Continuous Improvement of a Concept Inventory: Using Evidence Centered Design to Refine the Thermal and Transport Concept Inventory

Dr. Brian Douglas Gane, University of Illinois at Chicago

Dr. Brian Gane is a Visiting Research Assistant Professor at the Learning Sciences Research Institute, University of Illinois at Chicago. Dr. Gane's research focuses on psychological theories of learning and skill acquisition, assessment, instructional design, and STEM education.

Ms. Dana Denick, National Science Foundation
Miss Natalie Jorion
Prof. Louis V DiBello
Prof. James W Pellegrino, University of Illinois, Chicago
Dr. Ruth A. Streveler, Purdue University, West Lafayette

Continuous Improvement of a Concept Inventory: Using Evidence Centered Design to Refine the Thermal and Transport Concept Inventory

Concept inventories (CIs) are increasingly being developed and used in engineering courses to assess student learning and understanding and to evaluate instructional practices. CIs differ from typical STEM assessments in that they tend to focus on a small set of key constructs and conceptual understandings within the domain—such as "the concept of force" in physics (FCI[1]); the area of "statics" (CATS[2]); or "digital logic" in computer science (DLCI[3]). The questions are frequently based on science and engineering education research, including research on misconceptions and common student errors.

CIs often have substantial research guiding their development. Nevertheless, validating an assessment involves explicating the proposed uses and interpretation of test scores and marshaling evidence to support the acceptability and plausibility of particular claims about the meaning of those scores.[4] As part of a larger research project we have developed an analytic framework to assess the validity of classroom assessments such as CIs.[5] In developing this framework we have conducted validity analyses with four CIs: Conceptual Assessment Tool for Statics, Statistics Concept Inventory (SCI),[6] Dynamics Concept Inventory (DCI),[7] and TTCI. The results of the analyses indicated that the different inventories have varying levels of evidentiary support for claims as to the inferences one can make from scores.[5]

History of TTCI and Potential for Redesign

The Thermal and Transport Concept Inventory (TTCI) was developed in iterative cycles over multiple years. The TTCI has three sections (thermodynamics, fluid mechanics, and heat) and items were created after subject matter experts performed a domain analysis and following several rounds of development and testing.[8,9] Instructors can use any combination of the three sections and receive reports on their students' performance on each of the three sections.

With respect to our validity analysis of the TTCI, there was moderate support for some of the authors' claims about the meaning and use of inventory scores. In fact, the evidence suggested that the instrument had a solid conceptual foundation but might be improved to better support specific interpretive claims. For example, development of the initial version of the TTCI arose from a strong design process; the developers employed a Delphi process whereby content experts identified the critical ideas in the domain.[8,9] In addition, preliminary analysis of TTCI student performance data indicated that some of the evidentiary weaknesses might be related to a limited number of items per content category; two of the categories had fewer than five items. Further instrument development was supported by interest from the TTCI developers in refining their instrument, enabling collaboration between domain experts, learning scientists, and psychometricians.

For reasons outlined above, we choose to redesign the TTCI. The redesign effort focused on only the heat section of the TTCI. The original heat section contained 18 items spanning three central concepts: steady state & thermal equilibrium (11 items), heat & transfer of energy (3 items), temperature & amount of energy (4 items). This distribution of items illustrates the difficulty of using TTCI scores to create structural models of students' conceptual knowledge – two of the categories have only 3 or 4 items. From a measurement perspective, estimating students' proficiency on a concept (e.g., heat & transfer of energy) is difficult when there are relatively few observations (items). Such measurement is critical if the CI is to be used to make inferences about students' understanding of a single concept and/or for diagnostic purposes.[5] We reasoned that one main goal for the redesign would be to ensure that there would be sufficient items for each category/concept that was a target for measurement and reporting.

Method of Redesign

We did not immediately start writing items for the categories that were deficient. Instead, we began a more fundamental instrument (re)design effort, based on application of evidence-centered design (ECD) principles. One guiding principle in ECD is that task design is evidence-driven; after identifying the concepts that students should understand, one works backwards to consider what evidence could be used to support the claim that students understand the concept, and based on the evidence one seeks, one then designs assessment items (tasks) that could produce that evidence.[10] ECD is a powerful method for developing assessments because it provides a structured design process and it aligns well with contemporary views of assessment as an inferential process of reasoning from imperfect evidence, and is a means of implementing the general rationale described by the assessment triangle.[11]

Several critical procedures and principles of our redesign emerged over this multi-year effort. We first provide a sketch of these principles because we think they will be useful to other CI developers. We then explore three of these redesign procedures in depth, using examples from our TTCI work.

Overall, our development process was **iterative.** Including the initial version of the TTCI, we produced and administered three versions of the heat section of the TTCI. Separating production of these three versions of the instrument were multiple rounds of iterative revision of individual TTCI items. Consistent with ECD, our design work was built upon a **domain model** that emerged from **continuous domain analysis**. A domain model identifies the concepts in the domain and forces one to explicate claims about what "proficiency" in the domain looks like.[10] The initial domain analysis was completed using a Delphi process,[8] however, an ECD-formatted domain model had not been built. Thus, we used the existing TTCI heat section to reverse-engineer its domain model. We created a design pattern that, among other things, described the *focal knowledge, skills, and abilities* (FKs) that we were interested in assessing. These FKs included both normative scientific concepts and also potential student misconceptions (i.e., alternative frameworks or preconceptions). Importantly, this domain model was not static; as we developed and refined items we revised the domain model. Therefore, we continued our domain analysis as we simultaneously developed items. Often, this domain analysis was advanced through item review meetings and cross-disciplinary discussions about the structure of the domain and the critical concepts represented therein.

Additionally, we used **multiple methods and data sources**. We had item review panels in which multiple people considered features of items' comprehensibility and content alignment. We conducted think-aloud studies in which engineering students explained their thinking as they worked through items. These studies were helpful in identifying comprehensibility issues and with providing evidence that student thinking when completing the task (item) was consistent with the type of reasoning that developers intended.[12] Additionally, we completed two rounds of data collection with the full instrument (i.e., all items) and conducted quantitative analyses of these larger data sets to assess item and instrument performance.[13, 14] These analyses results helped us revise items and update the domain model. Finally, we used a **Q-matrix** to document the mapping between item answers (correct and incorrect) with the normative concepts and misconceptions specified in our evolving domain model.

Examples of Redesign Processes

Q-matrix

To provide more detail, we walk through examples of some of these critical redesign processes. First, one product that became useful in designing the final instrument was a Q-matrix that we updated throughout the redesign. A Q-matrix[15, 16] is similar to a table of specifications[17] except that it is a matrix of concepts (horizontal) and items (vertical). A Q-matrix can be used to represent the mapping between items and FKs. We had two different versions of Q-matrices, one at the *item level* and one at the *item response level* (e.g., “A”, “B”, “C”, etc.; our items were multiple-choice). Table 1 shows a portion of one of our item level Q-matrices. In this table, we have four items, four concepts (“FK.c#”), and four misconceptions (“FK.m#”). The cells are coded dichotomously: a “1” indicates that solving the item requires proficiency with that concept. An item can be coded for multiple concepts (e.g., Item 28). In contrast, Table 2 shows a portion of one of our item response level Q-matrices. The items, concepts, and misconceptions are the same in Tables 1 and 2, but Table 2 has a row for each items’ response option. An “X” indicates that the concept (or misconception) is required to endorse that response option. An “-x” indicates that an *incorrect* application of that concept is required to endorse that response option. This coding allows us to map all of an item’s correct and incorrect responses to conceptions and misconceptions. Table 1 is constructed from Table 2, where “-x” codings are ignored and the presence of 1 or more “X” per item is coded as “1”.

Our Q-matrix served as running log of our alignment between draft items and our domain model, and pointed to domain categories that still needed items to be developed. Additionally, we used it as a focal artifact around which we could conduct item reviews. Finally, it served as a guide when we needed to select items (and remove items) for the final instrument. In the future, the Q-matrix will allow us to conduct diagnostic classification modeling.[16,18] Our next two examples of redesign activities further explain our process; importantly, each activity is rooted in using the Q-matrix.

Table 1. Portion of a Q-matrix at item level.

Item	FK.c1	FK.c2	FK.c3	FK.c4	FK.m1	FK.m2	FK.m3	FK.m4
4.A	1	0	0	0	0	0	0	1
5.K	1	1	0	0	1	1	0	0
28	1	0	1	0	0	0	1	0
9b	0	0	0	1	0	0	0	1

Table 2. Portion of a Q-matrix at item response level.

Item	Response option	answer/ (dis)tractor	FK.c1	FK.c2	FK.c3	FK.c4	FK.m1	FK.m2	FK.m3	FK.m4
4.A	a	dis								X
4.A	b	dis								X
4.A	c	answer	X							
4.A	d	dis								X
4.A	e	dis								X
5.K	a	answer	X	X						
5.K	b	dis	-x					X		
5.K	c	dis					X			
5.K	d	dis								
5.K	e	dis								
28	a	answer	X		X					
28	b	dis			-x				X	
28	c	dis			-x					
28	d	dis			-x				X	
9b	a	dis		-x						
9b	b	dis								X
9b	c	dis				-x				
9b	d	answer				X				
9b	e	dis	-x							

Item review panels

Developing and refining items occurred in multiple ways, but often resulted directly from item review panels. Item review happened at two levels before students ever saw items. After an initial draft of an item was created, several project members first reviewed the item for clarity and comprehensibility issues. Following this, we reviewed an item for its alignment with the FKs specified in the domain model. We mapped each item response to the FKs in the domain model, and recorded this mapping using our Q-matrix format (e.g., Table 2). This mapping was always done independently by at least two raters (sometimes three) who have a background in thermal science. When raters disagreed, they, along with other team members, discussed and ultimately resolved the disagreement. These resolutions took different forms: Sometimes the item stem and/or response would be revised, sometimes the mapping to an FK would be revised, and sometimes the FK itself (wording or conceptual category) would be revised so that it could be consistently applied and distinguished from other FKs. In a few cases, team members could not agree on a single mapping and so we included those items in think-aloud studies. We then used the data from students' conceptual reasoning to help decide on an appropriate mapping. The Q-matrix codings were continuously updated to reflect the modifications we made to the domain model and/or the items.

Domain analysis & model

One substantial portion of the redesign effort was revision of the domain model. This revision occurred as a result of continued, iterative domain analysis, using a Q-matrix to track alignment between items and FKs, conducting think-aloud studies using individual TTCI items, and conducting larger-scale pilot studies using the full set of TTCI Heat items. Revisions to the domain model occurred primarily in terms of the number and content of the FKs. This included revising the boundaries of the FK, creating new FKs, and condensing multiple FKs into a single FK.

Our first step was to develop a domain model in terms of FKs and then map the original TTCI items to these FKs using a Q-matrix. One conceptual distinction between our domain model and the original TTCI domain model is that we allowed an item to measure more than one concept (FK). In the original version of the heat section of the TTCI, items were grouped according to topics, with each item mapped to a single topic. In our Q-matrix, we allowed an item to map to more than 1 FK if needed. This flexibility is consistent with how we conceptualize FKs; they are "knowledge, skills, and abilities" that students must have to answer a question. It is reasonable (and common for complex items) that successfully answering the item requires more than one piece/type of knowledge, skill, and ability. For instance, one item asked students to reason about why it feels cooler to walk across a tile floor than carpet. This item deals with both the rate of energy transfer and sensation of hot and cold. In our first version of the Q-matrix these two ideas were separate FKs, so the item mapped to both. In the developers' original domain model this item was linked to a single topic "heat and transfer of energy." Table 3 shows the original TTCI Heat Section recoded to our ECD domain model using a Q-matrix. Note that two of the original 18 items were dropped, leaving 16 items, and 5 FKs (labeled in this table as "Skills"). Our first domain model did not separate student misconceptions. With these 16 items and 5 FKs there were still several FKs that had too few items (e.g., Skill 2 and Skill 5). We thus began to write more items and revise existing items.

In authoring and reviewing these new items, we had many discussions where the conceptual basis for an FK was clarified (i.e., what the FK "means" was modified). Also, in critiquing newly written items we sometimes had to add novel FKs to represent conceptual ideas that we were not previously considering. In other words, the item review served as a knowledge elicitation method[19] that continued the domain analysis and allowed us to update the domain model. Based on writing these items and then aligning them to our domain model (revising our domain model or items when necessary) we created a new domain model that had 12 FKs and 5 misconception FKs (see Table 4). In other words, as we were developing new items we were also refining our domain model by updating and adding FKs and misconceptions.

Table 3. Original version of the TTCI Heat section, recoded into domain model with 5 FKs.

Item	Skill1	Skill2	Skill3	Skill4	Skill5
5.A	1				
4.A	1				1
5.Qp	1	1			
5.Qp	1	1			
5.B	1				1
5.Gp			1		
5.Gp			1		
7.Fp				1	
7.Fp				1	
7.Cp				1	
7.Cp				1	
4.F	1				
4.G		1			
5.KA	1				
5.Op			1		
5.Op			1		

We piloted a second version of the TTCI heat section that had 10 new items and 9 stalwarts from the original TTCI heat section. We were primarily interested in collecting item performance metrics on these items so that we could determine if any were problematic or poorly functioning. A secondary consideration was to complete structural analyses. This second version had items that performed well but unfortunately did not clearly demonstrate the hypothesized conceptual structure. In hindsight, one of the limitations of our second version was that we did not use the Q-matrix to guide which items to select for the 19-item test. Although we had been using the Q-matrix to represent item-to-FK mappings, we only aggregated item counts across FKs and misconceptions at the data analysis stage, not at the instrument construction phase. We modified this process for our next version of the TTCI section so that we kept one eye on the item revisions and kept one eye on how those items were mapping to the domain model (i.e., if we had too many or too few items for a given FK). Additionally, we refined our domain model so that it had only a small number of critical concepts (FKs) and misconceptions. It was productive for our domain analysis to add FKs, but to create a final instrument we needed to simplify our FKs to focus only on core ideas in the domain. We wanted to accurately measure a few core ideas, but to do so we needed enough items per concept, a somewhat equivalent number of items across concepts, and an instrument that would take less than 50 minutes to complete (so that students could complete it within one class session if needed).

To create the final instrument with enough items for only a few core ideas in the domain, we distilled, condensed, and eliminated FKs (and misconceptions). We achieved this through using the Q-matrix to see which concepts had many items and which had few, through talking with item developers to see which concepts were difficult to write multiple items for, and conversations with domain experts to decide which concepts were really fundamental and which concepts were more fine-grained, that might be able to be integrated into more core ideas. In the end, we settled on a domain model with four concepts and four misconceptions (see Table 5).

Table 4: Expanded TTCI ECD Focal Knowledge and Misconceptions

FK1	***Rate*** – movement of heat (thermal energy) from a hot object to a colder object is not instantaneous, but occurs at a rate dependent on key factors
FK2	Heat transfer by ***conduction*** – the rate is proportional to temperature gradient, heat transfer contact area, and physical properties of materials involved
FK3	Heat transfer by ***convection*** – the rate is proportional to temperature difference, contact area, fluid velocity, and physical properties of the fluid
FK4	***Thermal conductivity*** – material property that describes how well the material will conduct heat; heat transfer rate is proportional to thermal conductivity
FK5	***Heat capacity*** – material property that describes how much energy a unit of an object can absorb (or release) per unit of temperature change; relates temperature change to the change in the amount of thermal energy stored
FK6	***Amount*** – thermal energy stored in an object is proportional to mass or volume of the object
FK7	***Surface area to volume ratio*** – determines the rate of temperature change in a solid object that is cooling or being heated
FK8	***Temperature*** – macroscopic property that emerges from the random motion of molecules with a narrow distribution of kinetic energies (i.e. temperature emerges as a measurable property of random molecular motion)
FK9	***Thermal resistance*** – reduces the rate of conductive or convective heat transfer
FK10	***Thermal equilibrium*** – two objects or systems are in thermal equilibrium if they are at the same temperature (so no net heat transfer occurs between the two)
FK11	***Steady-state and unsteady-state –*** an object or system is at steady-state if <u>none</u> of its variables (e.g. temperature, pressure, composition, flowrate) are changing with time at a fixed position in the object or system; an object or system is at unsteady-state if at least one of its variables (e.g. temperature, pressure, composition, flowrate) is changing with time at a fixed position in the object or system
FK12	***Phase change*** – phase change of a substance (e.g. evaporation, condensation) at a solid surface increases the rate of heat transfer at the surface
FK13	**Misconception**: Energy and temperature are equivalent (ignores heat capacity); temperature is not the only variable that determines the thermal energy content of an object or system (heat capacity and phase are also important variables)
FK14	**Misconception:** More energy is exchanged if the rate is faster (rate = amount)
FK15	**Misconception:** A system at steady-state is not transferring energy to another system or the surroundings
FK16	**Misconception:** Heat is a substance transferred between objects or stored in an object
FK17	**Misconception:** Temperature sensation measures how ‘cool’ or ‘warm’ an object feels; feeling colder means being colder; wind chill effect; heat index effect
FK18	**Misconception:** Steady-state and thermal equilibrium occur together; one can’t occur without the other

Table 5. Condensed TTCI ECD Focal Knowledge and Misconceptions

FK.c1. ***Rate processes* –** Movement of heat (thermal energy) from a hot object to a colder object is not instantaneous, but occurs at a rate dependent on key factors including temperature gradient, heat transfer contact area, and physical properties of materials involved; includes convective effects of fluid properties and object geometry (surface area to volume ratio).
FK.c2. ***Amount* –** Thermal energy stored in an object is proportional to the mass or volume of an object; heat capacity describes how much energy a unit of an object can absorb (or release) per unit of temperature change.
FK.c3. ***Thermal equilibrium* vs. *Steady-state* –** steady state refers to system parameters (temperature and heat transfer rate) which do not change with time at a specific position in the system. Thermal equilibrium refers to two systems or bodies that are at the same temperature, so no net heat transfer occurs between them.
FK.c4. ***Phase change* –** Phase change of a substance (e.g. evaporation, condensation) at a solid surface increases the rate of heat transfer at the surface.
FK.m1. ***Misconception*:** Energy and temperature are equivalent (ignores heat capacity); temperature is not the only variable that determines the thermal energy content of an object or system (heat capacity and phase are also important variables).
FK.m2. ***Misconception*:** More energy is exchanged if the rate is faster (rate = amount).
FK.m3. ***Misconception*:** A system at steady-state is not transferring energy to another system or the surroundings; steady-state and thermal equilibrium occur together; one can't occur without the other.
FK.m4. ***Misconception*:** Temperature sensation measures how 'cool' or 'warm' an object feels; feeling colder means being colder; wind chill effect; heat index effect.

Now that we had a clean domain model and many items from different rounds of item authoring and revisions we needed to select a final set of items to include in the Q-matrix. Our goal was to have at least 6 items per concept, but allow concepts to have more items since an item can map to multiple FKs. Using the Q-matrix, results from our previous analyses of item performance (quantitative pilot tests and think-aloud studies) and our experts' rational analysis, we selected 24 items for the instrument. Table 6 shows a summary Q-matrix where each item (not each response) is coded to the respective FKs. This third version of the TTCI Heat section has seven items that were retained from the original version.

Next Phase of Development and Validation

We are currently completing field-testing with this version of the instrument, and to date have collected data from almost 100 undergraduate students at four universities. In the next year, we hope to collect data from approximately 500 students so that we can do more extensive validity analyses. Based on those data and analyses, we will then make final changes to the inventory. We expect this next set of revisions to be minor compared to the prior design efforts. Specifically, we will identify whether any items perform poorly and need to be removed.

Additionally, we plan to conduct structural analyses (e.g., exploratory and confirmatory factor analyses) to test the hypothesized conceptual structure represented in our domain model (i.e., the 4 FKs). The Q-matrix specifies a priori predictions about the structure of the domain represented by the items. We can use the Q-matrix codings to specify the item loadings in our confirmatory factor analysis (i.e., items coded as ‘1’ are hypothesized to load an that factor). Moreover, we can assess the diagnostic capacity of the TTCI by using the version of Q-matrix with item response levels (see Table 2) to create diagnostic cognitive models that take into account which multiple-choice option was endorsed.[18] We have done similar structural analyses with other concept inventories,[5] but to do this with the TTCI will require a large sample.

Table 6. Q-matrix for the latest iteration of the TTCI: Heat section

Item	Rate processes (FK.c1)	Amount (FK.c2)	Thermal equilibrium vs. steady state (FK.c3)	Phase change (FK.c4)
4.A	1	0	0	0
4.F	1	0	0	0
1-flat	1	0	0	0
8	1	0	0	0
14	1	0	0	0
5.G	0	1	0	0
5.O	0	1	0	0
2a	0	1	0	0
6	0	1	0	0
5.A	1	1	0	0
5.K	1	1	0	0
5.Q	1	1	0	0
29	0	0	1	0
30	0	0	1	0
35	0	0	1	0
36	0	0	1	0
9a	1	0	1	0
28	1	0	1	0
9b	0	0	0	1
11a	0	0	0	1
32	0	0	0	1
37	0	0	0	1
16a	1	0	0	1
33	0	0	1	1
Total	**11**	**7**	**7**	**6**

These future item, test, structural, and diagnostic analyses will provide the empirical evidence needed for a robust validity argument.[5] We expect the structural analyses will support the use of sub-scale scores (one sub-scale for each FK) and that the diagnostic modeling will support the use of the item responses to diagnose problematic student thinking. Having empirical support for the use of either sub-scale scores or the diagnostic capacity would highlight the utility of the TTCI heat section redesign.

Conclusion

Our development process ensured that all three vertices of the assessment triangle (cognition, observation, and interpretation)[11] and their interconnections were fully considered and worked in concert. The cognition corner (our domain model) was constantly modified through continuous domain analysis. The observation corner (tasks/items) was reflected in item revision and item review processes in which careful consideration was given to whether an item elicited the desired type of evidence about student thinking. This process was aided by creating a design pattern[20] to describe features of a class of items. The interpretation corner explicates how observations are used as evidence for students' competency, and is partly represented by development of the Q-matrix. Statistical models will need to be applied to student performance data as part of interpretation. Not surprisingly, because the assessment triangle expresses interrelationships and dependencies among cognition-observation-interpretation, as we changed one part of our assessment argument (model, task, or evidence rules) we had to modify the other parts of the assessment argument accordingly. In the end, we feel the redesign equally valued all three vertices of assessment triangle because we extended and refined the domain analysis, iteratively developed items, and used a Q-matrix to keep track of item and instrument design.

The current version of the instrument is available at www.thermalinventory.com. If you are an instructor (or researcher) and want to use it with your students, please contact Dr. Ronald Miller (rlmiller@mines.edu) and we can provide access. Students can complete the instrument during class or outside of class, and we will generate score reports for the class' overall performance. Some instructors may wish to use the instrument in a pre/post design while others may use it only once a semester for formative or summative purposes.

While CIs are increasingly being developed in engineering domains,[21,22] their use to advance instructional change lags behind. We hope that our rigorous design process and forthcoming validity analyses of the new version of the TTCI heat section will spur innovative uses of the TTCI (and other CIs) by instructors in in their classrooms and as part of research-based educational interventions. For instance, in mechanical engineering, faculty have used the Statics Concept Inventory to test the effectiveness of an intervention using worksheets and class discussions to identify and rectify misconceptions.[23]

Acknowledgements

We thank the National Science Foundation for funding this work through grant NSF 0918531, 0918436, 0918552, and 0920242. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

[1] Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30*, 141–166.
[2] Steif, P. S. & Dantzler, J. A. (2005). A statics concept inventory: Development and psychometric analysis. *Journal of Engineering Education, 94*, 363-371.
[3] Herman, G. L., Zilles, C., & Loui, M. C. (2014) A psychometric evaluation of the digital logic concept inventory, *Computer Science Education, 24*, 277-303, DOI: 10.1080/08993408.2014.970781
[4] Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1-73.
[5] Jorion, N., Gane, B. D., James, K., Schroeder, L., DiBello, L. V., & Pellegrino, J. W. (under review). An analytic framework for evaluating the validity of concept inventory claims.
[6] Stone, A., Allen, K., Rhoads, T. R., Murphy, T. J., Shehab, R. L., & Saha, C. (2003). The statistics concept inventory: A pilot study. In *Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference* (Vol. 1, pp. T3D-1–T3D-6). IEEE.
[7] Gray, G. L., Costanzo, F., Evans, D., Cornwell, P., Self, B., & Lane, J. L. (2005). The dynamics concept inventory assessment test: A progress report and some results. In *Proceedings of the 2005 American Society for Engineering Education Annual Conference & Exposition* (electronic). ASEE.
[8] Olds, B. M., Streveler, R. A., Miller, R. L., & Nelson, M. A. (2004). Preliminary results from the development of a concept inventory in thermal and transport science, In *Proceedings of the 2004 American Society for Engineering Education Annual Conference and Exposition* (electronic).
[9] Streveler, R. A., Miller, R. L., Santiago-Román, A. I., Nelson, M. A., Geist, M. R., Olds, B. M. (2011). Rigorous methodology for concept inventory development: Using the 'assessment triangle' to develop and test the Thermal and Transport Science Concept Inventory (TTCI), *International Journal of Engineering Education, 27*, 968–984.
[10] Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–67.
[11] National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Committee on the Foundations of Assessment, J. Pellegrino, N. Chudowsky, & R. Glaser (Eds.) Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences Education. Washington, DC: The National Academies Press.
[12] Denick, D. & Streveler, R. (2014, May). Qualitative analyses of students' conceptual reasoning. In J. W. Pellegrino (Chair), *Evaluating and Improving Concept Inventories as Assessment Resources in STEM Teaching and Learning*. Symposium conducted at the meeting of the American Educational Research Association, Philadelphia, PA.
[13] Jorion, N., James, K., Schroeder, L., DiBello, L. V., & Pellegrino, J. W. (2013, April). Statistical and diagnostic analyses of student performance on concept inventories. In J. W. Pellegrino (Chair), *Evaluating the Validity of Concept Inventories as Aids for STEM Teaching and Learning*. Symposium conducted at the meeting of the American Educational Research Association, San Francisco, CA.
[14] Jorion, N., Gane, B. D., James, K., Schroeder, L., Pellegrino, J., & DiBello L. (2014, May). Quantitative analyses of student performance on concept inventories. In J. W. Pellegrino (Chair), *Evaluating and Improving Concept Inventories as Assessment Resources in STEM Teaching and Learning*. Symposium conducted at the meeting of the American Educational Research Association, Philadelphia, PA.
[15] Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49*, 95-100.
[16] DiBello, L., Stout, W., & Roussos, L. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates.
[17] Oosterhof, A. (2001). *Classroom applications of educational measurement*. Upper Saddle River, New Jersey: Prentice-Hall, Inc.
[18] DiBello, L. V., Henson, R. A., & Stout, W. F. (2015). A family of generalized diagnostic classification models for multiple choice option-based scoring. *Applied Psychological Measurement, 39*, 62-79. DOI: 10.1177/0146621614561315
[19] Cooke, N. J. (1994). Varieties of knowledge elicitation techniques. *International Journal of Human-Computer Studies, 41*, 801-849.
[20] Denick, D., Miller, R., Streveler, R., Pellegrino, J., DiBello, L. Schroeder, L. Gane, B. D., James, K. & Jorion, N. (2014, April). The use of evidence centered design to reverse engineer an existing CI and develop new questions for diagnostic reporting. In J. W. Pellegrino (Chair), *Evaluating and Improving Concept Inventories as Assessment Resources in STEM Teaching and Learning. Symposium conducted at the American Educational Research Association annual meeting*, Philadelphia, PA.

[21] Reed-Rhoads, T., & Imbrie P. K. (2008). Concept inventories in engineering education. Paper presented at the *National Research Council's Workshop Linking Evidence to Promising Practice in Stem Undergraduate Education*, Washington, DC.

[22] National Research Council. (2012). Discipline-based education research: Understanding and improving learning in undergraduate science and engineering. S.R. Singer, N.R. Nielsen, and H.A. Schweingruber (Eds.) Committee on the Status, Contributions, and Future Directions of Discipline-Based Education Research. Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

[23] Steele, K. M., Brunhaver, S. R., Sheppard, S. D. (2014). Feedback from in-class worksheets and discussion improves performance on the statics concept inventory. *International Journal of Engineering Education, 30*, 992-999.