in Engineering Education from Purdue University. Her primary research interest is investigating students' understanding of difficult concepts in engineering science with underrepresented populations. She also teaches introductory engineering courses such as Problem Solving and Computer Programming, Statics, and Mechanics.

Mary A. Nelson is senior instructor in the Applied Mathematics Department of the University of Colorado, Boulder, USA where, in addition to teaching, she is the Director of Assessment and conducts educational research to transform the teaching of introductory calculus classes. She has received numerous teaching awards, including the 2006 College of Engineering Peebles Award and the 2010 Boulder Faculty Association Teaching Award. Her current research focuses on helping students develop deeper conceptual understanding through the use of small group oral reviews.

Monica R. Geist earned her Ph.D. in Applied Statistics and Research Methods at the University of Northern Colorado in 2008. She teaches mathematics at Front Range Community College in Westminster, Colorado, USA. Her methodological interests include mixed methods research and evaluation, instrument development, focus groups, and any type of educational research.

Barbara M. Olds is acting Deputy Assistant Director in the Directorate for Education & Human Resources at the United States National Science Foundation (NSF) in Washington, DC. Prior to joining NSF on a full-time basis, she served as Associate Provost for Educational Innovation and Professor of Liberal Arts and International Studies at the Colorado School of Mines. She returned to CSM in 2006 after spending three years at NSF where she served as the Division Director for the Division of Research, Evaluation and Communication (REC) in the Education & Human Resources Directorate. During the 2006–2007 academic year she was a visiting professor in Purdue University's Engineering Education Department. Her research interests are primarily in understanding and assessing engineering student learning. She has participated in a number of curriculum innovation projects and has been active in the engineering education research and assessment communities. She is a Fellow of the American Society for Engineering Education and was a Fulbright lecturer/researcher in Sweden.