8-1-2013

Development and evaluation of a thermochemistry concept inventory for college-level general chemistry

David A. Wren

UNIVERSITY OF NORTHERN COLORADO

Greeley, Colorado

The Graduate School

DEVELOPMENT AND EVALUATION OF A THERMOCHEMISTRY CONCEPT INVENTORY FOR COLLEGE-LEVEL GENERAL CHEMISTRY

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy

David A. Wren

College of Natural and Health Sciences
Department of Chemistry and Biochemistry
Chemical Education

August 2013

This Dissertation by: David Wren

Entitled: *Development and Evaluation of a Thermochemistry Concept inventory for College-Level General Chemistry*

has been approved as meeting the requirements for the Degree of Doctor of Philosophy in College of Natural and Health Sciences in Department of Chemistry and Biochemistry, Program of Chemical Education

Accepted by the Doctoral Committee

_____
Jack Barbera, Ph.D., Research Advisor

_____
David L. Pringle, Ph.D., Committee Member

_____
Kimberly A. O. Pacheco, Ph.D., Committee Member

_____
Steven Pulos, Ph.D., Faculty Representative

Date of Dissertation Defense _____

Accepted by the Graduate School

_____
Linda Black, Ed.D., LPC
Dean of the Graduate School and International Admissions

ABSTRACT

Wren, David A. *Development and Evaluation of a Thermochemistry Concept inventory for College-Level General Chemistry*. Published Doctor of Philosophy dissertation, University of Northern Colorado, 2013.

The research presented in this dissertation culminated in a 10-item Thermochemistry Concept Inventory (TCI). The development of the TCI can be divided into two main phases: qualitative studies and quantitative studies. Both phases focused on the primary stakeholders of the TCI, college–level general chemistry instructors and students. Each phase was designed to collect evidence for the validity of the interpretations and uses of TCI testing data. A central use of TCI testing data is to identify student conceptual misunderstandings, which are represented as incorrect options of multiple–choice TCI items. Therefore, quantitative and qualitative studies focused heavily on collecting evidence at the item–level, where important interpretations may be made by TCI users.

Qualitative studies included student interviews ($N = 28$) and online expert surveys ($N = 30$). Think-aloud student interviews ($N = 12$) were used to identify conceptual misunderstandings used by students. Novice response process validity interviews ($N = 16$) helped provide information on how students interpreted and answered TCI items and were the basis of item revisions. Practicing general chemistry instructors ($N = 18$), or experts, defined boundaries of thermochemistry content included on the TCI. Once TCI

items were in the later stages of development, an online version of the TCI was used in expert response process validity survey ($N = 12$), to provide expert feedback on item content, format and consensus of the correct answer for each item.

Quantitative studies included three phases: beta testing of TCI items ($N = 280$), pilot testing of the a 12-item TCI ($N = 485$), and a large data collection using a 10-item TCI ($N = 1331$). In addition to traditional classical test theory analysis, Rasch model analysis was also used for evaluation of testing data at the test and item level. The TCI was administered in both formative assessment (beta and pilot testing) and summative assessment (large data collection), with items performing well in both. One item, item K, did not have acceptable psychometric properties when the TCI was used as a quiz (summative assessment), but was retained in the final version of the TCI based on the acceptable psychometric properties displayed in pilot testing (formative assessment).

ACKNOWLEDGMENTS

I have many people to thank for my growth, both personally and professionally, during the last four years. To save time, I will thank those who have been my biggest supporters, as well as those who are most likely to actually read this section.

To Jack, I have much to say, but will keep it brief, as I need to turn this dissertation in soon, and you know I've never been great with deadlines. You are the reason I applied to the chemical education program at the University of Northern Colorado (UNC) and the reason I am leaving UNC for a faculty position in great chemistry department. I believe my growth as a researcher, teacher, and future faculty member are the result of your hard work and dedication. I have learned much from observing you as a new faculty member, researcher, advisor and teacher—probably more than you know. However, your biggest gift to me was your trust and faith in my ability and the freedom you provided me in my graduate studies. Becoming a husband and a father during my time at UNC was not without challenges, and I felt fully supported taking on these new roles. The last year I was able to spend with Charlie was invaluable. Amy and I are both indebted to your faith in my ability to get things done on a non-traditional work schedule.

To Amy, you have been my constant supporter throughout my entire graduate career. I am glad I now have something to show for your constant encouragement, specifically, a real job. Of all the wonderful memories I will leave Colorado with,

meeting you that one summer night in Boulder will be the one for which I am most grateful. Your support in our new adventure in North Carolina means so much to me, as you are making more sacrifices than either one of us are willing to admit. Your love and support has made it possible for me to get through this last year, which has been the most rewarding and challenging year of my life

To Charlie, if you are reading this, it probably means you have found a dusty book with your dad's name on it. Though you will not remember me as a graduate student, I will remember the encouragement you provided every time you laughed. Even if you take after your mom and have no interest in chemistry, I hope you will look at this dissertation and know that if you do what you love, anything is possible. Find what you love, no matter how unorthodox or counter–intuitive it may be, and you will never regret your decision to follow your passion.

TABLE OF CONTENTS

LIST OF TABLES

TABLE

LIST OF FIGURES

FIGURE

FIGURE

CHAPTER I

INTRODUCTION TO STUDY

PROBLEM

Chemistry is Hard

Chemistry is a historically difficult subject. General chemistry courses frequently have high D, F, and withdrawal rates and low student retention.[1, 2] This is an inherent problem for both students and departments, since many degree programs require successful completion of general chemistry at an early stage in the program.[1] Many studies have focused on course-specific characteristics, including mathematical demands, vocabulary, levels of representation, levels of abstraction, course content, and class size. Other studies have looked at student-specific traits and skills as predictors of success in general chemistry. These include mathematical skills,[3, 4] verbal and language skills,[3] formal reasoning ability,[5, 6] affective traits,[7-9] personality traits,[10] chemistry content knowledge,[2] chemistry conceptual knowledge,[4] among many others. Many studies have focused on several of these attributes, comparing predictive capabilities[11] or creating complex predictive mathematical algorithms.[1, 12, 13] Conceptual understanding is the culmination of overcoming many of these challenges. The scope of the research in this study was a small part of a large body of work. Specifically, this study focused on conceptual understanding in general chemistry. The purpose of this introduction is to

provide context and background for student conceptual understanding research in general chemistry, state research goals, and explain why this research is not only important but also why it is needed by the chemistry education community.

Course-Specific Challenges

Mathematical demands. Most general chemistry courses have some type of mathematical pre-requisite, either recommended or enforced by the department.[11] This is because many chemical problems involve basic algebraic manipulations, use of exponents, calculations using exponential and logarithmic functions, linear regressions and associated equations, or use of percentages and evaluation of chemical formulas.[1] Thus, mathematical competency is required for chemistry problem-solving success. Chemistry teachers do not have time to go over these critical mathematical skills in class, so the responsibility of having or obtaining these skills lies with students.

Chemistry vocabulary. Chemistry involves the use a very specific, technical vocabulary, as illustrated by the large glossary of any general chemistry textbook. Communicating chemistry concepts involves the heavy use of chemistry-specific vocabulary, as well as terms that have wide-spread colloquial use in everyday language.[14] For chemistry-specific terms, students need to conceptualize a term and what it means in a chemistry context. An example would be the term molarity. A definition for molarity is "moles of solute per volume of solution in liters".[15] A student must previously know the terms, moles, liters, solvent and solute, to make sense of this definition. In addition, this definition provides a good example of the complexity of chemistry definitions. Chemistry vocabulary also uses terms shared with everyday vernacular, but sometimes has a completely different definition.[14] For example, the term heat is treated most often as a

noun in common everyday usage, but is used to describe a process in chemistry. Even within chemistry, words such as weak and strong are used in different contexts with different meanings. A strong bond is used to describe a bond that requires a significant physical force to disrupt it. However, a strong acid dissociates readily in water.[16] Chemistry word problems that use technical terms need to be translated such that important information can be identified and the correct problem-solving strategy can be implemented.[14]

Levels of representation. The words that chemistry instructors use are not the only way they communicate with students. Chemistry content is often presented in three forms of representation: macroscopic, sub-microscopic, and symbolic.[17] Macroscopic representations are at the time and scale of everyday life and include classroom demonstrations (physical observation). The strength of macroscopic representations is that they are concrete; however, they are not good for explaining chemical phenomena. Sub-microscopic (or particulate) representations depict phenomena or objects that cannot physically be observed, such as atomic motion. Though sub-microscopic representations are often used to explain macroscopic observations, the concepts and information are often abstract and difficult for students to conceptualize.[18] Symbolic representation encompass chemical symbols; formulas; and structural and graphical representations of chemical species, reactions, and phenomena. Often during a typical chemistry lecture, all three levels of representation will be used. Sometimes multiple levels of representation (e.g., a chemical formula and a graph) are used at the same time.[19] New to chemistry instruction is the use of computer-simulated representations that can have dynamic interfaces or show dynamic motion that cannot be represented in pictures.[20] Thus,

information presented in general chemistry classrooms can be in multiple levels of representation, sometimes including multiple forms of representation for the same process.

Levels of abstraction: Concrete versus abstract variables. Chemistry has many variables that can be experienced by students in the real world, such as time, force, temperature, mass, and length. All of these are examples of concrete variables that can be measured by students. Other variables in chemistry are less concrete and more abstract, such as velocity, which is calculated by measuring the distance covered over a certain amount of time. Students cannot measure velocity directly, but they can calculate it from concrete variables (time and length). Many variables in various topics of chemistry are calculated from variables that cannot be directly measured.[21] Moreover, many topics in general chemistry cover phenomena and processes that are not easy to relate to personal experiences (e.g., atomic motion), based on size, speed, scale, or scope.[14, 20-23]

Depth and breadth of content covered. The amount of content that is taught in most general chemistry classrooms is significant and has increased over the past 50 years.[24-26] Over 20 years ago, upon analyzing the last century of how chemistry is taught, Spencer concluded, "the world has changed but general chemistry has not, except to add more topics".[25] To illustrate this point, simply look at the content coverage for the summative general chemistry American Chemical Society (ACS) exam.[27] This large quantity of content demonstrates the breadth of content typically addressed by chemistry instructors. Many individuals[25, 26, 28, 29] and an officially sanctioned committee[30] have called for a reduction of course content to focus on depth and providing more opportunity to focus on developing cognitive skills and discussing application-based topics. Specifi-

cally, calls for removing content only pertinent for chemistry majors have been made,[29, 30] since most students taking general chemistry are not chemistry majors.[28] However, currently no major professional chemistry organization has provided official guidelines for changes in general chemistry curriculum content, so instructors are challenged with sacrificing depth of coverage for breadth of coverage.[28]

Pedagogical dogmas: Educational inertia. With all of the calls for curricular and instructional changes for the general chemistry series in the last 20 years, why haven't more instructors taken action in their own classrooms? Two main factors affect instructors' implementation of new instructional techniques and curriculum: extrinsic and intrinsic factors.

Extrinsic factors that general chemistry instructors face include physical barriers, such as classroom design and class size and departmental and institutional barriers. A physical barrier to instructional change can limit what changes are practical or possible. If a class is taught in a classroom that does not support use of a technological teaching techniques (e.g., projectors, clickers, etc.), then these techniques cannot be utilized.[31] This is especially challenging with large class sizes, where a room change is not an option. Class size also affects what kind of feedback an instructor can give for student-generated work resulting from a teaching innovation (e.g., concept maps).[32] Institutional and departmental barriers include lack of incentives (e.g., promotion, recognition, financial),[33, 34] lack of support (e.g., technology, teaching assistants, etc.),[31, 35] and structure of course teaching responsibilities (e.g., team-taught courses).[36] Studies have shown that though most faculty members value both teaching and research, they believe

that only their research is rewarded by promotion and prioritize their efforts accordingly.[33, 34]

Intrinsic factors that general chemistry instructors face include affective barriers and resource barriers. Instructor affective barriers include lack of motivation[37] and fear of poor course evaluations.[33, 34] Resource barriers include lack of knowledge and lack of time. Not all chemistry instructors are familiar with the wealth of research literature on effective instructional strategies and interventions or have the support to implement these techniques in their courses.[26, 31, 35-37] Even with the motivation, knowledge, and ability to try new instructional strategies, instructors may be limited with how much time they can invest. This is especially true for tenure-track faculty members, who do not see this extra preparation time as a good investment toward promotion.[31, 33, 34]

In general, a cost/reward argument is often used by faculty to evaluate the time and effort of implementing new instructional techniques. How much effort is required (cost) and what tangible benefits will result (reward), and what probability that the new teaching technique will be better than the current technique? Studies have found that unless the benefits were intrinsically-based (e.g., instructor satisfaction of a job well done), most faculty will not risk the threat of poor course evaluations and use of valuable research time to implement new instructional techniques.[34] So, even as the chemical education research community provides more instructional tools and teaching strategies for general chemistry instructors, many will continue teaching the way they were taught.[28]

Course-specific challenges summary. General chemistry requires many skills outside of understanding factual knowledge. There are several course-specific challenges that students face, which have been discussed above. These include mathematical

demands, copious technical and sometimes confusing vocabulary, multiple levels of representation and abstraction, large breadth of course content, and many factors that discourage the implementation of instructional techniques that could help students with these demands. Thus, success in general chemistry requires both content knowledge and other skills to aid in understanding and applying this knowledge.

Student-Specific Challenge

Variability of student populations. Students enrolling in general chemistry courses have a wide range of personal experiences and prior chemistry content knowledge, vary in age, and come from a large array of degree programs.[1] This heterogeneous student population provides unique challenges for instructors and students. The following highlight student-specific challenges associated with chemistry. Some of these challenges directly relate to course-specific challenges discussed above (e.g., mathematical, vocabulary, levels of representation).

Affective domain. A student's attitude in a general chemistry course can be another predictor of success in chemistry.[6, 8-10, 38, 39] The term, attitude, includes the mental constructs of beliefs, interests, values, self-concept, self-efficacy, self-esteem, motivation, and anxiety.[38] Most studies focus on one or two of these constructs in relation to an instructional intervention or academic achievement. Both intellectual accessibility and emotional satisfaction correlated moderately with achievement on the summative ACS exam, and both increased the predictive power of mathematical Scholastic Assessment Test (SAT) score and achievement on the ACS exam.[7] Studies of Turkish high school students suggest a link between student motivation and anxiety with course grades,[9] but this has not been studied with a college student population in the United States. The

degree to which the results of this study reflect trends in United States college chemistry classrooms still needs to be determined and is outside the scope of this work.

Student self-concept, "the evaluation one makes regarding one's own ability and performance in a subject area",[8] has also been linked with success in general chemistry. Specifically, students who were classified as having a low self-concept had predictably lower ACS test scores, even when the researchers controlled for mathematics SAT scores.[8] Self-concept might be able to be increased throughout the course of the semester using student-focused teaching strategies (e.g., process oriented guided inquiry learning [POGIL]).[8]

Lastly, student's Jungian personality type has also shown to correlate with general chemistry achievement.[10] Personality types have four main categories: Attitude (Extra-verted or Introverted), Perception (Sensing or iNtuitive), Judging (Thinking or Feeling), and Dominant (Judging or Perceiving). Students who were characterized as I-N-T-J personality typed were most likely to be found in the top 10% of course grades, while students with E-S-F-P were most likely to be found in the bottom 10% of course grades. Interestingly, most professors were *- *-T-J but not E- *- *-P personality types. The traits most often associated with people with I-N-T-J personality type include comfortable with studying alone, able to handle abstract concepts, and prefer conclusions reached using mathematical and logic.

Levels of representation and student learning. All levels of representation in chemistry aim to help illustrate and explain abstract and complex content in meaningful and easier–to–understand formats. However, no level of representation is without problems. Macroscopic representations provide no explanation for chemical phenomena,

so students might misinterpret the "why" and "how" of their observations. Sub-microscopic representations are the most abstract and most difficult for students to relate to concrete observations. The translation of knowledge gained from macroscopic representations to sub-microscopic representations is not always easy. Often students have difficulty ascertaining the dynamic nature of particulate motion with pictorial sub-microscopic representations. Symbolic representation can often be interpreted incorrectly by students and lead to conceptual misunderstandings. In addition, the use of multiple levels of representation at the same time does not always benefit students of all abilities. Research has shown[19] that students of low ability benefit from instruction that focuses on symbolic representation, which helps with conceptual understanding. However, students of higher ability benefitted most from sub-microscopic representations, as they had already obtained the sufficient prior knowledge to understand these representations.

Prior knowledge. Student prior knowledge is just as important as cognitive skills for success in general chemistry.[2, 4] While many studies discussed so far have demonstrated that chemistry content knowledge is not the only factor in achieving success in general chemistry, it is still a major component. Given that not all institutions enforce high school chemistry (or equivalent preparatory course) prerequisites, and given that all students who meet the prerequisites do not have the same level of prior knowledge, knowing what students know can be quite challenging.[11-13, 43] The influential constructivist David Ausubel famously stated, "The most important single factor influencing learning is what the learner already knows. Ascertain this and teach him accordingly".[44]

Mathematical ability. Mathematical literacy is critical to success in general chemistry and can vary greatly among general chemistry students. Most chemistry

problems involve some type of mathematical interpretation or calculation. Conceptual understanding alone will not guarantee success, as many studies have shown.[1, 2, 4, 11, 43, 45-47] Studies that aim to measure students mathematical ability have either used strictly math-based assessments with no chemistry content, which include mathematics SAT scores,[6, 48-50] and a diagnostic algebra test.[1] Many other studies have used placement exams that include both items that test mathematical competency and items that test chemical knowledge. These include the Toledo Chemistry Placement Exam,[45] California Placement Exam,[4] Fullerton Test,[2] and the Student Pre-Semester Assessment.[43]

A common result from all of these studies was that mathematical proficiency, as determined by any of the above measures, improved the ability to predict student success in general chemistry. In one example, low mathematics SAT scores strongly correlated with low first semester general chemistry success rates, but high mathematics SAT scores did not correlate as strongly with high first semester general chemistry success rates.[48, 49] In general, the conclusion from all of these studies is that mathematical proficiency is not the only predictor of success in general chemistry, but is a good predictor of failure in general chemistry.

Chemistry language: Vocabulary, context, and syntax. The language of chemistry poses several different challenges for students. First, students need to be fluent in chemistry-specific vocabulary to understand concepts that use these terms. If a student does not understand what the term, mole, means in a chemistry context, she cannot understand the definition of molarity.[14, 51] Alternatively, many terms used in general chemistry are familiar to students outside of chemistry, but have slightly or completely different meaning in chemistry (e.g., equilibrium).[51] Thus, students' prior knowledge of

colloquial terms used in chemistry are sometimes at odds with chemistry-specific definitions and lead to cognitive conflict.[16, 51, 52] In addition, terms used within a chemistry context have different meanings, such as strong, weak, powerful,[16] or volatile[53] or pure.[54] Furthermore, when English is not the primary language for students, certain words caused problems for students, such as disperse, composition, consistent, isolate, proportion, efficient, and reference.[54] Research shows that students can lose up to 25% of their available working memory space when processing information not in their primary language.

Lastly, the syntax used in word and multiple-choice chemistry problems can have an effect on students. A question stem that uses complex syntax is more difficult than one that has a simple structure.[55] Specifically, the number of words has less effect on difficulty than grammatical structure of a sentence.[52, 55] If unfamiliar words are used in a sentence with a complex grammatical structure, students will have a challenging time processing the information within the problem. These problems are exacerbated when English is not a student's primary language.

Information processing ability. The theory of information processing was developed from the awareness that chemistry is difficult for many students, and this perceived difficulty had a strong negative influence of students' affective dispositions.[51] Information processing theory pools the chemistry-related difficulties under one theory and focuses broadly on two main student-specific mental constructs: information stored in long-term memory (e.g., prior knowledge) and available space for information to be stored and processed in short-term memory, or working memory space.[17] The flow of information is shown in Figure 1. A critical component of this theory is that working

memory space is limited, and when overloaded will severely impede student learning. Current theory speculates that students can only hold 7 ± 2 pieces of information in their working memory at one time and only process 5 ± 2 pieces of information.[56] Thus, students could be expected to remember and repeat 7 ± 2 random digits, but if asked to repeat these digits in reverse order (e.g., process this information), could only repeat 5 ± 2 digits. Overloading the working memory space is commonly known as cognitive overload, which is known to occur readily for many general chemistry students.[51, 54, 57-59]

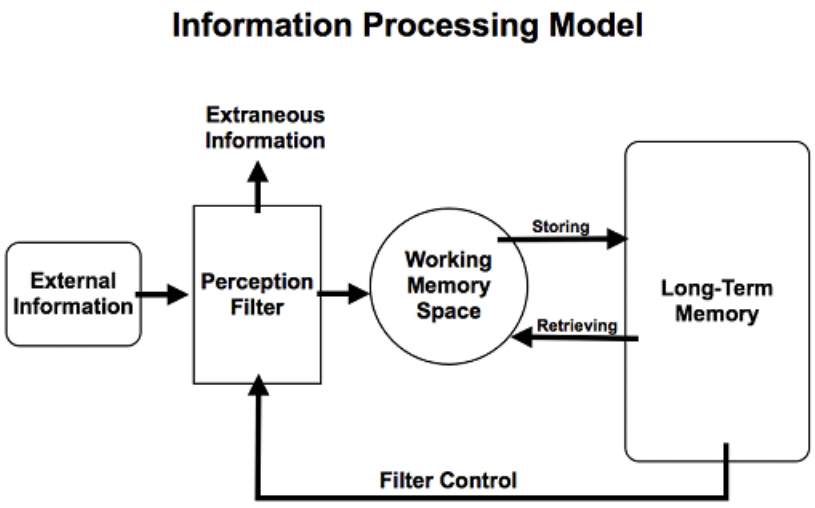


Figure 1. Johnstone's information processing model.[51]

Prior knowledge plays a critical role in reducing cognitive overload. Learners can filter out extraneous information using their prior knowledge, including cognitive skills such as pattern recognition. This model also helps explain why chemistry problems with chemistry-specific vocabulary and complex syntax quickly lead to cognitive overload. If students are unfamiliar with terms in the question, they are not filtered and go straight to

the working memory, even if the terms are not critical to solving the problem.[17] Additionally, if the students have to process complex question syntax or if English is not their primary language, the available space in the working memory decreases to 5 ± 2 pieces of information.[52, 54] This model is an explanation of the Ausubel quote given above: student prior knowledge is a limiting factor for the ability to process new information. If instruction continually overloads student working memory, then no meaningful learning can occur.[44, 60]

Problem solving ability. The information processing model can be extended to understand the difficulties students face in solving chemistry problems. How prior knowledge is organized will affect how easily it can be recalled for processing information in the working memory space. Examining differences between experts and novices, specifically with regard to prior knowledge organization and problem solving, can be very illustrative. Experts tend to organize knowledge by grouping similar information into chunks that can help with faster recall.[56,61] This chunking of knowledge demonstrates deep understanding of concepts by experts. In contrast, novice learners do not have as complex and meaningful knowledge structures, so chunking is primarily based on superficial characteristics of concepts. Likewise, experts can flexibly and quickly retrieve chunked knowledge, where novice learners often take more time and struggle to retrieve chunked knowledge. One reason for this difference is that experts can process information with greater ease because of an ability to filter out extraneous information and focus on the important aspects of new information. Novices do not have the same deep content knowledge base and struggle to filter new information and can easily overload their working memory.

Therefore, problem solving ability can be dramatically decreased with cognitive overload, which can be influenced by a lack of prior knowledge. Prior knowledge can help filter extraneous information in word problems as well as provide cognitive strategies to process information in working memory space. The more this information is chunked and organized in meaningful ways, the easier this information can be recalled and used for problem solving.

Another important student-specific factor in problem solving ability is student metacognition. Metacognition includes the ability to reflect upon one's thinking, ability to regulate this thinking, and actively reflect and regulate cognitive processes based on knowledge of these processes.[62, 63] Metacognition has been shown to be important in student ability to create successful problem-solving strategies, which can lead to successful problem solving.[62-64] Student confidence in answers to multiple-choice questions is a measure of the reflective aspects of metacognition. High-performing students' confidence is in almost complete agreement with their answers, but low-performing students are over-confident and overestimate their ability to answer questions correctly.[63] This reflects the need to provide feedback to students, to help develop accurate metacognitive skills, especially for students who perform poorly on classroom assessments.

Conceptual understanding. The basis of both Johnstone's information processing model[17] and Ausubel's work on student cognition stem from a constructivist theoretical perspective that began with the seminal work of Piaget.[65] George Bodner eloquently states the tenets of constructivism in the following excerpt: "Knowledge is constructed in the mind of the learner.[66] This theoretical framework assumes that we don't discover knowledge; we actively construct it. We invent concepts and models to make sense of our

experiences. We then continually test and modify these constructions in the light of new experiences."[67]

The assumption that knowledge is constructed by learners, from both their formal and informal experiences, is critical in trying to explain conceptual understanding. Construction of knowledge also implies that the connection between chunks of knowledge are made in meaningful, non-arbitrary ways.[44] Thus, for meaningful learning to occur, students must have some prior knowledge to help contextualize and anchor new knowledge in ways that make sense of both prior knowledge and new knowledge.[44, 68] What defines conceptual understanding is a difficult question to answer, as most science education researchers would argue because it is multifaceted. Some of these facets include: (1) mindful memorization of conceptual knowledge, (2) integrating knowledge of multiple concepts to construct a knowledge framework, (3) being able to transfer and apply conceptual knowledge to solve novel problems, and (4) being able to think globally about a concept in the context of a larger system.[69] Most of these facets involve higher-order cognitive processes. To be able to mindfully memorize conceptual knowledge, students need to understand the concept well enough to contextualize it with other concepts in a way that is both accurate and meaningful. This is the beginning of creating a knowledge structure, which can then be used for transfer tasks with novel problems. Because many students struggle with the mindful memorization of conceptual knowledge, all higher-order cognitive processes become that much more difficult.

However, breaking down conceptual understanding into four stages is still an over simplification of a very complex process. Arguably, this is an improvement from the dichotomous, you understand it or you don't understand it classification, but does not

reflect the continual learning processes of constructing and reconstructing knowledge from experiences. A more accurate description of conceptual understanding is a continuum. For simplicity, we can assume that this continuum is finite and has two opposing ends. On one end of this continuum lies how an expert in chemistry would understand a given concept. This would include all four milestones of conceptual understanding mentioned above, but not put emphasis on the distance between these steps on this continuum. On the opposite end of the conceptual understanding continuum is alternative conceptions, which is the most extreme example of a conceptual misunderstanding. A representation of this continuum is shown in Figure 2.

Conceptual misunderstanding. If we assume that students actively construct knowledge from their experiences, this construction can either be in agreement with what an expert believes to be true (conceptual understanding) or will be is someway in disagreement (conceptual misunderstanding). The inclusive term, conceptual misunderstanding, will be used to describe all forms of student knowledge that vary from those that are accepted by the scientific community. Alternative conceptions are conceptions that vary from what is accepted by the scientific community. What makes alternative conceptions unique is that students have taken the steps to contextualize and chunk knowledge into existing knowledge structures, but this knowledge is in disagreement with content experts' conceptual understanding. Students with alternative conceptions believe this knowledge is true and integrate these conceptions into their knowledge structure, chunking with other knowledge. The consequences being, once integrated, alternative conceptions can be very difficult to replace. However, not all student conceptual misunderstandings are as ingrained or established as alternative conceptions. As seen in Figure

2, conceptual misunderstandings include mismemorization, incomplete conceptions, incomplete knowledge integration, and incorrect knowledge integration. This, again, is an oversimplification of a very complex process, but can be used as a device to help understand what exactly conceptual misunderstandings entail.

The study of conceptual misunderstandings in science has been a focus of the science education research community for the past 75 years.[71] Much of this research has focused on the identification of specific conceptual misunderstandings students have in a particular content area, including many in the field of chemistry.[42, 71-95] This body of research has demonstrated that many student conceptual misunderstandings are pervasive, resistant to instruction, and in some cases, may reduce or block the adoption and utilization of accepted expert conceptions and limit the effectiveness of instruction. Additionally, conceptual misunderstandings found in student populations have also been identified as being used by pre-service secondary instructors,[72, 90] secondary instructors,[79] and even post-secondary instructors.[86]

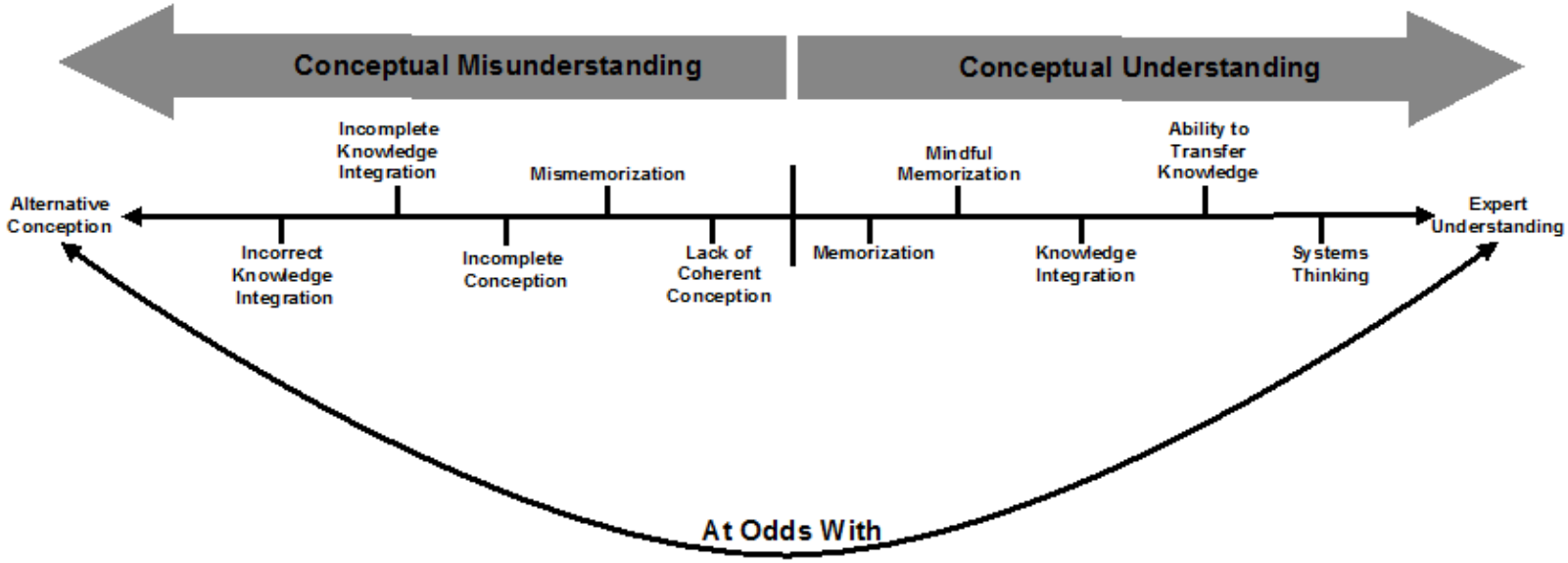


Figure 2. Conceptual understanding and misunderstanding. It is represented as a continuum with key facets represented on this continuum. Spacing of key facets on continuum is arbitrary. Possible connections between facets are not shown for simplicity.

Challenges of conceptual change. A logical question is why are some conceptual misunderstandings resistant to change, even with targeted instruction? Conceptual change is difficult because it is not a one-step process. If this were true, students would simply replace their conceptual misunderstandings with the correct conceptions when taught by instructors. However, the fact that many students complete general chemistry, and sometimes a degree in chemistry, and still hold and use certain conceptual misunderstandings is an argument against a one-step process. Instead, conceptual change has been described by many as a multi-step process. Posner[96] argued that for conceptual change to occur, students must first be dissatisfied with their existing conception, understand the new correct conception, associate meaning and context of the new conception in their existing knowledge framework, and believe that adoption of the new concept will be beneficial.

Dissatisfaction with the existing conceptual misunderstanding is crucial for the process of conceptual change to begin. The most common way this can happen is through an anomaly.[96] An anomaly is an event where new information cannot be assimilated within a student's current knowledge structure. Or put more plainly, the new information does not make any sense.[96] When students try to use their existing conception to explain a phenomena or assimilate new information and find that there is disagreement or incongruity, a decision needs to be made. Either the current conception needs to be modified or changed, or the use or phenomena trying to be explained is anomalous. Students cannot easily modify or change their conception unless they know what part of their existing conception needs to be modified or have a new conception ready to replace their existing conception. This is not as simple as it sounds. Modification of the existing conception can

be challenging because students need to know why their current conception is not viable and know what modifications need to be made to make the conception viable. Alternatively, replacing the current conception is not easy because students generally do not have a back-up conception ready in case of an anomalous event. In addition, the new conception needs to be intelligible and be seen as a viable replacement for the current conception. Often, the new (correct) conception may be more difficult to understand. If students do see the need to replace their current conception with a new intelligible conception, they still need to accommodate this conception into their existing knowledge framework. This is hard. Accommodation of the new conception means rebuilding connections within an existing knowledge structure, which can take significant mental effort. This is why the last step of conceptual change is important. Accommodation will only occur if the student believes that the new conception will be beneficial. Otherwise, the effort for accommodation will not be justified and the current conception will be kept.

An illustration as to how difficult conceptual change can be for students has been shown by using anomalous data. A study[97] found that college students had eight possible responses to anomalous data: (1) ignoring the data, (2) rejecting the data, (3) professing uncertainty about the validity of the data, (4) excluding the data from the domain of the current theory, (5) acknowledging the data do not agree with the theory but hoping new information will surface to explain this discrepancy (6) reinterpreting the data, (7) accepting the data and making peripheral changes to the current theory, and (8) accepting the data and changing the theory. Only the last response will allow for conceptual change to occur, while the other seven responses avoid cognitive conflict and uphold the use of a conceptual misunderstanding.

Student-Specific Challenges Summary

This section illustrates how diverse the student population taking general chemistry can be and how these differences can affect learning and success in chemistry. Many students do not have the required skill to be successful in general chemistry. These skills can be both cognitive and affective. Some of these skills can be fostered or taught (e.g., affective domain skills, metacognition, problem solving strategy, knowledge chunking), while other are much more difficult for general chemistry instructors to address (e.g., mathematical ability, prior knowledge, working memory space). Conceptual understanding is a very complex process, and many outcomes can result from shared experiences such as instruction. Such outcomes include conceptual understanding as well as conceptual misunderstanding. Conceptual misunderstandings can be resistant to instruction and can be detrimental for student learning. However, changing students' conceptions, especially those embedded in student's knowledge framework, can be very difficult from both the perspective of the instructor and the student.

PROBING STUDENT KNOWLEDGE

Many Tools for Many Problems

The course-specific and student-specific challenge addressed above have been the focus of many studies, demonstrating that the chemical education research community is both aware of and actively addressing these challenges. The pedagogical toolbox available for instructors has never been bigger and continues to grow each year.[95] Some of these tools are diagnostic instruments, designed to provide feedback to both instructors and students. This information can be used to help instructors make informed decisions about instructional practices and design, specifically with regard to student-specific

challenges that can vary from one student population to another. This information can be used by students, both before and during the learning process, by providing feedback regarding deficiencies in certain cognitive and affective skills, as well as conceptual understanding, which can be developed throughout the course of the semester.

A concerted effort has been made in the last 10 years to create diagnostic instruments for conceptual misunderstandings.[95] Given the difficulties associated with conceptual change, diagnostic instruments that can help both students and instructors identify and address conceptual misunderstandings are critical. These diagnostic instruments are commonly called concept inventories, as their goal is to identify what conceptions students are using in a course.

Concept Inventories

Concept inventories were popularized over 20 years ago when the Force Concept Inventory was published for use in physics classrooms.[98] A concept inventory question (or item) is very similar to a concept question found in traditional quizzes and tests as a way to measure conceptual understandings. What makes concept inventories unique is the use of identified student conceptual misunderstandings as distracters in multiple-choice items. In this way, a concept inventory provides information about what, if any, conceptual misunderstandings a student is using based on what distracter (incorrect answer) they choose. Over the last two decades, many concept inventories have been developed[95] using an assortment of developmental methods.[99, 100] A reason for this growth in the area of concept inventory development is a call by the National Research Council (NRC), Committee on the Foundations of Assessment, for more assessments that measure higher cognitive processes in easy-to-administer formats [pg 3].[64] Assessment of higher cogni-

tive processes is distinct from simply testing competency or ability to memorize discrete chunks of knowledge without context.

Concept inventories can also be utilized as an instructional tool when used as a formative assessment instrument. Formative assessment is meant to provide feedback to both instructors and students throughout the learning process. This is in contrast to summative assessments that are generally cumulative and used for student evaluation at the end of a course. Giving a concept inventory directly after instruction can inform the instructor of persistent conceptual misunderstandings that are still being used by students and provide students with direct feedback with what concepts they have yet to fully understand. This can be a very efficient and effective way to provide feedback, which can be obtained multiple times through the learning process. This is critical for addressing conceptual misunderstandings that have been shown to be resistant to instruction in student populations.

Concept inventories have been created and used in many science, technology, engineering and mathematical (STEM) disciplines.[99-101] Given all of the difficulties facing students in a general chemistry course, concept inventories can be a practical and effective tool for instructors to identify and confront student conceptual misunderstandings. Potential targets for concept inventory development include topics in general chemistry, where students historically are known to hold many conceptual misunderstandings and in topics that have concepts important for subsequent chemistry courses. Thermochemistry is taught in first-semester college-level general chemistry and has many topics that are the building blocks for understanding concepts in thermodynamics, which is taught in second-semester general chemistry and physical chemistry. Student concep-

tual misunderstandings in thermochemistry are well documented,[73-75, 78-80, 82-87, 90, 93-95, 102-105] yet no chemistry-specific thermochemistry concept inventory has been developed.

Probing Student Knowledge Summary

Chemistry courses and associated content can challenge different students for different reasons. Students' wide range of cognitive skills and affective traits both help determine how well students overcome these course-specific challenges. However, instructors and students can benefit from student-specific information obtained from diagnostic instruments assessing both cognitive skills and affective traits. Diagnostics assessing basic cognitive skills (e.g., mathematical ability, working memory space) and affective traits (e.g., motivation, self-concept, personality type) are good assessments at the beginning of courses.[64] These assessments give instructors baseline information of the student population taking a course and give students feedback on their preparedness coming into the course. On the other hand, diagnostics focusing on cognitive processes (e.g., conceptual understanding) are useful during the learning process and at the end of instruction.[64] These assessments are more content-specific and relate to expectations of what students should be able to do or know at the end of instruction. Conceptual understanding is a cognitive process, which is multifaceted and encompasses complex and dynamic relationships that can be very difficult to measure. The study of conceptual misunderstandings, specifically using diagnostic assessments such as concept inventories, can be used as formative assessment and as a tool to evaluate new instructional techniques aimed to target specific conceptual misunderstandings.

PURPOSE STATEMENT

The purpose of this research was to create an assessment instrument to help identify student conceptual misunderstandings pertaining to the topic of thermo-chemistry taught in college-level general chemistry. This assessment is referred to as the Thermochemistry Concept Inventory (TCI). Obtaining evidence for the reliability and validity of the use and interpretation of testing data generated by the TCI for the target population is a crucial aspect of the evaluation of the TCI's utility.

RESEARCH QUESTIONS

- Q1 What conceptual misunderstandings in thermochemistry are students using in college-level general chemistry classrooms?
  - Q1a What thermochemistry topics are taught in most general chemistry classrooms?
  - Q1b Of these topics, which are classified as most important by practicing general chemistry instructors?
  - Q1c What conceptual misunderstandings (from the important topics) are being used by students?
- Q2 What development methodology is necessary to create the Thermochemistry Concept Inventory?
  - Q2a How many items are needed to cover the most important conceptual misunderstandings?
  - Q2b What is the best format for these items?
  - Q2c What are the most important criteria to evaluate these items in pilot tests?
- Q3 How do items in the Thermochemistry Concept Inventory perform in pilot studies?
  - Q3a How will the performance of items be measured?
  - Q3b With what student population will item performance be evaluated?

Q3c How will these performance measurements be evaluated?

Q3d What changes to items will be implemented to improve performance?

Q3e How do these changes affect item performance?

Q4 What are the intended uses and interpretations to be made using data collected from the Thermochemistry Concept Inventory, and what evidence is there for the validity and reliability of these uses and interpretations?

Q4a What types of validity need to be established for the uses and interpretations of testing data collected using the Thermochemistry Concept Inventory?

Q4b How will these types of validity be established and with what evidence?

Q4c What threats to validity and reliability are expected?

Q4d How will these threats be mitigated or addressed?

RATIONALE

As evidence that chemistry is a difficult subject, go to any social event (e.g., party, barbeque, etc.) and proclaim that you are a chemist. General reactions from strangers may involve increased physical distance, followed by a slightly awkward moment of silence, and then an exclamation of how they (the newly befriended stranger) “were never any good at chemistry”.[106] The STEM disciplines have been losing potential college majors, in part, because of a lack of instruction to both facilitate meaningful learning and to motivate students.[107] Creating positive, meaningful learning environments in chemistry classrooms can promote learning as well as help retain students within the sciences and possibly within chemistry. Sometimes for instructors, what is known about students can be just as important as content knowledge.[107] This knowledge of students can be obtained

through frequent feedback from students, such as formative assessments or student-generated responses. Students should leave chemistry courses with not only conceptual understandings and chemical literacy, but also with a positive experience interacting with the material. These goals can be promoted and cultivated through thoughtful, purposefully-designed and evidence-based instructional practices.[103, 107]

Evidence-Based Instructional Practices: Where is the Evidence?

Calls for general chemistry instruction to be more evidence-based have persisted for the last 25 years.[26, 28, 32, 51, 103, 107, 108] Yet many instructors do not adhere to repeated calls by NRC reports,[61, 64, 103, 107] commissioned panels,[26, 95] and concerned chemical education researchers[28, 32, 53, 67, 108] to change instructional practices in light of what we know about students and student learning. Why is this? The forces that resist change in instructional techniques discussed above (e.g., educational inertia) are sure to contribute. However, one problem still remains for many general chemistry instructors: lack of evidence. Many instructors know the challenges facing students as discussed in the previous section. In addition, many instructors will design new instructional techniques to address these challenges and concerns. But collecting evidence as to whether or not new instructional techniques are better than current practices (or worse) can be very difficult.

There are challenges associated with creating an experimental design to evaluate a new technique; specifically, having to teach for a control group (using an existing method) in addition to teaching with a new instructional technique. Even more challenging is having an instrument to accurately measure changes between the control group and the treatment group. The instrument needs to be designed specifically to measure changes

in a student-specific attribute (e.g., affective trait, cognitive skill, etc.) that the new instructional technique is trying to target. Using course grades to measure the effectiveness of instructional interventions lacks resolution and has inherent lack-of-generalizability issues associated with how course grades are determined.[6] In addition, validity and reliability for the intended uses and interpretations of instrument-generated data need to be established for the instrument's initial development and each time it is used for a new target student population. Many diagnostic instruments, including concept inventories, have been specifically developed for just this purpose.[95, 100, 109] However, the methods used to develop these instruments vary considerably, sometimes with little evidence for validity or reliability reported.[100, 109]

Thus, for evidence-based instruction to take place, instruments that have obtained validity and reliability evidence for collecting data to measure the effectiveness of new instructional techniques are needed. This evidence is crucial to promoting new effective instruction techniques and discouraging the use of those that are less effective than traditional teaching methods.

Evidence. Traditional testing mainly focuses on factual and procedural knowledge.[64, 110] However, many instructional interventions target student-specific skills and traits and do not target recall of factual and procedural knowledge. Thus, student performances on traditional classroom evaluations (e.g., quizzes and tests) are not a good source of evidence for the effectiveness of instructional interventions. An illustration of how concept inventories can be used to collect evidence in evaluating new instructional techniques can be found in physics. The Force Concept Inventory (FCI)[98] has been used as a source of evidence for student conceptual understanding in physics.[111] The FCI was

developed from identified conceptual misunderstandings related to the concept of force used by physics students. The FCI found widespread use among physics educators in evaluating student conceptual understanding. In one study, evidence was needed to evaluate the use of interactive-engagement methods that specifically aimed to develop student conceptual understanding in physics, so the FCI was used.[111] Data collected from student responses to the FCI questions provided evidence for the effectiveness of the interactive-engagement instructional method to promote student conceptual understanding.[111] This is just one example of how a concept inventory can be used as evidence for evaluating a new instructional technique.

Need for a Thermochemistry Concept Inventory. If recommendations by the NRC that instruction and assessment need to focus more on cognition and less on factual and procedural content are to be implemented by instructors, instruments that measure student cognition will be needed[64] to both support learning (formative assessment) and evaluate new teaching methods. Concept inventories can be very useful and efficient instruments to measure student conceptual understanding (and misunderstanding). This is especially true in general chemistry classrooms, where student enrollment can be large and there is a need for assessments that minimizes data collection and analysis time and maximizes the information collected by the assessment. Creating a cumulative concepts inventory for one or both semesters of general chemistry is difficult, because the number of concepts taught in general chemistry is large. Furthermore, the required length for such a concepts inventory would be impractical in most instructional situations. In contrast, concept inventories focusing on a singular topic or several topics taught in a general chemistry course could have fewer items and require less administration time. In addition,

concept inventories focusing only on a single topic are ideal for formative assessment, given the targeted content and shorter administration time. Deciding which topics in general chemistry should be targeted for development of a concept inventory useful to the chemistry education community requires certain selection criteria. The topic should be taught in most general chemistry courses, central to future chemistry conceptual understanding, and have known associated student conceptual misunderstandings. These three criteria help ensure that the concept inventory will be relevant in most chemistry classrooms, focuses on important general chemistry concepts, and has enough associated student conceptual misunderstandings to warrant the creation of a concept inventory.

Thermochemistry is a historically challenging topic for students,[18, 21,57] and is taught in most general chemistry courses. Thermochemistry topics are important in second-semester general chemistry, along with many other upper-division chemistry courses. Thermochemistry also has many concepts that are abstract in nature, as seen in Figure 3. Abstract variables are more difficult to conceptualize, especially if students do not have a solid conceptual understanding of variables of lower abstraction. For students to understand the concept of enthalpy, they need to understand every concept of lower abstraction below enthalpy, as shown in Figure 3.

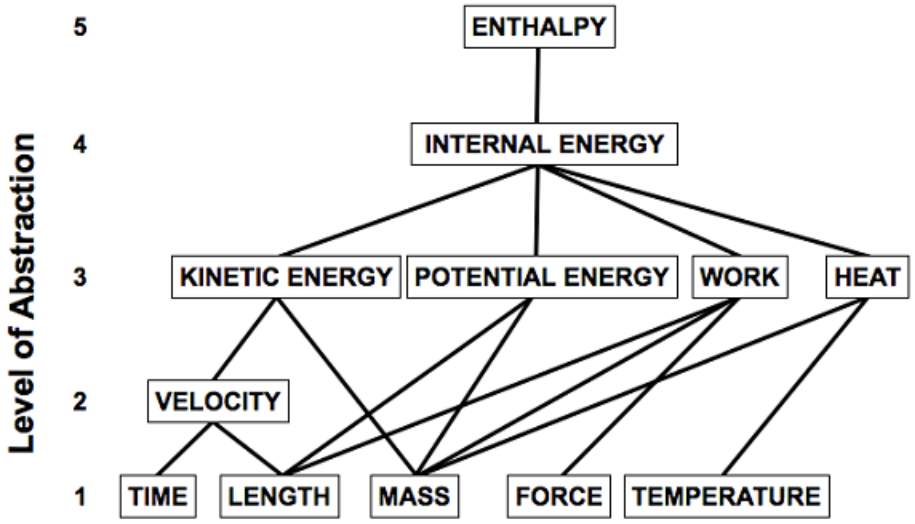


Figure 3. Levels of abstraction related to thermochemical concepts. Adapted from Dixon, J.; Emery, A., Jr. Semantics, Operationalism, and the Molecular–Statistical Model in Thermodynamics. *Am. Sci.* **1965,** *53,* 428–436.

In addition, if students try to understand a concept of high abstraction but lack understanding of concepts of lower-order abstraction, they might be forced to memorize factual knowledge. The difference being that having a conceptual understanding implies that some contextualization and integration into existing knowledge occurs. This process also can screen new knowledge for inconsistencies with prior knowledge and prior knowledge structures, helping to prevent adoption of conceptual misunderstandings. If students do not have the ability to screen new information and instead memorize it, adoptions of conceptual misunderstandings can occur more readily. Therefore, the levels of abstraction related to a specific topic can be an indicator of the difficulty of that topic for students to conceptually understand. Thus, it might be expected that there will be

many different levels of conceptual understanding and misunderstanding within a student population with regard to the topic of thermochemistry.

Many research studies have demonstrated students have difficulty understanding core concepts in thermochemistry and have identified many student conceptual misunderstandings.[73-75, 78-80, 82-87, 90, 93-95, 102-105] Not all of these studies focus on college-level first-semester thermochemistry taught in a general chemistry course. Many of these studies focus on secondary-school students, some from countries outside the United States. Currently, a concept inventory specifically developed for first-semester college-level thermochemistry taught in general chemistry has not been developed.

Development of the Thermochemistry Concept Inventory. The intended use of a concept inventory determines what format, length, content, and validation protocols are required.[112] Thoughtful and rationale design methodology is important for establishing validity and reliability for an assessment that will provide useful and informative data.[113] The TCI is intended to identify conceptual misunderstanding related to the most important thermochemistry topics being used by students in college-level first-semester general chemistry.

Rationale for the creation of the Thermochemistry Concept Inventory summary. There is no silver bullet that will make learning chemistry easy. A body of research has given guidance to instructors as to ways to make instruction more effective and concepts easier for students to learn in meaningful ways. Student conceptual understanding is a major goal for chemical educators, which is reflected, in part, by the amount of research that has focused on studying student conceptual misunderstandings. One way to improve student conceptual understanding is to use evidence-based instructional

practices that have been shown to work for a given student population. This evidence can be collected through the use of carefully-designed, specifically-targeted assessments that measure student skills or traits that instructional practices aim to target. The need for such assessments is clear, especially for historically difficult courses such as college-level general chemistry. Thermochemistry, taught in first-semester general chemistry, is an example of a content area that has demonstrated a need for evaluation of student conceptual understandings. This need is highlighted by the importance of topics taught in thermochemistry for future understanding in later chemistry coursework and the large body of literature on identified student conceptual misunderstandings. The creation of a thermochemistry concept inventory can help provide evidence necessary to evaluate new instruction strategies to improve students' conceptual understanding of thermochemistry concepts.

LIMITATIONS

This study uses qualitative studies to obtain rich and descriptive information about student conceptual understanding of thermochemistry. However, due to the small sample size and variability of students in any sample, the generalizability of the results of this study are limited. Attempts were made to increase variability by sampling students at institutions that serve different student populations. However, because convenience sampling was used, these most likely are not representative samples.[114] However, quantitative studies provide some insight on the generalizability of results from qualitative studies. For example, a student conceptual misunderstanding identified in student interviews and used as a distracter for an item in the TCI, but was not attractive to students in multiple populations, was an indication that the conceptual misunderstanding

is not widespread in the target population. Another limitation is that not every identified conceptual misunderstanding could be used in the TCI. Choosing which conceptual misunderstandings that were included was based on how many that were identified and on expert importance ratings of thermochemistry topics. Because the expert sample is a convenience sample (and not representative), there is some bias as to these importance ratings, and therefore, selection of conceptual misunderstandings reflect any biases present. However, a bias on which conceptual misunderstandings were used in the TCI would not affect any student population disproportionately, and therefore, is not a major risk to threaten the utility of the TCI.

DEFINITION OF TERMS

Alternative conceptions. The most extreme form conceptual misunderstanding, which is fully integrated into a learner's knowledge framework and can be resistant to conceptual change.

Concept inventory. Diagnostic assessment instrument that uses identified student conceptual misunderstandings as distracters in a multiple-choice item format.

Conceptual misunderstandings. All forms of knowledge that vary from those that are accepted by the scientific community.

Conceptual understanding. Ability to accurately recall conceptual knowledge from long-term memory and applying it to solve novel problems.

Concurrent validity. Evaluates the degree to which the construct being measured by a test correlates to a related criterion.[117]

Consequential validity. Evaluates to what degree does the test display bias to a specific group or sub-group of test-takers that may have adverse affects.[117]

Construct validity. "Construct validity is evaluated by investigating what psychological qualities a test measures, i.e., by demonstrating that certain explanatory constructs account to some degree for performance on the test. . . Essentially, in studies of construct validity we are validating the theory underlying the test".[115]

Content validity. Evaluates how well an assessment instrument defines the content to be evaluated and the criteria used to determine this content.[116]

Convergent validity. Evaluates the degree to which the measured construct correlates to other theoretically-similar or linked constructs, measured by other tests.[117, 118]

Discriminant validity. Evaluates the degree to which the measured construct correlates to other theoretically-dissimilar constructs measured by other tests.[112]

Expert. Faculty member who has traditionally taught in the general chemistry series and who currently teaches first-semester (or equivalent) general chemistry.

Expert response process validity. Evaluates if an item portrays the content correctly, and if it is in agreement with the correct answer and incorrect answers.

Formative assessment. Assessment that is utilized throughout the learning process, providing feedback to both the instructor and the learner.

Item difficulty. Calculates the proportion of correct responses for each item (number correct responses/number total responses).[117]

Item discrimination. Degree to which an item differentiates students with a high total score from those with a low total score.[117]

Item validity. Evaluates if item stems and responses are using only construct-relevant content, if construct-irrelevant content is being used, and to what degree.[117]

Logit. Contracted form of "log-odds unit", commonly used as the unit of measurement for Rasch analysis.

Meaningful learning. Process where new information is contextualized and connected to existing knowledge in substantive, meaningful ways.[44]

Novice response process validity. Evaluates how students understand the item stem, any associated figures, and language and wording of the responses.

Predictive validity. Evaluates the degree to which the construct being measured by a test correlates to a criterion to be measured in the future.[117]

Response process validity. Can be evaluated by either content experts (e.g., instructors) or by content novices (e.g., students), and assesses if the test is measuring what it claims to measure.[117]

Rote learning. Process where new information is imbedded into knowledge structure, but in arbitrary, verbatim ways that isolate new knowledge from pre-existing knowledge structures.[44]

Structural validity. Degree to which the actual test structure matches the designed theoretical structure based on the construct being measured.[116]

Summative assessment. Assessment that is cumulative in nature and generally used for evaluation at the end of an instruction section or course.

Test reliability. Extent to which differences in examinee's observed scores can be attributed to differences in their true score, rather than differences in their error scores.[117]

Thermochemistry. Topic generally taught in the first semester of a two-semester college-level general chemistry series, focusing on foundational topics surrounding the first-law of thermodynamics.

Validity. Degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests.[112]

CHAPTER II

REVIEW OF LITERATURE

The goal of this research was to create a diagnostic concept inventory for thermochemistry as taught in a first-semester college-level general chemistry course. To create a concept inventory, this research needed to be informed by (1) the nature of conceptual understanding and conceptual misunderstanding, (2) the process of conceptual change, (3) the measurement of conceptual understanding, (4) the assessment of construct validity of these measurements, and (5) the psychometric evaluation of concept inventory testing data. Clearly, creation of a concept inventory requires knowledge of both qualitative and quantitative research methodologies. This chapter reviews the necessary background required to understand the foundation of both qualitative and quantitative methodologies employed in this study. This review starts the theoretical framework used in this study. This theoretical framework helps provide context and a rationale for this research.

THEORETICAL FRAMEWORK: CONSTRUCTIVISM

The theory of constructivism can be considered a classical theory, in that its development and use is both historical and the foundation of many contemporary learning theories.[119] The theory is also incredibly useful, as it is intuitive and mirrors many of the same parameters of scientific theories.[66] This is, in part, why the theory of constructivism

is so commonly used within science education research community.[103] This research relies heavily on the theory of constructivism, but acknowledges that learning theories that have grown from constructivism are not without merit and utility. Theories published by Vosniadou, diSessa, and Chi are all contemporary examples of learning theories. These theories' increased level of complexity and abstraction can limit application to cross-discipline research such as this study, but are very important in explaining the complexity of human cognition.[120-122]

An epistemologist named Jean Piaget developed the theory of intellectual development by trying to answer the question, "How is knowledge gained by children?".[65] In doing so, he created a classification of development in children centered on intellectual development. The last two stages of development, concrete operational and formal operational, were important to understanding students studying chemistry. Concrete operational learners transition to formal operational learners about the time students are taking high school and college chemistry courses. Concrete operational learners focus on concrete objects and events of the present and have difficulty with abstract thinking and thinking of the possible (prediction).[123] Formal operational learners have the potential to think of abstract concepts without the aid of visual props and thinking of the possible. However, students who are formal operational thinkers can revert to concrete operational thinkers when encountering unfamiliar concepts.[123] This differentiation in students' ability to understand abstract concepts is important to help explain student difficulties in learning chemistry, where many concepts are abstract and lack visual representations.

Along with the advent of developmental classifications, Piaget also introduced terminology that would be used heavily in future learning theories, including the terms

adaption, assimilation, disequilibration, and accommodation.[65, 123, 124] However, Piaget's theory did not include interactions with peers or teachers, only with the physical environment.[65, 123] A contemporary of Piaget, Lev Vygotsky, introduced the idea of how interactions between instructors and students can allow for learners to achieve success with cognitive challenges above their ability.[125] The work of Piaget, Vygotsky, and their contemporaries, laid the foundation for the theory of constructivism.

Chemistry education researchers began to promote and apply the theory of constructivism, with early proponents being Dudley Herron[123] and George Bodner. Bodner[66] argued that the Piagetian idea of classification of learners (e.g., concrete operational and formal operational) should not be the central focus of instruction, rather the processes that students are using to construct knowledge from their experiences and observations.[66] In addition, some believe that grouping students by Piaget's classifications can be difficult, and might be an oversimplification of the complex process of cognition.[125]

To understand the theory of constructivism, a brief review in some background of the prominent learning theories that were prevalent at the beginning of the constructivist movement is helpful. The traditional notions of learning and knowledge held by many instructors, including many chemists, treated students as being blank slates (tabula rasa). Learners were believed to adopt knowledge verbatim without processing, such that the information passed from instructor to student was identical.[66] The realist perspective used by many, believed that there was only one reality of the world, and if one understands the world, they understand this reality. Thus for learning to occur, knowledge held by the learner needed to match reality.[66] In contrast, constructivists believed that for learning to

be meaningful, it has to fit within existing knowledge and cognitive structures[44] (e.g., schema,[65] frameworks,[126] etc.). The main assumption of the theory of constructivism is that students have prior knowledge that has been constructed based on previous experiences. For new knowledge to be incorporated into preexisting knowledge structures, it must be processed by the learner to some degree. Radical constructivists believe that (1) "Knowledge is seldom transferred intact from the mind of the teacher to the mind of the student", and (2) "Useful knowledge is never transferred intact".[124] The second statement implies that, for knowledge to be useful, it must be able to accurately predict phenomena, to transfer to new problems, and to fit within the existing knowledge structure, rather than be isolated through rote memorization.

Principles of the constructivist theory have been recommended by the NRC, including two key findings from a large body of research of learners and learning:

> (1) Students come to the classroom with preconceptions about how the world works. If their initial understanding in not engaged, they may fail to grasp the new concepts and information that are taught, or they may learn them for the purpose of the test but revert to their preconceptions outside the classroom.
>
> (2) To develop competence in an area of inquiry, students must: (a) have a deep foundation of factual knowledge, (b) understand facts and ideas in the context of a conceptual framework, and (c) organize knowledge in ways that facilitate retrieval and application.[61]

The first finding suggests that instruction must recognize and address student prior knowledge if meaningful learning is to occur, as students will need to process new information such that it can be incorporated into existing cognitive structures. Instruction should help facilitate this process. The second statement highlights the importance of contextualizing factual knowledge within existing cognitive structures, such that mean-

ingful connections to prior knowledge are built and allow for accurate and fast retrieval and application.

In addition to a large body of research that supports the statements above, the constructivist theory of knowledge mirrors the scientific notion of theories.[66] Students acquire knowledge and construct meaning by contextualizing and processing it such that it fits into preexisting cognitive structures. This is all the result of trying to explain observed phenomena and help make predictions of future events. This is analogous to how scientists build theories. If the predictive power of a conception is weak or inaccurate, it will not be useful and will be reevaluated, just as scientific theories are revised or replaced when scientists are presented with anomalous experimental data.

In summary, the constructivist theory of learning is founded on the belief that knowledge is constructed in the mind of the learner.[66] Knowledge is rarely transferred intact from the mind of the instructor to the mind of the learner, and if it could be, it would not be useful to the learner. Rather, new knowledge must be processed by the learner, through contextualizing and integrating into existing cognitive structures. These cognitive structures are dynamic and can adapt to anomalous observations and information.

DEFINING STUDENT UNDERSTANDING

The theory of constructivism explains the process of student learning, but does not provide as many details of the cognitive structures that have been built through the learning process. The phrase meaningful learning has been used in describing the constructivist theory, but it has yet to be defined in this dissertation. David Ausubel's

seminal work in cognitive psychology helped to define meaningful learning. Ausubel defined meaningful learning as the process where new information is contextualized and connected to existing knowledge in substantive, meaningful ways.[44] For meaningful learning to occur, Ausubel argued three conditions need to be met:

> (1) A student must have some relevant prior knowledge to which the new information can be related in a non-arbitrary manner.
>
> (2) The material to be learned must be meaningful in and of itself; that is, it must contain important concepts and propositions relatable to existing knowledge.
>
> (3) A student must consciously choose to non-arbitrarily incorporate this meaningful material into his/her existing knowledge.[68]

The first condition for meaningful learning is critical to the second and third conditions. According to Ausubel, prior knowledge is the key component to meaningful learning. Instruction, therefore, should be informed by students' prior knowledge, to facilitate making meaningful connections between new and prior knowledge. Peter Novak built upon Ausubel's definition of meaningful learning, arguing that additional conditions need to be met outside the cognitive domain, including affective and psychomotor domains.[68] One result of meaningful learning is higher rates of retention and transfer of knowledge.[127]

Rote learning is the process where new information is imbedded into knowledge structure, but in arbitrary, verbatim ways that isolate new knowledge from pre-existing knowledge structures.[44] This is a counterexample to meaningful learning, which non-arbitrarily makes meaningful connections to prior knowledge. Two key differences between rote learning and meaningful learning are a lack of contextualization (e.g., arbitrary relation of new and prior knowledge) and isolation of new information from

existing knowledge structures. Rote learning often will not lead to long-term incorporation of knowledge due to this lack of integration.[44] Unfortunately, many science classrooms, including chemistry, condition students to revert to rote learning.[128] However, this often does not reflect the learning goals of chemistry instructors. If students' grades are based on test scores, and the tests are primarily assessing factual, algorithmic or recognize and regurgitate knowledge, then rote learning will be promoted and rewarded. This is not to say that students who have interacted and learned the course content in a meaningful way will not do well on tests. However, these types of tests might not differentiate the rote learners from the meaningful learners and will not promote meaningful learning. Thus, if the learning goals for instruction include meaningful learning, tests should include questions that make students apply knowledge to novel problems, where transfer of knowledge is necessary.

The complexity of categorizing human cognition and the emphasis of meaningful learning is illustrated in the revision of Bloom's Taxonomy. The original Bloom's Taxonomy (*Taxonomy of Educational Objectives*)[129] was published in 1956 and aimed to provide educators with a hierarchal system to classify educational goals and objectives. The hierarchy included six levels: knowledge, comprehension, application, analysis, synthesis, and evaluation. This taxonomy was designed to be unidimensional, but this was only accomplished by reducing the knowledge category to fit the structure of the other levels.[130] All five other categories represent different cognitive processes. However, the knowledge category was an amalgam of all types of knowledge and did not differentiate rote learning (e.g., memorizing factual information) from meaningful learning (e.g., meaningful understanding conceptual knowledge).[130] This problem was addressed 45

years later with the Revised Bloom's Taxonomy (*A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*)[131] which created a second dimension to the original taxonomy: knowledge. The knowledge dimension includes four categories: factual, conceptual, procedural, and metacognitive. The cognitive process dimension has six categories: remember, understand, apply, analyze, evaluate, and create. Thus, this new taxonomy can distinguish remembering factual knowledge from understanding conceptual knowledge, or in other words, rote learning from meaningful learning. This revised Bloom's Taxonomy also underscores the difficulty in creating a hierarchical taxonomy for educational objectives of understanding.

CONCEPTUAL UNDERSTANDING

The study of student conceptual understanding has been of interest to the chemical education research community for over 25 years. The first attempt to measure conceptual understanding can be traced back to Nurrenbern and Pickering,[46] who first attempted to differentiate students' ability to solve traditionally-formatted open-ended problems and matched conceptually-based multiple-choice questions containing no mathematical calculations. This study was a nucleation site for future research that centered on conceptual understanding of chemistry topics.[40, 88, 89, 132-137] Many studies followed the hypothesis put forth by Nurrenbern and Pickering, that conceptual understanding might not be required to successfully answer traditional chemistry problems. These studies still continue today, including much larger data sets and more sophisticated statistical analysis methods.[132] The results of the Nurrenburn and Pickering study, which found that students could indeed successfully solve traditional problems but not conceptually-based problems, led to the core question of student conceptual understanding: How can researchers

measure student conceptual understanding if not by traditional open-ended questions, given the time-consuming nature of evaluating these types of question? This led to an even more basic question: What are conceptual understandings? This section examines how the chemical education research community grappled with defining, explaining, and examining student conceptual understanding to accurately assess this understanding.

Ironically, conceptualizing what a concept is can be rather difficult. Defining a concept operationally is actually easier. Vygotsky believed that concepts allowed for the simplification and generalization of reality as to allow for a uniform meaning to be used by members of a culture.[138] In addition, Ausubel believed that formation of concepts allows for (1) creation of cognitive structures, where generic constructs can be correlated, differentiated, and applied in meaningful associations, and (2) the ability to apply knowledge through cognitive processes to form hypothesizes and solve problems.[44] Concepts are more than just arbitrary facts, because they are generalizations often used to simplify complex processes into meaningful chunks of knowledge. Conceptual knowledge, then, would be knowledge that is in the form of a concept.

Conceptual understanding relates to the ability to use conceptual knowledge. Specifically, conceptual understandings can be defined as the ability to accurately recall conceptual knowledge from long-term memory and apply it to solve novel problems. The ability to quickly and accurately recall conceptual knowledge demonstrates some contextualization of knowledge.[61] The ability to apply knowledge to novel problems discriminates conceptual understanding from employment of problem recognition and procedural knowledge recall.[139] Thus, conceptual understanding is directly related to an ability to apply conceptual knowledge, which is a cognitive process.[69] There is a spectrum

of cognitive processes, as highlighted by the revised Bloom's Taxonomy. The process of creation using conceptual knowledge is more difficult than simply remembering a concept,[130] demonstrating a higher level of conceptual understanding in the revised taxonomy. As discussed earlier, conceptual understanding can be described as a continuous process, which is theoretically infinite. However, for conceptualization purposes, a finite continuum based on expert conceptual understandings presented in Chapter I (Figure 2). This expert-based continuum is applicable to teaching environments where teachers are considered experts.

To summarize, conceptual knowledge is to know and conceptual understanding is to use.[69] There are varying degrees to what someone knows and how much they can use what they know. The revised Bloom's Taxonomy is a very good demonstration of this distinction, where it separates the knowledge domain from the cognitive process domain. An important feature of conceptual understanding is that it involves cognitive processes with distinguishable benchmarks. To measure conceptual understanding is to measure the degree of conceptual understanding on a continuum.

CONCEPTUAL MISUNDERSTANDING TERMINOLOGY

Terminology for describing student conceptual understanding difficulties in science and chemical education research has been confusing and sometimes contentious, to put it mildly. Calls for some consensus on what to call student conceptual misunderstandings have occurred for the past 20 years. [23, 140]

Nomothetic Versus Idiographic Studies

A categorical approach to provide rationale for terminology has been provided, which differentiates terms associated with nomothetic studies from those associated with idiographic studies. Nomothetic studies are grounded in the belief that knowledge is assessed based on the agreement with accepted principle (e.g., accepted scientific knowledge). Terms associated with nomothetic studies include misconceptions, mistakes, erroneous ideas, errors, naïve conceptions, misunderstandings, preconceived ideas, preinstructional ideas, persistent pitfalls, conceptual difficulties, children's conception, preconceptions, prescientific conceptions, naïve theories, and conflicting schemas.[23, 140] Nomothetic studies generally check for congruence and are often quantitative in nature, such as paper-and-pencil tests.[23] Alternatively, idiographic studies are grounded in the belief that knowledge is individualized, but may contain common themes that can be identified and addressed in order to facilitate meaningful learning. Terms associated with idiographic studies include alternative conceptions, alternative frameworks, alternative ideas, developing conceptions, personal constructs, intuitive beliefs, existing or prior conceptions.[23, 140] Idiographic studies are generally qualitative, and based on thick and descriptive data from individuals, such as from interviews. This classification system does not work in many research studies.[23] Specifically, some studies are a hybrid of both frameworks, and choosing a term becomes somewhat arbitrary, as is the case for many studies in chemical education.[74, 76, 86, 93, 98, 137, 140-143]

Misconception Versus Alternative Conception

The two most common terms used in chemical education literature are misconceptions and alternative conceptions; however, similar or sometimes identical definitions are cited for these two terms. For example, Mulford and Robinson define alternate conceptions as "concepts that are not consistent with the consensus of the scientific community".[88] Yet, Nakhleh defines a misconception as "any concept that differs from the commonly accepted scientific understanding of the term".[141] Basically, the same definition for two terms that, many argue, have very different theoretical implications.[141] Further confusion over these terms may be due to authors who try to distinguish between alternative conceptions and misconceptions, but do not put forth any theoretical framework as the basis for their definitions. For example, Abimbola defines a misconception as, "an idea that is clearly in conflict with scientific conceptions and is therefore wrong" and an alternative conception as, "an idea which is neither clearly in conflict nor clearly compatible with scientific conceptions, but which has its own value and is therefore not necessarily wrong".[144] In this context, alternative conceptions are in conceptual limbo, neither right nor wrong.

The terms alternative conception and alternate conception are synonymous and convey a constructivist theoretical framework.[145] The term misconception is rooted in a positivist theoretical framework, where there exists the truth or expert conception and an incorrect conception which is simply not the truth.[114, 145] Often, as Abimbola's definition highlights, misconception is synonymous with mistake.[140, 145] This is in direct conflict with a constructivist theoretical framework, as learners who construct and make meaning

of their experiences, often construct concepts that vary from expert conceptions, but make sense from their own experiences.

Alternative Conception Versus Alternative Framework

The terms alternative conception and alternative framework have also been used synonymously in chemical education literature. The term alternative framework was first used by Driver and Easely,[146] who used the term framework as a reference to the constructivist theoretical framework, where conceptual frameworks are often used to describe conceptual knowledge. Driver defines alternative frameworks as "the situation in which pupils have developed autonomous frameworks for conceptualizing their experiences for the physical world".[146] However, the term has been commandeered by other researchers to imply that an alternative framework is a collection of alternative conceptions.[74, 142] This, not surprisingly, has caused confusion and some have argued that alternative conception is more inclusive and should be used instead of alternative frameworks.[140]

Conceptual Misunderstanding as an Inclusive Term

Is there a term that can be agreed upon and used by the chemistry education (and science education) community to describe student conceptual difficulties? Novak tried to answer this question 30 years ago, proposing the acronym LIPH (limited or inappropriate propositional hierarchies), which has not been used by his peers because it was too explicit.[145] Chemistry education researchers use multiple theoretical frameworks; thus, it is not surprising that there is disagreement of terms. To complicate matters, in some cases, constructivists even opt to use the term misconception, as it is already commonly

used in colloquial speech and can be used without definition when communicating outside the educational research community. In addition, some researchers have chosen to use the term incorrect idea.[143] rather than alternative conception, because alternative conceptions are understood to represent strongly held conceptions, where an idea is more malleable in the mind of the learner. Furthermore, the complexity of conceptual understanding cannot easily be summed up with one word; therefore, many terms have been created to distinguish types of conceptual difficulties. This study chooses to use the inclusive term conceptual misunderstanding, which represents all forms of student knowledge that vary from those which are accepted by the scientific community.

Though creating a new term might appear to increase the entropy of the conceptual terminology system, in actuality, it does the opposite. It allows for other terms to be included in a theoretical framework that both conceptual understanding and misunderstanding are on a single continuum, and this continuum has many benchmarks that can be defined by current terminology. Thus, researchers can describe any student conceptual difficulty as a conceptual misunderstanding, but elaborate on the degree of the misunderstanding and use whatever term they feel appropriate based on their own perspective.

CONCEPTUAL MISUNDERSTANDINGS IN THERMOCHEMISTRY

Conceptual misunderstandings have been reported in all major student populations (primary school through post-secondary) and for instructor populations (secondary instructors and post-secondary instructors). Many of these studies are for student populations outside the United States, where thermochemistry is sometimes taught in secondary school.

Table 1 summarizes the thermochemical conceptual misunderstandings indentified in qualitative research studies focusing on varying student populations. The thermochemistry topics are organized based on themes discussed in Chapter IV. Different levels of coverage (e.g., definition, sign, and understanding) are used to distinguish conceptual misunderstands further for broad topics, such as enthalpy.

Table 1. Review of Literature on Conceptual Misunderstandings in Thermochemistry

Topic	*Literature Reference*	*Student Population*
<u>Definition—Heat</u>		
Heat is a substance (noun)	94, 117; 147	S, Int
Heat and enthalpy are the same thing	73	PS
Heat and temperature are the same thing	75; 85	S, Int
Heat and thermal energy are the same thing		
Heat is energy that is added to something	93	PS
Heat can be quantified as a specific amount of energy possessed by a body with temperature being a measure of that quantity	85, 87	S, Int; PS
Heat is a state function	87	PS
Heat is a substance that resides within objects and can pass from one to another	94; 78	S, Int; PS, Int
Heat is an extensive property	85	S, Int
<u>Definition—Enthalpy</u>		
Enthalpy is the change in heat	73	PS
Enthalpy is a measure of heat contained within a system	73	PS
<u>Sign—Enthalpy</u>		
The sign of Δh_{rxn} cannot be determined without using tabulated values for enthalpy	93	PS
<u>Understanding—Heat of Reaction (Δh_{rxn})</u>		
Emthalpy change is the same as internal energy change	93	

Table 1. (continued)

Topic	*Literature Reference*	*Student Population*
Definition—Temperature		
Temperature is an extensive property of a substance or body	90	PS, Int
Temperature is an accurate measure of heat	85; 87	S, Int; PS
Addition of thermal energy to a system will always result in an increase in temperature	85; 148	S, Int
Bodies with the same temperature will have the same energy	85; 94	S, Int
The temperature of an object can be accurately be determined by touch	94	S, Int
Temperature is a property of the substance from which the body is made	94	S, Int
Understanding—Endothermic and Exothermic		
In an exothermic process heat enters the system, and in an endothermic process heat exits the system	90	PS, Int
Endothermic processes require energy to occur	76; 74	PS, Int; S
Exothermic reactions occur faster than endothermic reactions	149; 90	SI, Int; PS, Int
Understanding—Difference between q and T		
Heat and temperature are the same thing	75	S, Int
Definition—System and Surroundings		
The system includes everything you are studying and the surroundings is everything else	150	PS
Consideration of the surroundings is not required when evaluating energy transfer process, only the system	151	PS, Int
Understanding—Heat of Formation (Δh_r)		
$\Delta H°_f$ is always exothermic	90	PS, Int

Table 1. (continued)

Topic	*Literature Reference*	*Student Population*
Understanding Bond Disassociation Energy		
Heat is released when a bond is broken	74; 152; 90	S; PS; PS, Int
Chemical bonds are structures that require energy because they need to build	74	S
Definition—Work		
When work is done on a system, heat is added to the system	87	PS
Work is a state function	87	PS
Understanding—Conditions for Thermal Energy Transfer		
Once thermal equilibrium is reached, small differences in temperature between two bodies can exist	85	S, Int
Other factors can affect thermal energy transfer besides temperature differences between two bodies	85	S, Int
The transfer of thermal energy is synonymous with energy conversion	83	PS

Note. S = secondary, PS = post secondary, SI = secondary instructor, PSI = post-secondary instructor, Int = international (outside United States)

THEORIES IN CONCEPTUAL CHANGE

Changing one's conceptual knowledge is difficult. Just how difficult and why it is difficult depends on the conceptual change theory utilized. Four different theories are briefly presented, one that was used for the framework for this research and three alternatives. All theories, however, agree that conceptual change is a process that is often difficult and gradual.

Classical Conceptual Change Theory Approach

This research relied most heavily on the classical conceptual change theory approach, which was presented in the introduction. The seminal paper of Kuhn[153] and expanded upon by Posner[96] were both a part of the constructivist movement and were heavily influenced by the work of Piaget. This theory put forth that conceptual misunderstandings are always resistant to change and need to be replaced. The theory promotes targeted instruction to make learners dissatisfied with their conceptual misunderstandings and provide intelligible conceptions for replacement. This dissatisfaction is believed to be a result of an anomaly of data or observations regarding phenomena related to the conceptual misunderstanding. In some instances, conceptual change can occur as quickly as learner dissatisfaction and adoption of new (correct) conceptions. This is in contrast to the more complex models of conceptual change, presented below.

Framework Theory Approach

The first alternative to the classical approach is the framework theory approach, championed by Vosinadou,[122] which proposes that knowledge systems are complex, hierarchical, and dynamic. Vosinadou puts forth the belief that knowledge is structured into domain-specific frameworks from an early age, initially through interaction and observations with the environment. As scientific knowledge, for example, is learned, incongruities will emerge with existing knowledge frameworks, resulting in two possible outcomes: (1) the existence of the inconsistency is acknowledged by the learner, and conceptual change can occur, much like in the classical approach; or (2) the inconsistency is not acknowledged, and a synthetic model is created. The idea of a synthetic model is

unique to this theory, where an instructionally-induced entity coexists with conceptual misunderstandings. Thus, conceptual change requires, in the case of synthetic systems, a much longer time to occur, as students might not be aware of inconsistencies in their knowledge framework.

Knowledge in Pieces Approach

In contrast to the framework theory approach, diSessa's[121, 154] knowledge in pieces approach does not assume that knowledge is placed into frameworks as it is acquired, rather it can initially be autonomous. This knowledge is gained from superficial interpretations of physical phenomena. These simple pieces of knowledge with individual, autonomous meaning are called phenomenological primitives (p-prims). Novice learners have knowledge systems composed of unstructured p-prims that they organized into meaningful knowledge structures. As p-prims are organized into meaningful knowledge structures, they lose individual meaning and gain the meaning of the structure as a whole. Conceptual change can be slow, since incorrect p-prims can be imbedded in knowledge structures, adding to the summative meaning of these structures. Therefore, conceptual change can involve changing knowledge structures, rather than individual p-prims.

Three Types of Conceptual Change Approach

Unique to the other conceptual change approaches, Chi[120] has put forth a theory that categorizes conceptual change into three different grain sizes, based on resistance to conceptual change. Conceptual understanding can be hierarchical, Chi explains, consisting of beliefs, mental models, and categorizations. A belief is a single idea within a specific domain of knowledge. A mental model is an organized collection of beliefs.

Categorization of knowledge represents the process of assigning beliefs (concepts) to specific categories. Chi also argues that categorical knowledge can be lateral or ontological, as well as hierarchical. Resistance to conceptual change will then depend on where the concept resides in the knowledge structure, and at what grain size the conceptual change is occurring. A belief revision demonstrates the least resistance, and is similar to conceptual change described by the classical model of conceptual change. A mental model revision includes revision of more than one belief, and is therefore, more difficult than a belief revision. A categorical shift is very resistant to change, as it includes reclassifying concepts, which can be very difficult.

To summarize, all theories agree that conceptual change does not simply imply conceptual exchange. The process of conceptual change can be slow, and can require significant facilitation from external sources, such as targeted instruction. How knowledge is acquired, processed, and structured varies between the four theories discussed, and determines how conceptual change can occur. The three alternative approaches are very theoretical and reflect the complex nature of human cognition. Empirical evidence in support of these theories continues to be collected,[119] which provides insight for future studies regarding conceptual change.

MEASURING CONCEPTUAL CHANGE

With over 75 years of research identifying student conceptual misunderstandings, there is still much to accomplish to create evidence-based instructional techniques that will facilitate conceptual change.[155, 156] This is partly because obtaining evidence of the effectiveness of new instructional techniques can be difficult to obtain. Developing and

implementing instructional techniques using true experimental methodology (e.g., treatment and control groups, random sampling, etc.) is difficult on a large scale. In addition, the availability of assessment instruments developed specifically to measure conceptual change is limited. Instructional techniques to promote conceptual change are discussed, along with ways to measure associated conceptual change.

Epistemological, Ontological, and Affective Focused Instruction

There are many proposed instructional techniques to facilitate conceptual change.[142] The focus of these different techniques will depend on what theory of conceptual change is being used, but can be generalized into three main categorizes:[155] epistemological focus, ontological focus, and affective focus. Instruction with epistemological focus seeks to examine how students can view the same conception in different contexts. For example, in thermochemistry the term heat refers to the process of thermal energy transfer from one thermal body to another. However, the term heat in a colloquial context has many meanings, some which are at odds with the scientific definition. Thus, epistemological instruction tries to identify alternative views of conceptions in different contexts, which can inform instruction that facilitates student understanding of the concept in a scientific context.

Instruction with an ontological focus looks to change the way students view reality. Using heat, again as an example, an ontological instructional focus would aim to change student perceptions of heat being a substance (incorrect) to heat being a process (correct). This is related to the epistemological definition, but involves possible

recategorization of the concept of heat, which is a more difficult form of conceptual change.[120]

Instruction with an affective focus tries to motivate, excite, or engage students in learning about conceptions. This is very important for any type of conceptual change to occur, because most theories have the assumption that students care about understanding and using conceptual knowledge, and that this conceptual knowledge is correct and can make accurate predictions. To extend the heat example, instruction with an affective focus would try to make students excited about and to see the relevance of the concept of heat.

Cognitive Conflict and Model Building Instructional Techniques

Two specific types of instructional techniques that can be used in epistemological, ontological, or affective focused instruction are cognitive conflict strategies and model building strategies. These strategies are not mutually exclusive and can be used synergistically, depending on the nature of the desired conceptual change.

Cognitive conflict was a term made popular by Piaget[65] and cognitive dissonance was later introduced by Driver and Erickson.[157] The terms are synonymous and relate to the classical approach to conceptual change. As Posner argued, students need to be dissatisfied with conceptual misunderstandings before the process of conceptual change can begin.[96] Cognitive conflict involves presenting situations where an anomaly can occur, generally a specific demonstration or problem where student conceptual misunderstandings will not provide the correct prediction or answer.[96, 155] Studies assessing the effectiveness of using cognitive conflict, however, have had contradictory results.[155] This could be

because conceptual change can be a slow process and might need iterative instructional interventions before resistant conceptual misunderstandings are replaced.[122] In addition, a lack of assessments sensitive to conceptual change could underestimate change or could be prone to a large error with small sample sizes. More experimental-design studies using assessments developed to be sensitive to the desired change need to be conducted before the efficacy of cognitive conflict can be measured. One type of assessment instrument that can be developed to measure conceptual change is a concept inventory. When common student conceptual misunderstandings are used as distracters in multiple-choice items, cognitive conflict can occur. When students choose a distracter that they believe is the correct conception, but find that it is incorrect, they are provided the opportunity ask why their conception is not correct. When used as formative assessment, concept inventories can motivate students to ask "why?", given that they might want to know the correct conception for future summative assessments (e.g., midterm or final exam). Because conceptual change is a self-motivated process, students need to be given a reason to care about knowing what the correct conception might be.[158] Given that formative assessments are designed to be used during the learning process, concept inventories, used as formative assessments, can also be effective for more resistant conceptual misunderstandings. Resistant conceptual misunderstandings might need iterative cognitive conflicts, since the conceptual process is difficult and might need multiple interventions.[120-122]

Model building instructional techniques are most often associated with the framework theory and the three types of conceptual change approaches.[158] Of the many types of possible models students can construct, semantic and causal models are more

commonly used to promote student conceptual change. Model building requires assembling elements together in meaningful ways. This requires making choices; making choices is where students are engaged in the learning process, including conceptual change. Semantic models are used to represent structural knowledge, components of a system, and semantic relationships between these components.[158] Concept maps are an example of semantic models. Concept maps can be used to examine student conceptual structure in a particular domain of knowledge. Certain conceptual misunderstandings can be identified as incorrect semantic relationships, which can be visualized by the learner using concept mapping. An important caveat of semantic models is that they do not provide causal relationships. Thus, conceptual knowledge represented in semantic models are necessary to measure conceptual understanding, but not sufficient. Causal relationships among components in a system allow for predictions to be made. The usefulness of many concepts can be related to how accurate and reliable are the predictions made using these conceptions. Conceptions that cannot accurately predict phenomena are not considered useful and cause dissatisfaction, a requirement in the classical conceptual change theory.[96] Thus, causal models show not only the structure of knowledge, but also system knowledge, specifically causal relationships. These models have become more popular with the advent of computer software packages that allow students to build and test models in the programming environment (*in silico*). Computer modeling has the distinct advantage of testing the predictive power of models as they are being built and revised.[158]

The measurement of conceptual change for both concept inventories and concept mapping requires assessing changes in conceptual understanding. Analysis of concept

inventory data must include assessment of the validity and reliability of interpretations made from this data. However, analysis of student responses, specifically changes in responses, can be used to assess conceptual change. Analysis of student-generated concept maps can be qualitative or quantitative,[142] and can be time consuming. However, use of computer software programs to analyze concept maps generated *in silico* have made analysis easier.[158]

Qualitative Versus Quantitative Conundrum

Arguably, the most informative and accurate measure of conceptual understanding and, therefore, conceptual change is one-on-one qualitative cognitive interviews. Experienced cognitive interviewers can obtain thick and descriptive details of student conceptual understanding, using real-time analysis of student responses and probing theses responses with insightful follow-up questions. However, the time required to recruit interview participants, conduct an interview, and analyze the data is extensive and impractical for diagnostic uses. Slightly less informative qualitative methods include computer-based open-ended assessments that can collect student-generated responses to conceptual prompts.[159] These assessments still require significant time for analysis, even with the aid of software to help code student responses. Student generated concept maps also fall into the category of time-consuming analysis, unless specific computer software is used to generate these concept maps. There is always a cost-benefit analysis on what is the most appropriate measure of conceptual change. Courses with large enrollments need more quantitative-based assessment (e.g., concept inventories) while smaller-enrollment courses have the option to use more qualitative methods. Ideally, quantitative-based

assessments should be developed and evaluated using qualitative methods (e.g., student interviews), to establish the validity and reliability of interpretations that can be made from quantitative data.

To summarize, there are different theories to explain and to measure conceptual understanding and conceptual change. What theory and method that are most appropriate depends on the type of conceptual understanding and conceptual change that instructors are interested in assessing. Limitations due to class size, instructional time, and assessment time will also be factors in deciding what methods are appropriate. However, the most important factor to assessing conceptual understanding and conceptual change is informing instructors about conceptual change theory and practices.[155] This includes making instructors aware of student conceptual misunderstandings and evidence-based instructional techniques to address these misunderstandings.

EVIDENCE FOR CONSTRUCT VALIDITY

Validity is the degree to which evidence and theory support the interpretations and use of test data.[112] Validity is not a property of a test; it is actually not even a property. A test can never be validated or claimed to be valid, because validity inherently focuses on the use and interpretation of test data and not the test itself.[160] Furthermore, test data cannot be considered valid, only the interpretation and use of that data. Likewise, a test item cannot be considered valid, only the use and interpretations of test item data. The key focus of evaluating validity is the interpretation and use of test data. The specific population the test was designed for, testing administration conditions and scoring of test responses, are all key factors influencing this evaluation. This information should be

indicated by test developers and followed by test users.[160] For example, administering the SAT to sixth-grade students to predict success in middle school would not be a valid use of test data. This is because the SAT was developed specifically to predict success (e.g., cumulative college grade point average) in college by high school students.[161] In addition, validity is not an all-or-nothing evaluation. There is no one empirical measure that can be used to evaluate evidence for the validity of interpretations of test data to provide a number for others to easily evaluate. Validity can be categorized as strong evidence for or weak evidence for specific types of interpretations of data, based on the holistic evaluation of all validity evidence, depending on the type of test and the intended use.[117] Establishing validity can include assessing item validity, especially if individual item responses are to be used in addition to, or in replacement of the total score, such as can be the case for diagnostic concept inventories.

The term construct validity was first proposed by an American Psychological Association (APA) Committee on Psychological Tests and Diagnostic Techniques in 1954,[115] specifically by Meehl and Challman.[116] This was later expanded upon by Cronbach and Meehl,[118] and even more so by Loevinger.[116] Cronbach and Meehl explained that "construct validity is evaluated by investigating what psychological qualities a test measures, i.e., by demonstrating that certain explanatory constructs account to some degree for performance on the test. . . Essentially, in studies of construct validity we are validating the theory underlying the test".[115] Loevinger argued that construct validity is made up of three components: a substantive component, structural component, and external component.[116]

This idea that construct validity is an over-arching validity trait of a test, which all other validities could be used to establish, was endorsed and expanded upon in the 1999 *Standards for Educational and Psychological Testing*, by the American Educational Research Association (AERA).[112] The components to construct validity are illustrated in Figure 4. This contemporary view of construct validity is composed of five categories: test content, response process, internal structure, association with other variables, and consequence of use. The new standard uses a more inclusive definition of a construct as " the concept or characteristic that a test is designed to measure".[112] The following sections will discuss each of the five categories used to establish construct validity for the use and interpretation of test data.

Test Content

Content validity. Content validity evaluates how well an assessment instrument defines the content to be evaluated and the criteria used to determine this content.[116] The process of defining and determining appropriate content for an assessment is just as important in establishing content validity as providing the content selected for an assessment. Specifically, what are the criteria for establishing what is within the universe of the construct and who established these criteria? In addition, what are the credentials required for those who make these decisions? If the entire universe of the defined content is not to be sampled equally (construct underrepresentation), then clear indication of what parts of the universe are represented in the assessment is necessary.[160] Experts in the field of the construct can be used, including rating the importance of content based on the use and interpretation of test scores.[112] Moreover, experts can evaluate test items for

construct-irrelevant content needed to answer the item, as this could give bias to specific populations of students.[112]

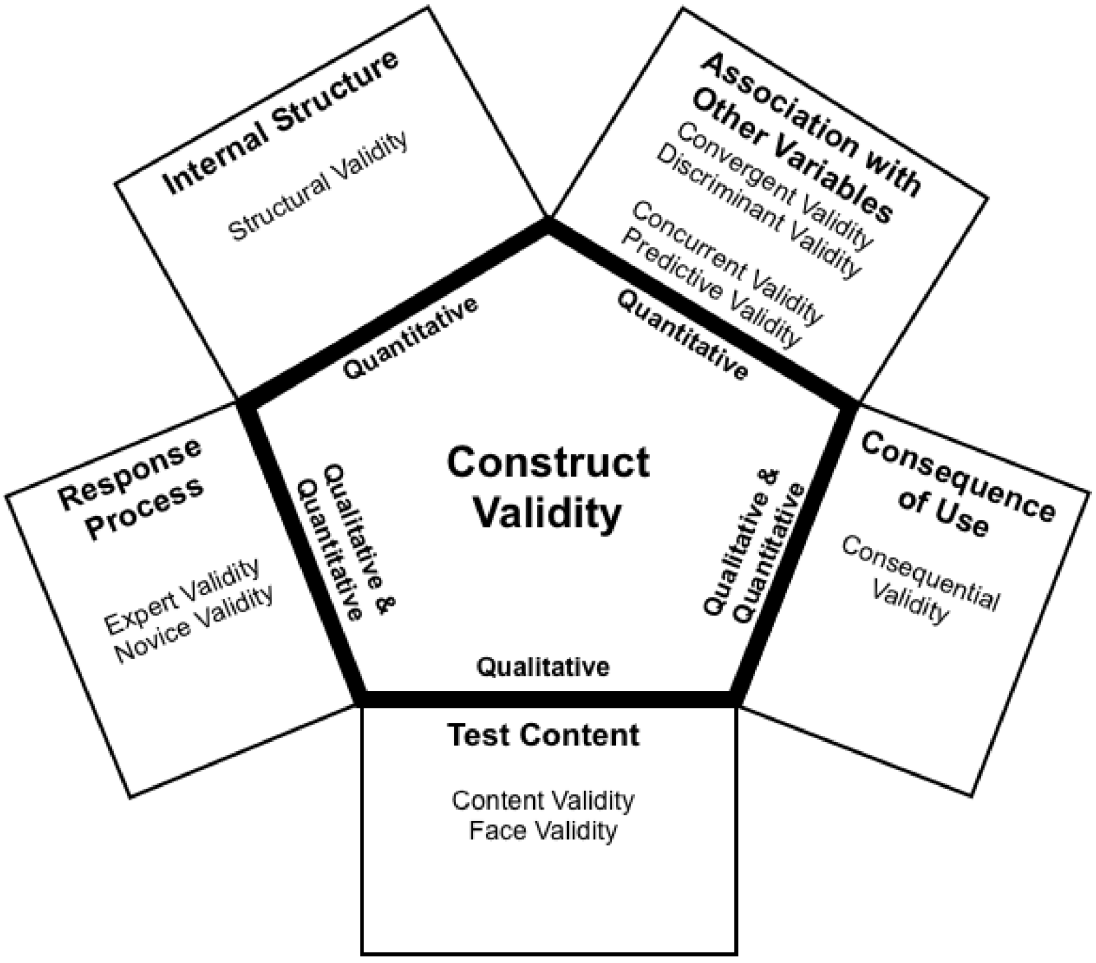


Figure 4. Lines of investigation used to establish construct validity using qualitative and quantitative studies.

Face validity. Face validity of a test can be evaluated by either content experts (e.g., instructors) or by content novices (e.g., students) to assesses if the test is measuring what it claims to measure.[117] This is important, because it could affect the motivation of

the instructor to give the assessment and students' motivation while taking the test and the honesty with which they answer items.[117] Evaluating face validity is subjective by nature, and is, therefore, a weaker form of construct validity evidence.

Item validity. Item validity is a higher-resolution form of content validity. Item validity evaluates if item stems and responses are using only construct-relevant content, if construct-irrelevant content is being used, and to what degree. In addition to what degree, is the content that should be evaluated by test items actually evaluated by test items? This evaluation includes to what degree the defined content is covered by test items and to what depth of content knowledge probed.

Internal Structure

Structural validity. Structural validity is the degree to which the actual test structure matches the designed theoretical structure based on the construct being measured.[116] If a test is designed to be unidimensional (only measure one construct), it should be established that only one construct is being measured. Likewise, if a test has two distinct and defined constructs being measured, the test structure should demonstrate that items associated with each construct be highly correlated with each other and not with other items.

Response Process

Evaluating the validity of the process that is used to answer test items is at the item level. Therefore, all three types of validity in this category are specifically at the item level.

Item validity. If data from individual items are to be interpreted and used, then evidence for these interpretations and uses need to be collected.[162] This is critical, because

errors associated with interpretation of individual item responses are susceptible to more error then test score data.[160] Both quantitative and qualitative evaluation can be used to provide evidence for item validity. Quantitative evaluation is most commonly used to inform qualitative studies. Classical test theory estimates of difficulty and discrimination might not match the test developer's expectations, and student interviews may be needed to clarify these discrepancies. More generally, qualitative evidence includes expert surveys, expert interviews, and student interviews. This evidence is commonly used to help establish expert and novice response process validity.

Expert response process validity. Expert response process validity specifically evaluates if an item portrays the content correctly, and if it is in agreement with the correct answer and incorrect answers. This can be an iterative process, especially when there is disagreement between experts.

Novice response process validity. Novice response process validity evaluates how students understand the stem, any associated figures, and language and wording of the responses. In addition, information about what specific content knowledge, conceptual knowledge, and cognitive process students are using to answer the item is collected. Importantly, to what degree does this evidence match what knowledge and conceptual processes were intended for students to use to answer the item by the item developers.

Evaluating response process validity for items, both at the expert and novice levels, provides meaningful information during the development process of a test. If the response process that students should be using to answer an item is different from the one students are using, then the validity of interpretations and uses of responses to that item is threatened.[117] Likewise, threats to validity can be evaluated by experts by determining if

the construct being measured is accurately portrayed by test items or if there are multiple correct answers or multiple ways to interpret item responses. Therefore, response process validity of items is more rigorous than response process validity for the test, relying on both detailed quantitative and qualitative analysis.

Association with Other Variables

If the psychological construct being measured by a test has other theoretical associations to other variables, then the test score should demonstrate those relationships with measures of those variables.[117] Evidence for validity underlying these relationships can be separated into criterion-related validities (concurrent validity and predictive validity) and those not related to any criterion, but to the actual construct (convergent validity and discriminant validity).

Convergent validity. Convergent validity evaluates the degree to which the measured construct correlates to other theoretically-similar or linked constructs, measured by other tests.[117, 118] An example of two theoretically-similar constructs that should correlate would be self-esteem and happiness.[117] Empirical estimates, such as correlative estimates, can provide convergent evidence.[117] Not all constructs will have known relationships with other constructs, or have tests available to measure related constructs.

Discriminant validity. In contrast to convergent validity, discriminant validity evaluates the degree to which the measured construct correlates to other theoretically-dissimilar constructs measured by other tests. If two constructs are theoretically distinguishable, a test measuring one construct should not measure the other construct. An example of two discriminant constructs would be self-esteem and intelligence.[117] This can

be evaluated by estimating the correlation of the test scores, which would be expected to be low if the two constructs are dissimilar and distinguishable.

Concurrent validity. Concurrent validity evaluates the degree to which the construct being measured by a test correlates to a related criterion. Both the test and the test measuring the related criterion need to be administered at the same time point, or concurrently. An example of a criterion is "competence in chemistry as measured by performance on the ACS standardized final exam". The evaluation of concurrent validity of a test is highly dependent of the validity of the criterion.[118]

Predictive validity. Predictive validity evaluates the degree to which the construct being measured by a test correlates to a criterion to be measured in the future. Examples of criteria for predictive validity could be success in college, job performance, or final course grade. These criteria can represent the reason the test has been created, thus should be very closely related to the construct. However, not all tests can be used to predict criterion, and therefore, predictive validity may be a large or very small component to establishing evidence for the use and interpretations of test data.

Though both concurrent and predictive validity evaluations are criterion-oriented, they are actually are a form of concurrent validity.[117] The difference within the criterion-related validities is whether the comparison construct is to be measured concurrently or in the future.

Consequence of Use

Consequential validity evaluates to what degree does the test display bias to a specific group or sub-group of test-takers that may have adverse affects.[117] This is especially important if a test is designed to distinguish test-takers by a designated

construct, but actually discriminated based on other traits besides the construct. If a test was designed to distinguish job applicants based on agreeability, but shows bias towards men, this would be an example of a threat to consequential validity.

Summary of Evaluating Construct Validity

Test data lends itself to many different uses and inferences. The purpose of evaluating evidence for construct validity is to determine which of these uses and inferences are valid, and which are not. A test can never be deemed valid, nor can an item, or test data. Only evidence to establish the validity of interpretations and uses of test data can be collected for evaluation. Thus, validity is not an all-or-nothing certification, but an evidence-based process that provides information relevant to many different characteristics of a test and test items.

DEVELOPING CONCEPT INVENTORY ITEMS

Developing concept inventory items needs to be a very thoughtful process. There is no standardized method for the development of concept inventories, which would be impractical because not every concept inventory has the same intended use. However, the developmental process should be guided by some key considerations. The first consideration is the intended use and interpretations of data collected from a concept inventory.[163] Test format, test length and item format will all be dependent on the intended use of the data.[113] The target population should be clearly defined using characteristics that are easy for both test developers and test users to identify[115]. This could be as general as college students in the United States, or more detailed as college students enrolled in first-semester general chemistry in the United States. A second consideration is a clear

definition of the construct or constructs to be measured by the concept inventory, including distinctions from other concepts.[115] The determination of construct-related content is a core concern of construct validity, which is especially critical for concept inventories. Enough evidence needs to be collected during the development process to allow for the evaluation of content validity.[112]

There are also many other important considerations in writing concept inventory items. Restriction of test length based on set administration times limits the number of items, and therefore, number of conceptual misunderstandings that can be targeted as item distracters. Choosing which conceptual misunderstandings are used, or used in duplicate, need to be rationalized for establishing content validity. Evaluating expert and novice response process validity through the item development process, sometimes iteratively, could be necessary, especially when individual item response is used in addition to or to replacement to the total test score.[160] Pilot testing of items and a complete inventory should utilize a standardize administration procedure. This administration procedure includes setting (lecture, lab, recitation, or on-line), allowed time for administration of test, test form (pencil-and-paper, online, etc.), and any instruction to be verbally provided to students. Using the standard administrative procedure during the development of a concept inventory to eventually be used for the completed instrument will help ensure that evidence for establishing lines of construct validity collect during development are still applicable.

ITEM EVALUATION: QUANTITATIVE MEASURES

During the development process and in characterizing the final version of a concept inventory, qualitative measures can be collected and used to make estimates of various test properties. The data generated from a test are not very complex, but how this data are analyzed can vary greatly in the complexity of the analysis and the output from the analysis.

TESTING DATA

Multiple-choice test responses can be represented as polytomous or dichotomous data. Polytomous data are discrete response-level data (e.g., a, b, c, d, or Likert-scale responses) for each item and for each student in a finite student population. In other words, polytomous data provide what option students' choose, but does not give information on if this response is correct. Dichotomous data are discreet item-level data (e.g., correct, incorrect) that can be reduced from polytomous data.

BASIC STATISTICS OF TEST DATA

One way to summarize raw data is through test-score histograms and item-response frequency plots. Dichotomous data can be used to calculate the total score of an exam, which can be plotted as a histogram to visualize test-score distributions. Polytomous data can be used to create an item-response frequency plot, which displays the frequency of students choosing each option for each item.

Another way to summarize data is by calculating different measures of central tendency, which are different ways of calculating the most typical test score in a

distribution of test scores.[117] These include the median, mode, and mean. The median represents the middle point of a test-score distribution. The mode represents the most frequent test score in a distribution. The mean of a test represents the average test score. How much test scores deviate from the mean in a distribution is characterize by variance and standard deviation.

Relationships between test data and other variables can be investigated using correlations and/or covariance. Analysis of covariance can provide information about the direction of association between two distributions of test scores generated by the same sample of students.[117] However, calculating the covariance of two test scores cannot provide information that can be used for clear interpretations. For this, a correlation coefficient should be calculated. The correlation coefficient indicates the direction and strength of linear correlations between two variables. Correlations between test scores, test items, and external criteria can all be used to calculate different types of correlation coefficients.[117]

DIMENSIONALITY OF TEST DATA

The dimensionality of test data evaluates whether the test is measuring one unique construct or multiple constructs.[117] If a test is designed to only measure one construct, it would be assumed to be unidimensional, and the total score would represent a measure of that construct.[113] The essence of structural validity is comparing the theoretical test structure (dimensionality) with the actual test structure.[112, 116] Alternatively, some advanced tests are designed to have unique tests-lets, aimed to measure different, unique constructs within a single test. This would be an example of a multidimensional test.

Using the total score for a multidimensional test is not possible, as it does not allow for interpretation of the separate constructs, so sub-scores need to be used.[113] Factor analysis is one technique that can be used to estimate the dimensionality of data.[117]

CLASSICAL TEST THEORY FOR ANALYZING TEST DATA

One higher-order method for analysis of test data utilizes Classical Test Theory (CTT). CTT includes the classical true score model, which assert that any observed test score (X_o) is a composite of the true score (X_t) and the error associated in measuring the true score (X_e), Equation 1.

$$X_o = X_t + X_e \quad (1)$$

There are two main assumptions of the classical true score model: (1) the error associated in measuring the true score is random and will have an average of zero for a population of examinees, and (2) the correlation between X_t and X_e is equal to zero for any population of examinees.[163] This is to say, that the error score can inflate or deflate the examinee's observed score, but being inherently random, will average to zero over several test administrations. However, the average of all examinee error scores will not be zero and is the basis for calculating a reliability index for a test.[163]

Reliability

Conceptually, the reliability of a test represents the extent to which differences in examinees' observed scores can be attributed to differences in their true score, X_t, rather than differences in their error scores, X_o.[117] Theoretical calculation for reliability can

focus on the correlation between X_o and X_t, (r_{ot2}), or the ratio between the variances of the true score and the error score (s_t^2/s_o^2). The difference between the two ways to calculate the reliability of the test are conceptual, as neither value can actually ever be calculated. This is because the true score can never be determined. However, there are many ways to estimate the reliability of a test based on estimates of the error score, which can be used to estimate the true score. Different methods for estimating reliability use similar ideas of parallel test forms, but collect different types of data and use different assumptions.[117] A parallel test form is a alternate form of the test that would give the same true score for each examinee and the variances in the error scores are equal.[163] Creating a parallel test form, as it turns out, is also very difficult to accomplish.[117] The ability to create parallel items and parallel test forms may not be possible for all types of assessments.

Item-level data can be used to estimate reliability as well, including Cronbach's alpha estimates. The convenience of not needing to create a parallel test form or conduct multiple administrations has made Cronbach's alpha a common estimate of reliability for assessments. There are multiple ways of calculating an alpha estimate, which vary in how data are transformed (raw Cronbach's alpha versus standardized Cronbach's alpha) or how the estimate is calculated (ordinal Cronbach's alpha).[117, 164] The ordinal Cronbach's alpha is most appropriate for tests with a small number of items, as it uses a polychoric correlation instead of a linear correlation. Polychoric correlations are less susceptible to underestimating alpha values when a test has a small set of items. Another benefit of using a Cronbach's alpha estimate is that it uses a less stringent assumption of test tau equivalence, instead of parallel test forms. The main difference is tests that are tau

equivalent simply need to measure the same construct, but do not need to have equal variance of error scores.

Item Difficulty and Item Discrimination

Determining item difficulty (p) and item discrimination (D) are two very important classical characteristics of test items. The item difficulty calculates the proportion of correct responses for each item (number correct responses/number total responses). If the item difficulty is multiplied by 100, it will represent the more common and equally useful percent correct for that item. Item discrimination represents the degree to which an item differentiates students with a high total score from those with a low total score ($D = p_{high} - p_{low}$), where p is the proportion of students within the high-score group or low-score group. Cutoffs for both groups depend on the number of students in a sample and use the total score on the test to differentiate students. Samples with less than 200 students should use the top 50% of the sample as p_{high} and the bottom 50% of the sample for p_{low}.[165] Samples with greater than 200 students should use the top 27% of the sample as p_{high}, and the bottom 27% of the sample as p_{low}.[165]

PROBABILISTIC MODELS FOR ANALYZING TEST DATA

Information gained from using probabilistic models can be used in addition to information from classical test theory estimates. These two theories can be complementary, with probabilistic models adding information and possibly helping to explain CTT estimates. Most importantly, no additional data need to be collected beyond what is used for CTT analysis. Both polytomous and dichotomous data can be utilized by probabilistic models, depending on the type of analysis.

Using students raw test scores as a measure of a construct assumes the spacing within the scale is invariant. In other words, the difference in student ability of two students who scored 45% and 55% on a test is the same as two students who scored 85% and 95%. However, it might be much more difficult to move 10 percentage points up the scale at the top of the scale, as opposed to the middle of the scale. Figure 5 provides a visualization of raw score versus a standardized scale.

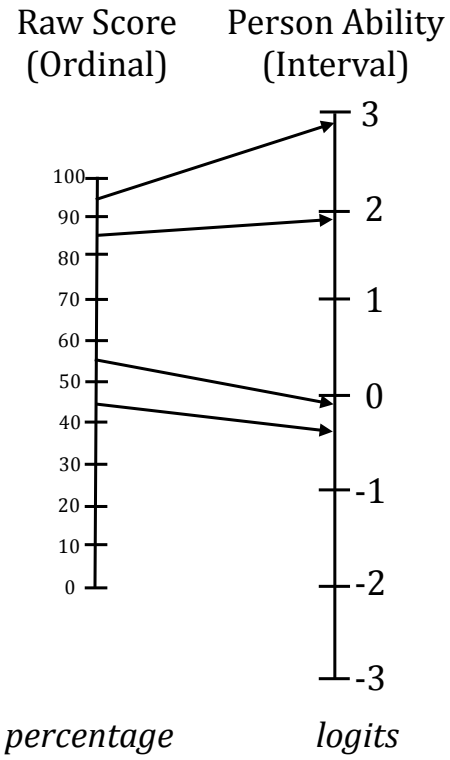


Figure 5. Ordinal raw scores (percentages) transformed to create an interval scale of person ability (logits). This demonstrates that the difference in student ability for higher raw scores is greater than for lower raw scores.

All that can be inferred from the raw scores is the ordering of student ability measured by the test, not the difference in abilities. The use of standard deviations to curve tests has long been a practice to address this issue, but this assumes a simple random sample and a normal distribution. Probabilistic models transform ordinal raw scores into interval log odds (logits), which is a simple logarithmic transformation of odds of success.[166]

The assumptions of probabilistic models are (1) each person is characterized by an ability measure, (2) each item is characterized by a difficulty measure, (3) both person ability measures and item difficulty measures can be represented as a number along an interval scale, (4) the difference between person ability measure and item difficulty measure alone can be used to calculate the probability of observing a specific response for a particular item, (5) the assessment is only measuring one construct (unidimensional), (6) the probability of getting one item correct is independent of the probability of getting another item correct (local independence), and (7) estimates of item difficulty based on testing data are independent of student ability (invariance), making item difficulty estimates less susceptible to differences in student abilities in different samples.[166]

One probabilistic model is the Rasch Model[117] that uses the following assumption about the relationship between item difficulty and person ability: "A person having a greater ability than another person should have the greater probability of solving any item of the type in question, and similarly, one item being more difficult than another means that for any person the probability of solving the second item is the greater one".[167]

Probability Estimates Using Item Difficulty and Person Ability Estimates

Item difficulty estimates and person ability estimates are calculated from ordinal raw data using a probabilistic model (e.g., Rasch Model), yielding interval data that can be placed on the same logit scale.[166] A logit is the contracted form of "Log-Odds Unit".[168] The log-odds for success is can be calculated using equation 2,[168] and has units of logits

$$P_{ni}\left(x_{ni} = 1\right) = \frac{\exp\left[B_n - D_i\right]}{1 + \exp\left[B_n - D_i\right]} \tag{2}$$

The fact that item difficulty and person ability measures have the same units, as any chemist would appreciate, means that these values can be directly compared. Probability estimates of a person answering an item correctly are made based solely on the difference between an individual's ability estimate and the item difficulty estimate. The probability of person *n* scoring a correct answer, $P_{ni}(x_{ni}=1)$, with ability B_n on item *i* with a difficulty D_i can be calculated as shown in equation 3.[169]

$$\ln\left(\frac{\text{Probability of Success}}{\text{Probability of Failure}}\right) \equiv \text{Ability} - \text{Difficulty} \tag{3}$$

Reliability Estimates

The reliability of a test when using probabilistic models depends on how well the item difficulties correlate to person abilities. Specifically, persons with abilities that are much higher or much lower than the average item difficulty will have very low reliability

estimates.[117, 166] Alternatively, students whose ability correlates closely to the average item difficulty will have relatively high reliability estimates. Thus, a test is not given a single reliability estimate, because it will be dependent on both person ability estimates and item difficulty estimates. Instead, two reliability indexes are given, person separation index and item separation index. The item separation index indicates how stable the item difficulty estimate is if given to an equivalent sample of students.[166] This index is sensitive to samples that have poor spread of student abilities to provide data for low- or high-difficulty items.[166] Person separation index indicates the spread of item difficulties and the capability of the test to differentiate among students of different abilities. Put another way, this index is an estimate of reproducibility of person ordering based on ability if given a parallel set of items measuring the same construct.[166] For example, if a test only contains items that have high item difficulty estimates, the test will not be able to differentiate students of low ability. This is because there is a high probability that all of the students will answer the high-difficulty items incorrectly.

SUMMARY

Raw testing data are very simple, and is the basis for all test-analysis and item-analysis methods. How this data are treated, transformed, and analyzed depends on the type of testing data, what interpretation is made with the data, and to what level of detail do researchers want to examine testing data. Classical test theory methods of analysis are still a staple for item and test evaluation and are highly relevant when developing a testing manual for public dissemination. The use of probabilistic models, though not as common, can inform results from CTT analysis and add insight into test and item characteristics and qualities. In both cases, quantitative analysis eventually leads to

questions that can only be answered by qualitative studies. Thus, item and test development needs to be an iterative process, such that multiple qualitative and quantitative studies may need to be conducted to produce a test with desired qualities.

CHAPTER III

METHODOLOGY

INTRODUCTION

The methods used to answer the research questions put forth in Chapter I are both qualitative and quantitative in nature. However, in some cases, both methodologies were used to answer the same question. Concept inventory development is both a linear and cyclic process. Importantly, for qualitative studies informed quantitative studies, and vice versa, reflecting the iterative nature of concept inventory development. The order of the research conducted is outlined in Figure 6. For clarity, this section follows the flow of the research, as it was conducted. As is clearly detailed in this section, information from qualitative and quantitative item evaluation studies were used to identify the need to revise specific items. Thus, criteria and parameters used to evaluate items are detailed in this section, along with the range of revisions that were made to poorly-functioning items.

EXPERT CONTENT
TOPIC SURVEY

To obtain evidence for the content validity of a thermochemistry concept inventory, the most important topics most often taught by chemistry instructors in first-semester general chemistry were identified. This is important for creating concept inventory items that use conceptual misunderstandings as distracters to minimize the threat to validity of construct underrepresentation.

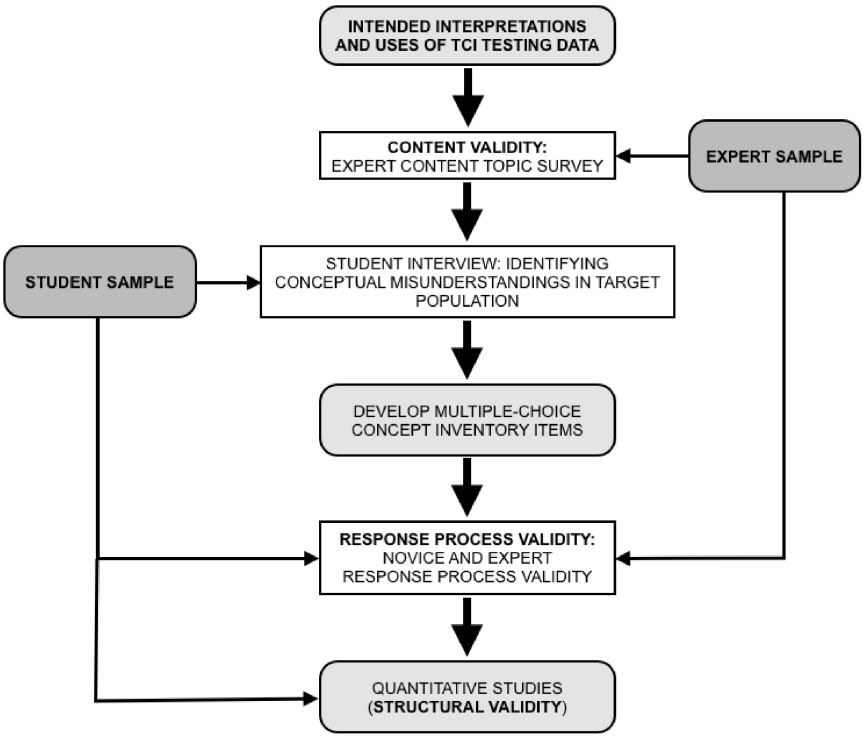


Figure 6. Overview of research and lines of evidence for the validity of uses and interpretations of testing data collected using the Thermochemistry Concept Inventory.

Participant Selection

To increase the generalizability and utility of the TCI, experts teaching at institutions located in different geographical regions having varying sizes and student populations were recruited to participate in the expert content topic survey (described below). A form e-mail was sent to individual professors that were currently teaching in the general chemistry series. This study defines an expert as a faculty member who has traditionally taught in the general chemistry series and who currently teaches first-semester (or equivalent) general chemistry. This qualification was clearly specified in the e-mail, to ensure that only the primary stakeholders of the TCI (e.g., general chemistry instructors)

provided feedback used to establish content boundaries. A link to an online survey was embedded within the e-mail. A follow-up e-mail was sent to encourage participation.

Determination of Thermochemistry Content

Most chemistry courses utilize textbooks, and most general chemistry textbooks cover thermochemistry. As a way to compile a list of topics possibly taught in general chemistry classrooms, topics that are found in the thermochemistry sections of the most widely-used textbooks[15, 170-173] were compiled. Specifically, topic selection utilized chapter section headers and then content covered in those sections. Certain topics, such as enthalpy, have multiple levels of coverage, including operational definitions, sign conventions, and conceptual understanding. For these topics, each level of coverage was classified (e.g., definition-enthalpy, sign-enthalpy, understanding-enthalpy).

Construction of an Online Survey

An online survey was built by the researcher on an internally-housed server to ensure data security. For each topic, a 5-point Likert scale (important = 1, slightly important = 2, neutral = 3, slightly unimportant = 4, unimportant = 5) and a comment box was provided for expert rating and input. In addition, a not covered radio button was the default position for each topic, to avoid any ambiguity between unimportant and not covered, as some topics could be important but not part of the thermochemistry section of general chemistry. The expert comments for each topic were used to help clarify this difference as well, and were important to validate importance ratings. For example, if an expert chose not covered for a topic, they could clarify if it was because it is taught in second-semester general chemistry, or because they do not cover it at any point in the

series. At the end of the survey, a comment box was provided to allow experts to add any topics that they cover in thermochemistry that were not part of survey.

Data Collection and Analysis

Data from the survey were downloaded into Microsoft Excel, including all scores and related comments. Expert importance ratings for each thermochemistry topic was used to create a percent importance ratings, shown in Figure 7. The percentage was determined by summing up the number of experts who rated a topic as either important or slightly important and then dividing this value by the total number of experts that responded. In addition to determining percent importance, expert comments were compiled for each topic and for topics not included in the survey. Specifically, expert scores and percent importance scores were compared to expert comments to check for congruency. Lastly, the response rate for the survey was calculated based on the e-mails sent out and the surveys completed.

$$\% \text{ Importance} = \frac{\overbrace{n_1 + n_2}^{\text{Number Important}}}{\underbrace{n_0 + n_1 + n_2 + n_3 + n_4 + n_5}_{\text{Number of Experts in Study}}} \times 100$$

n_0 = not covered
n_1 = important
n_2 = slightly important
n_3 = neutral
n_4 = slightly unimportant
n_5 = unimportant

Figure 7. Calculation of the percent importance for each thermochemistry topic using expert ratings.

STUDENT INTERVIEWS TO IDENTIFY CONCEPTUAL MISUNDERSTANDINGS IN THERMOCHEMISTRY

To identify what thermochemical conceptual misunderstandings are being used by students in our target population, student interviews were conducted.

Student Participants

Students were recruited from institutions within a reasonable driving distance along the Front Range of Colorado. Interviewing students from different populations at various institutions was important for the generalizability of the student conceptual misunderstandings identified in this study.

The timing of student interviews targeted the opening between instruction of thermochemistry in first-semester general chemistry and instruction of thermodynamics in second-semester general chemistry. Given that concepts taught in thermodynamic build upon some topics in thermochemistry (e.g., enthalpy), conceptual understanding can change upon learning thermodynamics concepts. This study specifically focused on student conceptual understanding of thermochemical concepts, which could be complicated by introducing more complex thermodynamic concepts. Thus, students currently enrolled in either first- or second-semester general chemistry that met the above requirements were eligible participants. The voluntary nature of student recruitment at targeted institutions represented convenience sampling.[114]

Participant Recruitment

A five-minute announcement was made during a lecture by the researcher at a time agreed upon by the researcher and instructor. This announcement gave a brief

overview of the purpose of the research, the interview, and the optional 30-minute free tutoring session following the interview. Emphasis was made by both the instructor and the researcher that participation in this study was optional, anonymous, and that it was not tied to the student's course or their grade in any way. An e-mail sign-up sheet was passed around the class and collected directly by the researcher. Students who signed up for the study were e-mailed by the researcher to set up an interview time. Students were recruited until no new student conceptual misunderstandings were identified.

Interview Design and Protocol

A semi-structured think-aloud interview protocol with probing questions was used for student interviews. This interview protocol has been used in other chemical education studies to explore students' cognitive processes when solving chemistry problems.[79, 174] Students were asked to verbalize their thought process while solving thermochemistry problems developed by the researcher. These open-ended questions focused on the most important topics identified in the expert topic survey. Because obtaining accurate understanding of student conceptions was the goal of this study, probing questions were utilized in this interview protocol.[175, 176] The researcher acknowledges that probing questions have the potential to introduce researcher bias, but that the clarity provided through probing student responses outweighs this risk.[175]

The interview protocol consisted of four parts: (1) overview of interview, Institutional Review Board (IRB) consent, demographics, (2) non-chemistry "warm-up" problem, (3) open-ended interview questions, and (4) recap session. An optional tutoring session was offered to all students at the end of the recap session. The first part of the interview aimed to reduce any student anxiety and build rapport; this has been shown to

be important for cognitive interviews.[176, 177] The IRB consent form was reviewed with students, and the researcher answered any questions the student had about the research or the interview process. Lastly, demographic information was collected, including student gender, major, year in college, and time lapse since instruction of thermochemistry. This took approximately five minutes to complete. During the second part of the of the interview, students were given a non-chemistry warm-up problem to become familiarized with the interview protocol[176] and allowed the researcher to provide feedback on student performance using this protocol before data collection.[178] Students were then given open-ended questions to answer using the think-aloud protocol. Five open-ended interview questions were created for the student interviews to maximize the number of topics that were covered in each student interview. Following the formal interview, students participated in a recap session where each question was reviewed by the researcher. This review allowed for the researcher to address specific conceptual misunderstandings students used during the interview and to obtain additional information about student responses from the think-aloud portion of the interview.

Data Collection

All interviews were video recorded (only the student workspace, with consent) and transcribed verbatim, including the recap session. Immediately following each interview, the researcher compiled field notes about the overall impression of the student and the interview, including any important aspects regarding conceptual understanding that might require specific evaluation in the coding process. All student-generated work was kept as artifacts to aid qualitative analysis.[179] Video files were transferred directly to the researchers password-protected computer immediately after the interview (camera

memory card was erased). Student participants were assigned an alphanumeric identifier to label data, organize data, and to keep student information anonymous.

Coding of Interview Data:
Identification of Student
Conceptual Misunderstandings

Interview transcripts, video files, student work and researcher field notes were imported into the qualitative analysis software package NVivo 8.[180] This software package allowed for all types of digital media to be coded, including sections of video clips and student-generated artifacts (e.g., student written work on open-ended questions). Each interview was coded through an iterative process. Initial coding placed excerpts of interview transcripts and video clips containing errors in bins based on a given thermochemistry topic (e.g., enthalpy, bond dissociation energy, work, etc.). These coded excerpts were then grouped into common themes (e.g., sign convention errors, definition errors, understanding errors, anomalous errors). From these themes, conceptual misunderstandings were identified and checked against those reported in the literature. Interrater reliability (Fleiss kappa) was used to evaluate half of the NVivo coding by five independent coders.

The second phase of analysis utilized interview transcripts, student-generated work and interview video footage to analyze coded excerpts. For each excerpt, the following questions were addressed: (1) Does the excerpt contain an error that is clearly articulated by the student? (2) Is the student using a conceptual misunderstanding to explain the error? (3) What is the conceptual misunderstanding being used? and (4) Does the student consistently use the conceptual misunderstanding throughout the interview?

Student excerpts that represented consistently used conceptual misunderstandings were compiled, grouped by thermochemistry topic, and evaluated for common themes.

Development of Multiple-Choice Items

This research aimed to develop and characterize 10 to 12 concept inventory items utilizing identified student conceptual misunderstandings. The item development process focused on the initial creation of 15 multiple-choice concept inventory items, expecting that some of these items would be combined or not included in the final thermochemistry concept inventory based on the results of qualitative and quantitative item evaluation studies.

Most traditional multiple-choice item development methodologies first construct an item stem, then the correct answer, and then plausible distracters.[113] Because the goal of a concept inventory was to use conceptual misunderstanding as distracters, items were designed around the use of identified student conceptual misunderstandings as distracters. Careful consideration was taken to incorporate conceptual misunderstandings related to topics determined to be the most important by the expert content topics survey. After items had been developed, student and expert feedback was obtained in novice response process validity and expert response process validity studies, respectively.

Evaluation of Novice Response Process Validity of Multiple-Choice Items

Novice response process validity is under the response process line of investigation to establish construct validity, illustrated in Figure 4 and focuses on the evaluation of individual items. Specifically, this process examines how students understand the item stem, any associated figures, and the language and wording of each multiple-choice

response. Student interviews were conducted to collect evidence for the validity of interpretations and use of item-level data, such as using student responses to identify the use of conceptual misunderstandings.

Participants

Students were recruited from institutions within a reasonable driving distance along the Front Range of Colorado. Obtaining the most diverse sample of students available to the researcher to evaluate novice response process validity helped increase the generalizability of concept inventory items. Specifically, student interpretations of item stems, figures, and responses could be influenced by prior knowledge, familiarity with certain types of visual representations, and differences in use of terminology for similar processes. An example of how this can vary at different institutions would be that of textbook usage.[15, 170-173] Most popular general chemistry textbooks[15, 170-173] commonly use either total energy or internal energy in describing the first law of thermodynamics. The textbook an instructor or department uses might affect how students interpret either term. If a student is unfamiliar with the term internal energy, he/she may not choose a response using this term, simply because they do not know what this term means. This can be problematic, as it would bias an item to students who are taught certain terminology.

Interviewing students from different schools and from different educational and demographic backgrounds help facilitate robust novice response process validity evaluations.[112, 117] This is the reason for interviewing students from multiple institutions, as detailed in Chapter IV.

Participant Recruitment

Participant recruitment protocols for novice response process validity interviews mirrored those used for the identification of conceptual misunderstandings student interviews.

Interview Design and Protocol

The purpose of novice response process validity studies is to probe students' response process, including interpretations of items and use of conceptual understandings and conceptual misunderstandings. In addition, these interviews aim to identify any distracters that do not seem plausible to students, or that do not seem relevant to answering the item. Thus, the interview protocol needed to simulate a testing environment, such that students use authentic test-taking mentality. Therefore, a retrospective semi-structured think-aloud protocol with probing questions was used. A retrospective interview, occurring directly after students answer test items, are most appropriate for evaluation of test items in a test-taking environment.[181] The students were provided a stack of ordered items on separate pieces of paper and instructed to answer these questions. Students were told that it was more important to thoughtfully answer each item, rather than completing all items. To ensure equal item coverage in interviews, item order was changed for each interview, and for each institution. Students were then asked to explain their response process in answering each item. This included interpretation of item stems, item figures, and item responses. Probing questions focused on identifying what conceptions students were using to choose their answer, or to eliminate responses. Plausibility and independence of item responses were also evaluated, as both increase item response process validity.[113, 181, 182]

Following the retrospective interview session, student answers were reviewed by the researcher with the student to address any conceptual misunderstandings that students used during the interview. Clarity on what conceptions students used was also sought, as necessary, during this recap session. The researcher also used this session to address any additional questions or concerns students had on the items. A 30-minute free voluntary tutoring session was offered to all participants.

Data Collection

Think-aloud and recap session portions of interviews were recorded using a video camera, recording audio and video of the student's workspace, with consent. Excerpts from interview and recap sessions were transcribed by the researcher. Relevant excerpts detailing student difficulties with items were documented. In addition, the reasoning for student responses were examined for concurrence with conceptions used to construct item responses, both correct and incorrect. Extensive field notes were taken immediately after interview completion, which was the basis for what sections of the interview would be coded later. All student-generated artifacts were collected for analysis. Video files were transferred directly to the researcher's password-protect computer immediately after the interview, and then the camera memory card was erased. Student participants were assigned an alphanumeric identifier used to label and organize data and keep student information anonymous.

Coding of Interview Data

Each test item had an individual data set with excerpts from student interviews specifically in reference to that item. Data were coded and analyzed immediately after each interview, such that if themes or trends emerged for problematic items,

modifications could be made. Modified items were reevaluated in additional novice response process validity interviews, in an iterative manner. This process continued until no more major problems were identified.

EXPERT RESPONSE PROCESS VALIDITY STUDIES

Expert response process validity studies collected evidence on agreement of the correct answer to the multiple-choice items, and whether items portray content correctly.

Expert Participants

This study defined an expert as college-level instructors who have recently taught or who actively teach courses in the general chemistry series. While instructors can teach at institutions of any level, diversity in institution classification was desired.

Expert Participant Recruitment

Expert recruitment targeted the same pool of experts used for the thermochemistry content topic survey. The same pool of experts who ranked topics based on importance were able to evaluate concept inventory items based on these topics. However, this pool was expanded to include more experts to increase institutional representation and number of experts from similar institutions. Experts were recruited using an e-mail form that clearly explained the desired participant qualification based on the study's definition of an expert. A follow-up e-mail was sent to maximize the survey response rate.

Expert Response Process Validity Survey Design

The Qualtrics®[183] web-based research suite was used for survey development, distribution, and response collection. An advantage to using a sophisticated survey client, such as Qualtrics, is the ability to create a professional-grade survey with many integrated

features. For example, the IRB consent form was included at the beginning to the survey and needed to be signed before expert participants could take the survey. This program also anonymously tracked who had not responded to invitations to take the survey and automatically sent reminder e-mails at predetermined times with custom messages by the researcher.

The expert response process validity survey had four main sections: (1) IRB consent, (2) demographic information and rating of importance of thermochemistry topics, (3) determination of topics significantly covered in thermochemistry section, and (4) TCI items. As mentioned above, those who did not sign the IRB by clicking an agree button were not allowed to take the survey and could not be included in the data set. All expert identities were anonymous; therefore, non-identifying demographic information was collected at the beginning of the survey. This demographic information included years teaching general chemistry, lecture class size, and textbook used in their general chemistry course. To compare with findings of the expert content topic survey, experts participating in this survey were asked to rate thermochemistry topics as significantly covered and not significantly covered, along with any topics not listed. This was used to compare results for similar samples for consistency. The remainder of the survey had one item per page, with radio buttons to select one correct response. A comment box was provided below each item allowing participants to elaborate on their answer choice, address content or wording issues, or provide general feedback on the item.

Data Collection

All data collected were reported with randomly-generated alphanumeric identifiers assigned by the Qualtrics survey platform. Data were downloaded from the Qualtrics

website as a comma-separated values (.csv) spreadsheet and analyzed in an item-by-item fashion.

Data Analysis

Expert selection of the correct answer for each item in the thermochemistry concept inventory was compared with what researchers agreed was the correct answer. Items deemed problematic, based on expert feedback, were used to create a follow-up survey to obtain additional feedback as to how to address issues and concerns brought up by experts. All comments and suggestions were used to create the final versions of the TCI items. The demographic data were compiled to report the variation in experts' institutions, teaching experience, and textbooks used. Finally, the response rate for the survey was calculated, based on the e-mails sent out and the surveys completed.

PILOT STUDIES USING THE THERMOCHEMISTRY CONCEPT INVENTORY

Once items for the TCI were developed and evaluated using expert face and novice response process validity studies, quantitative testing data were collected through beta and pilot testing. Items were characterized using both CTT and Rasch model analysis, and evaluated using methodology and parameters discussed in the following sections.

Student Participants

The student samples used for beta and pilot studies needed to be large enough to allow for use of certain statistical analysis techniques. Rasch model analysis required approximately 10 observations per multiple-choice item response. One way to reach this target is obtaining approximately 10 times the number of participants as items in the TCI

and evaluate each category for a minimum of a 10% item option response frequency. Therefore, pilot testing focused on institutions where a sample of greater than 100 students could be given the TCI, accounting for attrition due to IRB consent, incomplete test forms, and variable attendance.

Administration Guidelines for Pilot Testing

Meetings with instructors teaching first- and second-semester general chemistry courses occurred before the start of the semester targeted for collection of pilot study data. As with student interviews, the pilot testing needed to occur after students had instruction of thermochemistry in first-semester general chemistry, and before they had instruction of thermodynamics in second-semester general chemistry.

Secondly, whether the TCI was administered in lecture or in laboratory/recitation was decided. If the administration occurred in laboratory or recitation, teaching assistants (TAs) administered the TCI. This can introduce testing error and bias, given that the test administration could vary, including verbal instructions, testing atmosphere, and testing time. To minimize possible TA bias, efforts were made to obtain TA buy-in to the research. One way this was accomplished was for TAs to take the TCI the week before the administration, provide a detailed answer key, and encourage questions about the administration or purpose of the TCI.

The format for the TCI was a paper-and-pencil test form with an additional scantron for data collection. Students were asked to place identifiers on both and to mark answers on both the scantron and TCI test form to clarify mis-markings on scantrons. Administration time was 30 minutes, including the time it took to pass out the assess-

ment. Due to the variability in how long students take to answer all TCI items, the administration was best suited to be given during the last 30 minutes of class.

The announcement made by the instructor or the researcher to test administration conditions were standardized and had IRB approval (see Appendix A). This statement was read verbatim by the instructor or the researcher administering the assessment to minimize testing error due to administration conditions.

All test forms were reviewed immediately after administration for IRB student consent, signified by signing of the IRB consent form attached to the TCI test form. Any tests without IRB consent were destroyed immediately along with associated scantrons. This protected against non-consenting student data to be used in data analysis. If the TCI was administered by TAs, the researcher assigned a TA-specific code on each student scantron to check for TA bias during statistical analysis.

THERMOCHEMISTRY CONCEPT INVENTORY TEST AND ITEM ANALYSIS METHODOLOGY

Beta and pilot testing the TCI were used to collect student data that were analyzed at the test-level and at the item-level, using both CTT and the Rasch model analysis.

Classical Test Theory Analysis

Sample differences. When data were collected from the same institution but in classes of different instructors, an analysis of variance (ANOVA) was used to determine if all class data could be pooled into one data set. An ANOVA analysis of the means of student TCI scores was conducted if there were three or more classes sampled at one institution, or a *t*-test for two classes. A Tukey post-hoc test was used to identify significant differences between classes (alpha < 0.05). This analysis was conducted using

Predictive Analytics SoftWare (PASW) Statistics 18.0 statistical software program.[184] If significant differences were found for a class data set, then this data were analyzed separately. Alternatively, if no significant differences were found between class data sets, a single data set was compiled from class data for analysis.

Test analysis. Student polytomous testing data were imported into Microsoft Excel. All student information was replaced with an alphanumeric code by the researcher. For test analysis, only dichotomous data were used, so polytomous data were transformed into dichotomous data by the researcher in Excel.

Student total TCI score and how many items they answered correctly were calculated to allow calculations of the mean and standard deviation of the mean. These calculations were completed in Excel.

Item analysis. Item analysis provided very useful information for item development and refinement. The overall characteristics of the test originate from test items, including difficulty (p) and discrimination (D) estimates.[162] Item difficulty estimates provide the first level of detail by giving the proportion of students in the sample who answered the item correctly and an estimate of the performance of the entire target population.[113] For a concept inventory, if a prevalent conceptual misunderstanding is used as a incorrect response, then the item difficulty would be expected to be low. In addition, if one item has multiple prevalent conceptual misunderstandings as distracters, then the item difficulty could be very low (small proportion of sample chooses correct answer). Thus, a low item difficulty is not a huge concern, if the distracters students choose are an accurate refection of their conceptual misunderstandings. Conversely, if an item has a very high item difficulty estimate (high proportion of sample chooses correct answer),

then this may be a concern. If no students are choosing distracters representing conceptual misunderstandings, two things might be occurring: Either, the conceptual misunderstandings are not present in the sample and possibly the population, or students have conceptual misunderstandings, but are not choosing responses representing these conceptual misunderstandings. The latter is one of the concerns addressed in novice response process validity studies and were detected during early pilot testing and item analysis. Thus, no cut-off item difficulty estimates were used for item analysis, but items with high or low item difficulty were assessed in novice response process validity studies to verify student response process and rational.

Item discrimination. Item discrimination provides more detail about how an item can discriminate students in a sample based on their performance on the test. A simple way of doing this is by sorting students by total score, then separating students into high- and low-score groups. The cut-offs for data collected during pilot testing depended on how many students were in the sample. If there were less than 200 students, then the top 50% of students, based on their TCI total score, was the high-score group and the bottom 50% was the low-score group.[165] If there was more than 200 students in the sample, then this changed to top and bottom 27%.[165] Items with low item difficulty estimates (low proportion of sample choosing the correct answer) are bound to have low discrimination estimates. This is because the maximum discrimination will occur for items that are passed by 50% of the sample.[162] Items that are at either extreme end of item difficulty estimates are not as reproducible and are less informative and should be analyzed accordingly.[162]

Item response frequency. Item response frequency, the number of students choosing each item response, are important quality of concept inventory items, especially when the number of items are limited. One of the major concerns of content validity is sampling the entire content domain with test items. Because the number of conceptual misunderstandings that could have been used as distracters in the TCI were limited by the number of TCI items, deciding which conceptual misunderstandings were used was based on two main criteria. The first was the importance of the conceptual misunderstanding, based on expert ratings of topics taught in thermochemistry. The second was the relevance of the conceptual misunderstanding to the target population. The later was assessed, in part, by evaluating item response frequencies. Items having response frequencies lower than 10%, for example, were not considered attractive to students in the sample. Perhaps, other options were more attractive or that the unattractive response appeared implausible or incorrect. A threat to novice response process validity occurs if students had a conceptual misunderstanding, but something about the construction of the item (e.g., wording, response ordering, or obvious correct answer) compelled students to choose another response. As with item difficulty and discrimination, items response frequency was critical for early beta and pilot testing; therefore, follow-up qualitative studies were used to assess any concerns that arose. Thus, item responses with less than 10% response frequency were of particular interest for qualitative analysis.

Rasch Analysis

Rasch analysis provided insight into estimates from CTT analysis. Item difficulty estimates calculated using the Rasch model were compared with CTT difficulty estimates, but Rasch difficulties have the advantage of associated item fit statistics and being

on an interval scale. In addition, Rasch analysis included person ability estimates and associated fit statistics, which provided unique information compared with just using CTT total score as a measure of a person's ability. Most importantly, Rasch analysis allowed for the comparison of item difficulty and person ability estimates to evaluate if test items were too difficult or too easy for the ability range found in a sample from the target population.[166] Thus, the ability of a test and test items to provide accurate information about a given construct (e.g., thermochemical conceptual understanding) was evaluated using Rasch analysis.[166]

Data preparation. Polytomous data used from CTT analysis were imported into Winsteps 3.70.1.1 (www.winsteps.com) and converted to a .win file (Winsteps-compatible). All Rasch analysis of TCI testing data utilized the Winsteps software program.

Rasch model selection. There are many Rasch models that share the general form shown in Equation 2.3.[169] For TCI multiple-choice data, two specific models are relevant: the original dichotomous model put forth by Rasch[167] and the Partial Credit Model (PCM) put forth by Masters.[185] When dichotomous data were used for Rasch analysis, then the dichotomous model was used. However, if polytomous data were used, the PCM Rasch model was used. The PCM model uses parameters for the difficulty for each response of a multiple-choice item. The assumption that not all responses have the same difficulty yield three important traits of PCM analysis.[185] First, option probability curves have unique shapes based on individual item responses of an item, such that the difficulty of an item was analyzed at the response level.[185] This was useful to determine what responses were informative and which were not based on difficulty estimates of each response. This information helped fine-tune the difficulty and discrimination of an item. Secondly,

obtaining information at the response level provided information about sources responsible for misfit to the Rasch model.[185] This allowed for targeted revision of misfitting items at the item-response level. In other words, an item might not fit the Rasch model as a result of a problem with one the response categories, and this might imply that the category needed to be modified or replaced to increase item fit. As discussed below, item fit to the Rasch model was important and informative. Thirdly, TCI items can differ in the number of responses for each item.[166] This is important for analysis of TCI data, since TCI items did not all have the same number of responses.

Before the PCM model was used, TCI data were analyzed to make sure there was at least 10 observations per category for each TCI item response.[186] This decreased the error associated with making difficulty estimates for item responses. Items that did not have 10 observations per category were flagged and interpretations made using Rasch analysis were made with caution.

Check for unidimensionality and local independence. Verification that using the Rasch model was appropriate for data sets was the first step in the evaluation process, specifically, verifying the assumptions of unidimensionality of the construct and local independence of items. Uniquely, the Rasch model requires the researcher to evaluate data based on its fit with the model.[166] This is in contrast to item response theory, where the parameters within the model can be adjusted such that the model fits the data.[166] This research only used the Rasch model for test and item analysis, so it was critical that the data met the assumptions of the Rasch model. If the data did not meet the assumptions of the Rasch model, other probabilistic models would have been considered (e.g., 1-parameter item response theory model).

The basic questions in the analysis of test data dimensionality are (1) What is the difference between the observed outcome and the outcome predicted by the Rasch model (response residual)? (2) If residuals explained by the Rasch model are removed, is there any commonality among the remaining residuals? and (3) What underlying traits might explain these residuals? Commonly, factor analysis, and specifically, confirmatory factor analysis, is used to verify unidimensionality of testing data.[166] This analysis uses raw non-linear ordinal data, which can show a heavy dependence on the sample used to collect testing data.[187] In addition, there are no fit statistics generally assigned to factor loadings, which can limit the interpretability of this analysis.[166] For these reasons, confirmatory factor analysis was not used in this study. Instead principle component analysis (PCA) of the standardized residuals (information not explained by the Rasch model) was used to address the questions presented above. Rasch analysis included evaluation of item fit and person fit. Both items and persons that exceeded acceptable misfit parameters were removed from the data set.[188] Then PCA was run in Winsteps, which analyzed the correlated variance of the standardized residuals of items in the TCI not explained by the Rasch model. Any item with loading on a secondary contrast greater than ±0.4 was flagged[189] if the associate eigen value of the contrast was larger that 2.00. To assess the threat to unidimensionality by these flagged items, further analysis determined (1) What is the magnitude of the difference between the primary dimension and the secondary dimension? (2) How many students, and particularly which students, are being impacted by this secondary dimension? and (3) Is there enough evidence to support either addressing problematic items or distinguishing the secondary factor as a separate subscale?[190] Going back to the raw ordinal testing data and running Rasch analysis only on items

flagged by PCA analysis might produce secondary item difficulty and person ability estimates. Plotting the secondary person ability estimates against those from the original analysis including all TCI items would answer who was most impacted and how much.[191]

To verify that the local independence assumption of the Rasch model was met, inter-item correlations were evaluated using the same PCA analysis described above. Items that displayed a strong correlation ($R > 0.5$) among standardized residuals were flagged; especially those that did not have obvious content similarities.

Fit statistics. The residuals used for PCA analysis were also used to determine how well testing data fit the Rasch model. As the Rasch model provided estimates of two parameters, item difficulty and person ability, associated fit statistics were evaluated for each; specifically, how well did testing data for each item of the TCI fit the Rasch model, and how well did individuals taking the TCI fit the Rasch model? The two fit statistics to be used for this analysis were outfit and infit. Analysis of data fitting to the Rasch model focused on identifying observations that were outliers to the data set and on unexpected response patterns in observations. Identifying outliers, using outfit statistics, was the first step in the analysis of TCI data, followed by identification of unexpected response patterns, using infit statistics.

Both outfit and infit are chi-squared statistics and are reported with associated Z-statistics to assess statistical significance.[166] Outfit is calculated by summing the square of standardized residuals for either all responses by an individual or all responses to an item, and taking the average.[189] When the average, a chi-squared statistic, is divided by the degrees of freedom, the result is a mean-square statistic (MNSQ), which is reported by Winsteps.[189] MNSQ values have an expect value of 1.00, and have a range from 0 to

infinity. However, MNSQ values 1.00 ± 0.5 are generally acceptable, and 1.00 ± 0.3 are used as more stringent evaluation criteria.[166, 189] Every MNSQ value has an associated Z-standardized statistic (ZSTD) to assess statistical significance. MNSQ with ZSTD values greater than 2.00 represent $p > 0.05$.[189] However, it should be noted, that for large data sets, ZSTD values increase due to increased statistical power and should be evaluated only after observations displaying MNSQ misfit have been identified.[189] Conceptually, outfit is sensitive to outliers, which is good for identifying outlying observations, but outfit is also easily skewed by these observations. Issues that are identified by poor outfit are generally easy to diagnose and easy to address; thus, outfit is normally the first fit statistic evaluated. For example, high outfit (underfitting) MNSQs (> 1.5), can result from a few low-ability students answering high-difficulty items correctly. One way students can correctly answer an item above their ability is by guessing the correct response. These observations can be removed from the data set, improving outfit for those items.

The infit statistic was created to reduce sensitivity to outliers displayed by the outfit statistic.[189] The infit statistic is calculated the same as outfit, but is weighted by the statistical information (model variance) of observations.[189] This model variance is larger for observations where the Rasch model should provide an accurate prediction (e.g., when a student's ability is close to an item's difficulty) and smaller for extreme observations (e.g., when a student's ability is much less than item difficulty).[166] This makes infit sensitive to inlier observations that display an unexpected response pattern. Observations with misfitting infit statistics are more complex and more difficult to diagnose. High infit MNSQs (> 1.5) can result from items that are well-targeted to student ability, but poorly predict observed outcomes.[189] Determining why an item is misbehaving is much more

difficult, because it may involve some component of the item construction or some part of a student's response process. These generally cannot be answered solely by Rasch analysis.[166]

Item analysis. Item analysis included evaluating statistics for item fit, item difficulty, as well as item separation index and option probability curves for response analysis. Polytomous data sets were used if the 10 observations per item response criteria was met.

Item fit statistics were evaluated in the following order: (1) outfit MNSQ ($0.7 < MNSQ < 1.3$), (2) evaluated associated ZSTD for outfit MNSQ statistic ($ZTSD > 2.0$), (3) infit MNSQ ($0.7 < MNSQ < 1.3$), and (4) evaluated associated ZSTD for infit MNSQ statistic ($ZTSD > 2.0$). Items that showed infit or outfit misfit that were significant were flagged and addressed after student data had been evaluated, and poorly-fitting students had been removed from the data set.

Item difficulty measure, given in logits, were a key characteristic for analysis. Because the transformation of raw ordinal data into interval data for Rasch analysis did not change the ordinal-information (ranking of item difficulty), item difficulty measure alone did not provide much new information. In other words, items that were most difficult based on CTT analysis (item difficulty) also had the highest difficulty measure on the logit interval scale. However, the distance between items on this logit scale did provide key information to compare item difficulties, which was new and unique information provided by Rasch analysis. In addition, unlike raw ordinal data, Rasch item difficulty measures had an associate standard error estimate associated with each item difficulty measure. The standard error estimate is an estimate of precision for the

difficulty measure, where the fit statistics (e.g., infit) are an estimate of accuracy. Standard errors of item difficulty measures were not heavily relied upon in the analysis of testing data in this study for two reasons. First, the precision of Rasch estimates were generally very reproducible, given the testing data fit the Rasch model (acceptable infit and outfit statistics).[189, 192] Secondly, this study was not looking to make any criterion-referenced decisions, such as cut-off TCI scores based on item difficulty measures, so statistical differences between difficulty measures were not a key concern.[192] What was of greater concern, for test development and applicability of TCI testing data, was analyzing the range and coverage of TCI item difficulties. Specifically, item analysis focused on the following questions: Are items all the same difficulty? If item difficulties differ, what is the spread, the spacing, and are there any large gaps of coverage? Are there items that are extremely difficult or extremely easy? Do item difficulty measures have significant overlap with student ability measures, which are also on the same logit scale?

A way to answer these questions was to use a Wright map, which places all items and all students on the same logit scale for comparison and analysis. The Winsteps program was used to create Wright maps for data sets, as shown in Figure 8. The mean of item difficulty measures is centered at 0 logits,[166] which was compared with the mean of student ability measures for how well item difficulty overlaps with student ability. Sufficient overlap was based on an evaluation of the mean of item difficulty of TCI items and the mean of student ability from the target population taking these items: less than one logit separation were deemed sufficient overlap. If this did not occur, flagging items with extreme difficulties and persons with extreme abilities was the basis for future analysis.

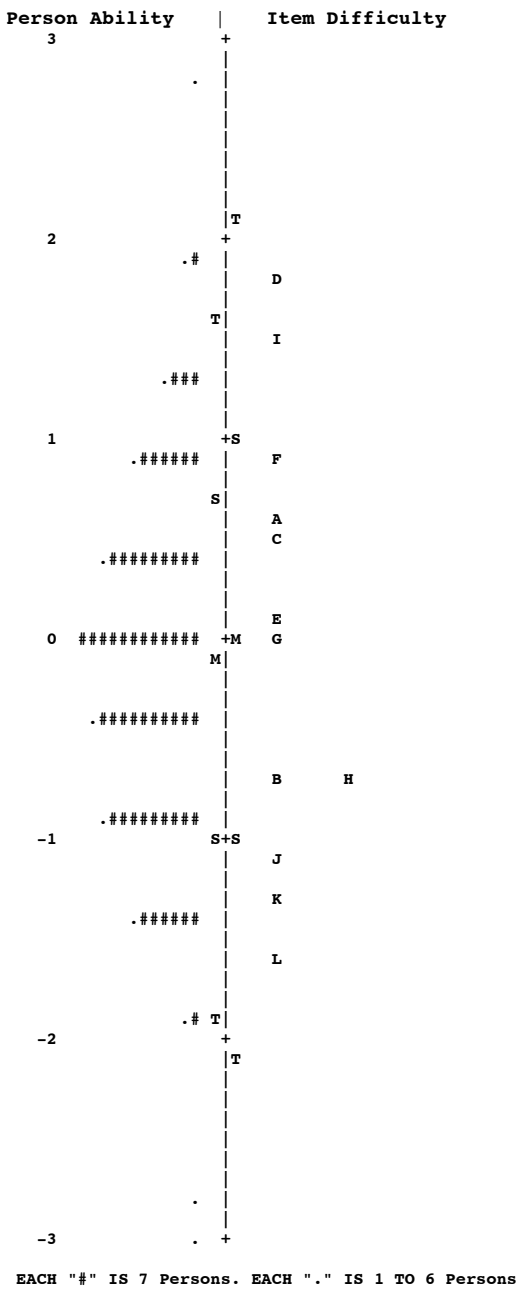


Figure 8. Sample Wright map output from Winsteps. This shows person ability distribution (left side) and item difficulty measure distribution and spacing (right side). Items labels are given at the far right (A-L), and the mean (M), first standard deviation from mean (S) and second (T) are shown near the centerline. Logit scale is given on the far left of the map.

The item separation (reliability) index, an estimate of the reliability of item difficulty estimates, is unique to the Rasch model and does not have a traditional (CTT) reliability equivalent.[189] If a test is given to equivalent samples of students, the item separation index reflects the stability of item difficulty estimates.[166] This measure is sensitive to samples of students who do not have a large enough spread of abilities to sample items of high and low difficulty. The item separation index can be increased by increasing the sample size of pilot studies and is generally not increased by increasing the number of test items.[189] For TCI item analysis, item separation indices larger than 3.0 would indicate sufficient item separation.[189] If item separation was less than 3.0, a larger sample size would have been collected.

Option probability curves (OPCs), along with associated item category misfit order, were used to analyze responses (categories) of an individual item. Winsteps provided OPCs for each item, which allowed for visualization of the probability of choosing each response based on student ability. An example OPC is shown in Figure 9. Students of the highest ability should display the highest probability of choosing the correct response. If this was not the case when viewing OPC plots, then further investigation using the item category misfit order was conducted. Outfit statistics for each item response was used to see how the data for the response fit the Rasch model. Outfit values were analyzed exactly the same as for items and persons, where overfitting (MNSQ < 0.7) and underfitting (MNSQ > 1.3) item categories was flagged for further analysis.[189]

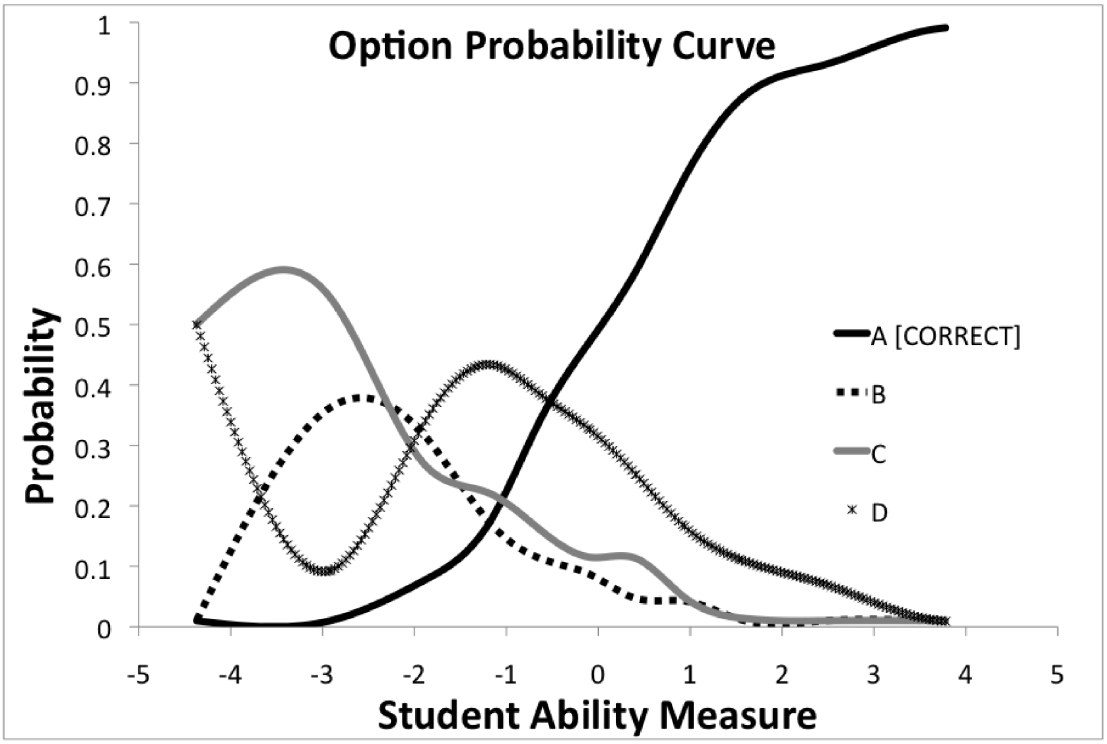


Figure 9. Probability of a category (e.g., item response) being chosen plotted against student ability measure. The probability of choosing the correct response increase to 1 as student ability increases. Response C is most attractive to students of the lowest ability, and becomes less attractive as student ability increases.

Student analysis. Each student who took the TCI had a student ability measure calculated using the Winstep software program. This ability measure was estimated using their total score on the TCI. For example, a student answering 50% of the items correct on the TCI would have a lower person ability measure than a student answering 70% of the TCI items correctly.

Person fit statistics for person ability estimates were analyzed in the exact same way as for item difficulty estimates. Students who displayed either overfit or underfit were flagged and in some cases were removed from the data set.[189] This generally

increased item fit statistics. If TCI pilot studies were administered by TAs in lab, TA bias was evaluated based on patterns of student misfit. Because TA information was embedded within the student identifier (TA code), this was easily evaluated by looking at all students displaying misfit. If a high proportion of students displaying misfit were from a particular TA, all student data from that TA, even persons who have acceptable fit, would have been removed from the data set. This was based on the assumption that if the TA did not take the administration of the TCI seriously, students may have not as well and might have answered in patterns that would not fit the Rasch model. This would be most apparent when evaluating the infit statistics.

The person separation (reliability) index indicates the ability of the instrument to distinguish high and low ability students. The person separation index is equivalent to traditional test reliability, estimated by Cronbach's alpha values.[189] However, Rasch person separation indexes generally do not use persons with extreme scores, which inflate this estimate. Traditional reliability estimates do use persons with extreme scores and generally have a higher reliability estimate when compared to the Rasch person separation index.[189] Just as with traditional reliability measures, the person separation index increases with an increase in the number of items.[166] The person separation index was used as one estimate of the TCI testing data reliability with respect to students.

SUMMARY

This section detailed the methodology used for creating the TCI. The linear and iterative nature of this process was necessary to collect evidence for the validity of the intended uses and interpretations of TCI testing data. For example, if the TCI is to be used as a diagnostic assessment for identifying student conceptual misunderstandings,

novice response process validity studies are essential for collecting this evidence. Specifically, verifying that student responses to item distracters can accurately and reproducibly be used to identify a specific student conceptual misunderstandings. The results of both qualitative and quantitative studies collected during all parts of this research were used to create a testing manual (e.g., peer-reviewed publication; Chapters IV and V). Figure 10 summarizes the research along with uses for the TCI. The results of this project is the publication of a psychometric evaluation of the TCI and creation of the final version of the TCI for use by chemical education researchers and educators.

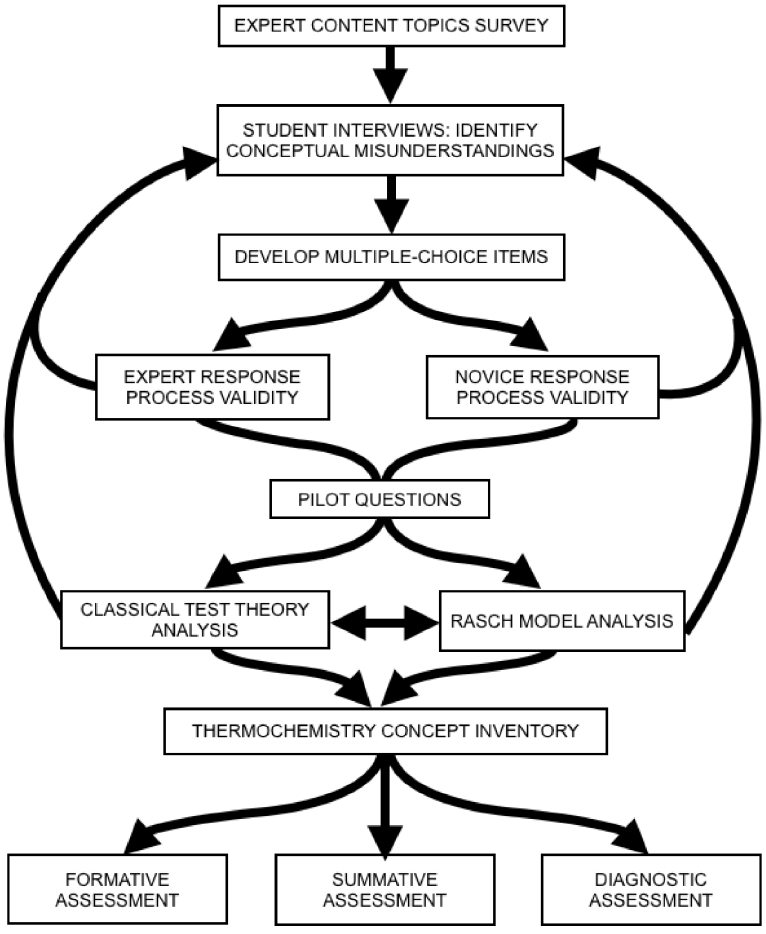


Figure 10. Summary of research used for development of the Thermochemistry Concept Inventory and future uses of this inventory by chemical education researchers.

CHAPTER IV

METHODOLOGY FOR THE DESIGN, DEVELOPMENT AND QUALITATIVE EVALUATION OF THERMOCHEMISTRY CONCEPT INVENTORY ITEMS

ABSTRACT

Assessment instruments developed specifically for chemistry classrooms have increased in number over the last decade. In the design, development, and evaluation of these instruments, the chemical education community has adopted many of the practices and standards used by the greater assessment communities. Methodologies for creating new assessment instruments now include collecting broad evidence for the validity of the uses and interpretations of data derived from an assessment instrument. The focus of this study was the design, development, and qualitative evaluation of concept inventory items for the Thermochemistry Concept Inventory (TCI). Qualitative research studies were used to obtain feedback from the primary stakeholders of the TCI. Evidence for content and response process validity are provided and used as arguments against the two threats to validity: construct underrepresentation and construct-irrelevant variance. In addition, a determination of the most important thermochemical topics taught in general chemistry classrooms is derived using feedback from general chemistry instructors' responses to an online survey. Semi-structured think-aloud interviews were used to identify alternative conceptions used by students answering open-ended questions. Qualitative data were used

to develop a series of multiple-choice items that were then evaluated by students in retrospective think-loud interviews. These student interviews, along with additional interviews with general chemistry instructors, provide evidence for the response process validity of the items.

INTRODUCTION

New Era of Assessment Development

Assessment of student learning in higher education plays a critical role in the evaluation of instruction, course-related learning goals, and departmental and institutional accreditation. The National Research Council's (NRC) 2001 report, *Knowing What Students Know: The Science and Design of Educational Assessment,*[64] using findings from cognitive science and measurement theory, put forth many recommendations for the development of educational assessment instruments. A central finding was a need for assessment to move beyond algorithmic or procedural knowledge[64] and probe for student cognition, as this is at the core of student learning. The report defines cognition as "the mental process and content of thought involved in attention, perception, memory, reasoning, problem solving, and communication".[64] An example of a cognitive process that could be targeted by assessment is conceptual understanding. This NRC report also recommended that assessment should be designed for practical use in classroom settings, should be aligned with curriculum and instruction to facilitate student learning and that the format of these assessments should match the intended use, including class size, administration time, etc. Along with these NRC recommendations, the chemical education research community has identified a need to develop assessment instruments with reliability and validity in mind.[100, 103, 143, 193, 194] The recent work by Arjoon, Xu, and

Lewis[195] highlight psychometric evidence needed to establish the measurement quality of an assessment instrument.

Establishing the measurement quality of an assessment instrument is critical to support the inferences and interpretations of student scores and uses of assessment data. The *Standards for Educational and Psychological Testing*, hereafter referred to as the *Standards*, published jointly by the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME),[112] refers to this as evidence of validity. Specifically, validity is defined as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests".[112] The *Standards* provide guidelines for the type and amount of evidence required, which depend on the specific interpretations and uses of testing data and the stakes of these interpretations and uses. As will be discussed in this paper, much of this evidence is collected during the development process. Therefore, collection of validity evidence should drive the development of assessment instruments.

Concept Inventories as a Diagnostic for Student Conceptual Understanding

Recently the NRC report on *Discipline-Based Educational Research (DBER)*[103] put forth the status, contributions, and future direction of DBER. This report found that the DBER community, which includes chemistry education research (CER), lacks an understanding of the scope and impact that the use of conceptual misunderstandings have among different groups of students (e.g., race, gender, academic ability, etc.) [pg. 72]. Furthermore, a need for more research on student conceptual change, a cognitive process,

specifically targeted by instruction and learning experience was also detailed. In essence, this report highlights that there is a lack of evidence of the effectiveness for new and innovative instructional techniques aimed to improve student learning outcomes, including those focusing on promoting conceptual change. Ironically, the lack of evidence for the effectiveness of pedagogical innovation is not due to a lack of effort on the part of the CER community. In the last decade, CER has focused on expanding the pedagogical toolbox for chemistry instructors by creating conceptually-based assessment instruments to evaluate new instructional techniques.[95, 195] Many of these instruments are concept inventories, which identify alternative conceptions used by students based on which distracters are chosen.

Concept inventories have been utilized in CER for over 25 years,[137] with a recent renaissance in the last decade.[95, 195] Concept inventories are generally multiple-choice assessments that use identified student conceptual misunderstandings as distracters. The development and use of concept inventories are rooted in the assumption that students are not empty vessels to be filled with knowledge, but each actively constructs their knowledge through meaning-making structures and schema.[65, 67, 124] This constructivist theoretical framework relies heavily on inductive research, many times using in-depth interviews to obtain rich and descriptive data about student learning. Research focusing on student conceptions has been a central focus in chemical education research.[42, 76, 81, 83, 85, 126, 137, 143, 196] The term alternative conception is most closely linked with the constructivist theoretical framework, describing a conception that varies from that which is accepted by the scientific community. Given the complex and sometimes abstract nature of concepts in chemistry, however, student conceptual understanding cannot be described by a single

term such as alternative conception. We propose that conceptual understanding can be described as a continuum, where alternative conceptions are on one end of this continuum and those accepted by the scientific community are on the other. In many cases students may have an incomplete conception, have incorrectly memorized a sign convention, confused two different conceptions, or simply lack enough knowledge to form a coherent conception. Though these forms of conceptual misunderstandings may not be as robust to conceptual change as alternative conceptions, they can be nucleation sites for future alternative conceptions and, therefore, should be targets for identification and instructional intervention. Therefore, we will use the inclusive term conceptual misunderstanding to describe all forms of student knowledge that vary from those which are accepted by the scientific community.

The role of concept inventories as a diagnostic tool for both instructors and students has a key role in the learning process, specifically with regards to conceptual change. David Ausebel famously said, "The most important single factor influencing learning is what the learner already knows. Ascertain this and teach him accordingly".[44] Though this was most likely in reference to correct conceptions, it is just as applicable to student conceptual misunderstandings. Conceptual misunderstandings that are connected or anchored to other conceptions constructed by learners cannot simply be replaced with correct conceptions.[61] This is because the entire knowledge structure around that conceptual misunderstanding must be rearranged and reorganized to accommodate a new, correct conception. Posner[96] argued that for conceptual change to occur, students must first be dissatisfied with their existing conception, understand the new correct conception, associate meaning and context of the new conception in their existing knowledge

framework, and believe that adoption of the new concept will be beneficial. Though this study is grounded in the seminal work of Posner, contemporary work by Chi[120] and diSessa[154] illustrate the complexity of the process of conceptual change.[119] In addition, students have robust rationalization schemes to adapt anomalous data that should produce cognitive conflict and promote conceptual change.[97] Therefore, many students will retain strongly-held conceptual misunderstandings throughout a chemistry course, which can reduce or block the adoption and utilization of accepted conceptions and limit the effectiveness of instruction. Another challenge for students is many of the conceptual misunderstandings found in student populations have also been identified in pre-service secondary,[72, 90] secondary,[79] and post-secondary[86] science instructor populations. Therefore, conceptual misunderstandings used by students might also be used by their instructors.

Thermochemistry Concept Inventory

Thermochemistry, typically taught in the first semester of the general chemistry sequence, has been the focus of numerous research studies[73-75, 78, 80, 82-84, 86, 87, 90, 93-95, 102-105, 142, 197, 198] for many reasons, including (1) it contains many concepts that prove challenging for students to learn, and (2) concepts taught in thermochemistry are the foundation for concepts taught in thermodynamics and physical chemistry. Thermochemistry is also unique in that physics and chemical engineering courses also teach principals of thermochemistry, including thermal physics and thermal transport, respectively.

There are currently several concept inventories available that assess thermochemistry concepts, including Thermal Concepts in Everyday Contexts (TCE),[197, 198] Thermodynamics Diagnostic Instrument (THEDI),[199] Heat and Energy Concepts

Inventory (HECI),[200] and Thermal and Transport Science Concept Inventory (TTCI).[104] The TCE is targeted for secondary school students and both the HECI and TTCI are targeted to engineering students. The THEDI covers topics taught in both thermochemistry and thermodynamics in the general chemistry series, but the majority of the conceptual misunderstandings used as distracters focus on conceptions only taught in thermodynamics, which is a second-semester course in most United States universities. Currently, no thermochemistry-specific concept inventory is available, and no study has concentrated on the target population of general chemistry students in the United States. The goal of this study was to create the TCI, targeting the population of college-level students in general chemistry courses where thermochemistry is taught. The TCI test length is intentionally designed to be short. Targeted administration and testing time will be under 30 minutes, corresponding to approximately 10 to 12 single-tier multiple-choice items. This will make the TCI a practical instrument to use for formative assessment.

The Role of Validity in the Design and Evaluation of Assessment Instruments

A test can never be deemed valid nor said to produce valid data.[112, 117, 201] Rather, validity refers to the specified uses and interpretations of testing data for a specific target population. The *Standards* defines and describes validity by the following passage:

> Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of the test. Validity is, therefore the most fundamental consideration in developing and evaluating tests. The process of validation involves accumulating evidence to provide sound scientific basis for the proposed score interpretations. It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself.[112]

The evidence needed to establish the validity of specified interpretations and uses of test scores will vary, with some types of validity being more critical than others. It is essential that test developers establish the intended uses and interpretations of test scores before the development of an assessment instrument. This allows for test developers to determine what forms of validity are most critical and accumulate evidence for these facets of validity through the developmental process. Therefore, evidence for validity should not be an afterthought in the development of an assessment instrument, but be the driving force in the design and development process. The guidelines put forth by the *Standards* suggest test developers clearly state a set of propositions that support the interpretations and intended uses of assessment data.[112]

A Contemporary Conceptualization of Construct Validity

The traditional view of construct validity put forth in the seminal paper of Cronbach and Meehl[118] was initially expanded upon by Loevinger[116] and most recently by Messick.[201, 202] Cronbach and Meehl's view of construct validity as one of three types of validity (content validity, criterion validity, and construct validity) was expanded by Loevinger into three components of construct validity (substantive, structural, and external components) and further expanded by Messick to include six components (content, substantive, structural, generalizability, external, and consequential). In the *Standards*, a contemporary conceptualization of validity uses construct validity as the over-arching validity trait, which all other validities could be used to establish. This contemporary view of construct validity includes five sources of validity evidence: test

content, response process, internal structure, association with other variables, and consequence of use, as shown in Figure 11.

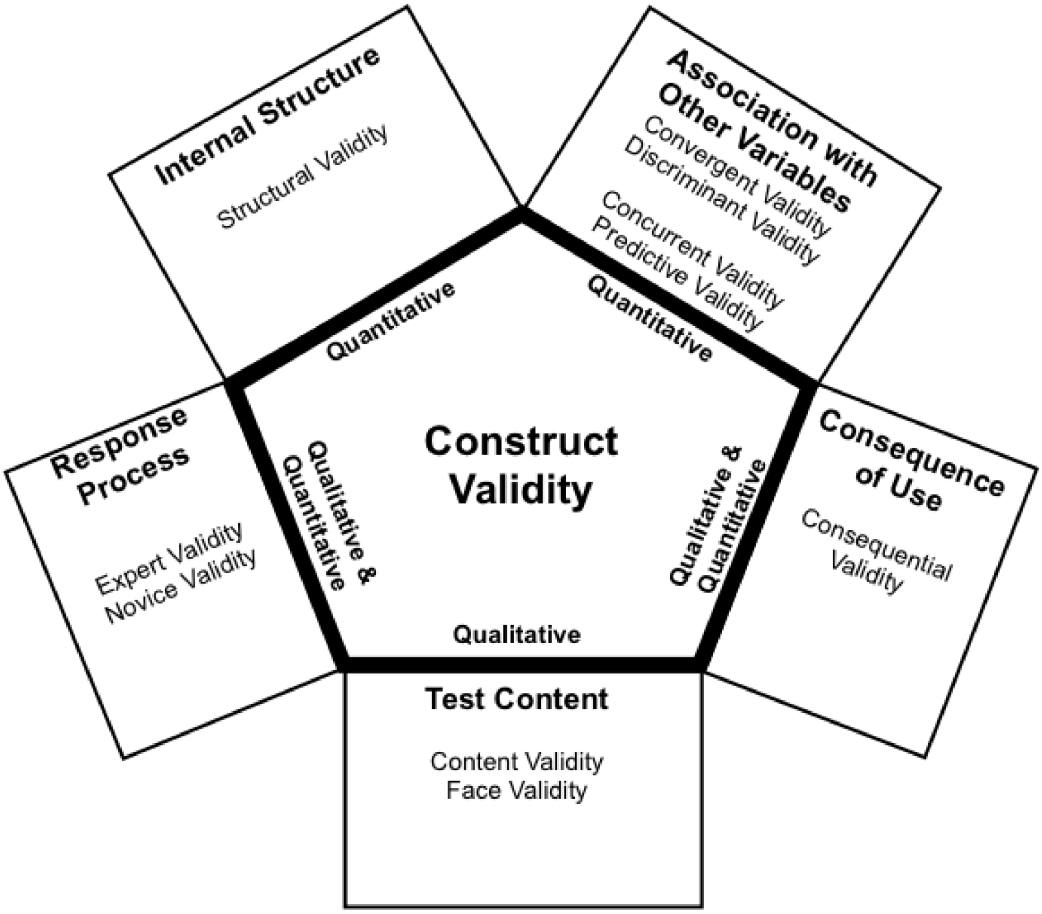


Figure 11. Evidence for construct validity provided from five unique evidence sources.

Both qualitative and quantitative data are sources for evidence of construct validity. Predictive validity, for example, generally requires quantitative data where response process validity could be evaluated using both qualitative and quantitative data. At what point in the development process evidence for validity is collected also largely

depends on the type of validity being evaluated. An in depth discussion of each type of validity has been published in this journal.[195] This dissertation specifically focused on addressing threats to validity, and what evidence for validity is most critical for the research questions proposed in the development of items for the TCI. Content validity and response process validity is addressed in this manuscript with regard to item development and qualitative evaluation.

There are two main threats to construct validity: construct underrepresentation and construct-irrelevant variance.[202] Construct underrepresentation can occur when an assessment is too narrow and fails to include all of the important aspects of the defined construct. Evidence for content validity can be used to argue against construct underrepresentation, which includes defining boundaries for test content, criteria for boundary selection, and criteria used for selection of experts used to help define content boundaries. Construct-irrelevant variance can occur when an assessment is too broad and requires knowledge or cognitive abilities located outside the defined construct. There are two main categories of construct-irrelevant variance: construct-irrelevant difficulty and construct-irrelevant easiness.[201] Evidence for response process validity and content validity can be used as arguments against both sources of construct-irrelevant variance. Construct-irrelevant difficulty can occur when extraneous knowledge or cognitive processes are required, making an item or test more difficult for a certain subset of the target population. An example of a source of construct-irrelevant difficulty would be a high level of reading comprehension for an item testing chemistry content knowledge or use of vocabulary not familiar to a group of students required to understand an item stem or multiple-choice options. Construct-irrelevant difficulty is not as critical for criterion-

based assessments but can be a major threat to validity for constructs that focus on knowledge or cognition. Construct-irrelevant easiness is generally associated with testing items that are familiar to a certain group of students or hints imbedded in an item or test that students use to answer an item correctly without using the construct being measured. Examples of sources of construct-irrelevant easiness would be using a chemical reaction where students already know the outcome or if the correct option for one item can be determined by information given in another item.

Overview of Research

The research detailed in this chapter proposed developing items for a TCI by using validity as the core consideration in the item-design, development, and evaluation process, as shown in Figure 12. By clearly stating the intended use and interpretations of TCI data at the beginning of this research study, a developmental scheme was developed to collect validity evidence most critical for these propositions. Evidence from testing-stakeholders, both students (novices) and general chemistry instructors (experts) from our target population, was used throughout the development and evaluation process, both in qualitative and quantitative studies. Figure 12 illustrates both the linear and circular nature of the development and evaluation of the TCI. This study focused solely on qualitative research used to develop TCI items with quantitative research and the final TCI items being published in the future.

Research Questions

Q1 What are the most important thermochemical topics taught in college-level general chemistry?

Q2 What evidence supports the intended uses and interpretations of student scores and item responses on the Thermochemistry Concept Inventory?

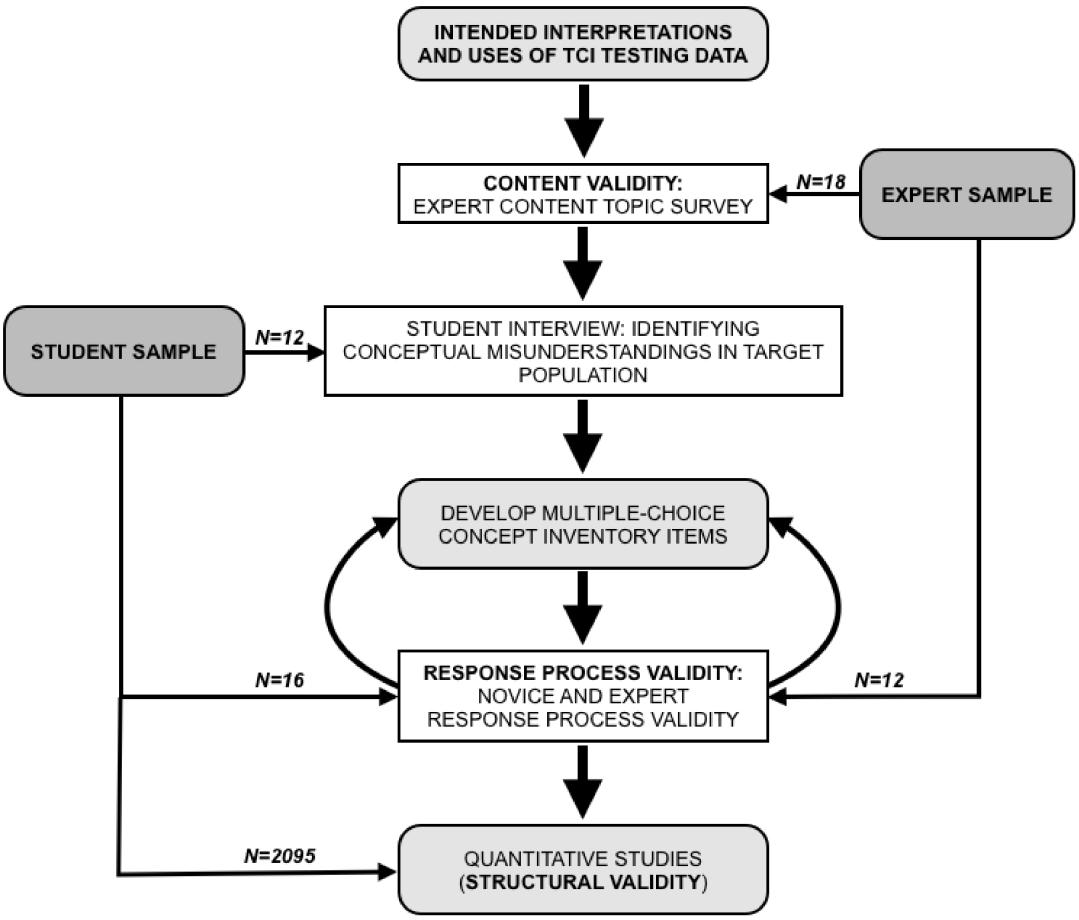


Figure 12. Overview of steps involved in development and evaluation of items for Thermochemistry Concept Inventory. Quantitative studies (structural validity) is published in a separate manuscript.

To address the first research question, we compiled evidence from content experts teaching at a variety of institutions. To address the second question, we interviewed students from the target population, as well as obtained feedback from content experts who evaluated TCI items. Propositions for the intended uses and interpretations of TCI testing data are given below.

1. The total score is an accurate measure of student conceptual understanding of college-level thermochemistry as measured by the TCI.
    a. The total score is based on the number of correct conceptions answered on the TCI.
    b. High total scores on the TCI most likely correspond to high conceptual understanding of thermochemistry.
    c. Low total scores on the TCI most likely correspond to low conceptual understanding/high frequency of using conceptual misunderstandings relating to topics covered in thermochemistry.
2. If a student chooses an item distracter, he or she is most likely using a singular conceptual misunderstanding that has been defined by the test developer.
    a. Item distracters correlate to a singular conceptual misunderstanding
    b. Students use the intended conceptual misunderstanding as a rationale when choosing a specific distracter.
3. Answering an item correctly signifies that students find the conceptual misunderstanding represented by distracters less attractive then the correct conceptions represented by the correct answer.

PARTICIPANTS, DATA COLLECTION, AND DATA ANALYSIS

This chapter summarizes several steps in the design and development process. For clarity, each step is discussed in order including the participants, the data collection methods, and the data analysis. The results and discussion section present a synthesis of how all data support the research questions.

Expert Thermochemistry Topics Survey

This study defined content experts as faculty who regularly teach first-semester general chemistry, as these are the stakeholders who use and make interpretations of TCI data.

Faculty were solicited by email to take an online survey on thermochemistry. To determine the most important thermochemistry topics taught in first-semester general chemistry classrooms, a topics survey was created in which experts were given a list of thermochemistry topics and asked to rate the importance of each topic using a 5-point Likert scale (important = 1 to unimportant = 5). An excerpt of the online survey, including survey instructions and survey items, can be found in Appendix B (B1). Topics were chosen by the researchers from the thermochemistry chapters of several of the most commonly used textbooks for general chemistry.[15, 170-173] Topic selection utilized chapter section headers and then content covered in those sections. Certain topics, such as enthalpy, had multiple levels of coverage, including operational definitions, sign conventions, and conceptual understanding. For these topics, the importance of each level of coverage was probed.

A total of 27 e-mails were sent to faculty at a variety of institutions. A total of 19 responses were received (70% response rate); however, only 18 of the experts completed the entire survey. Of those who completed the entire survey, 14 (78%) also provided comments. Participating experts taught at 8 different institutions having 6 different institutional Carnegie classifications as shown in Table 2. Data from the online survey

included expert's importance scores, comments associated with an individual topic, and information regarding any topics not covered in the survey.

Table 2. General Chemistry Instructors Participating in Content Topic Survey.

Institutional Carnegie Classification	*Expert Participation*		
	Content Topic Survey	*Expert Response Process Validity*	*In Both Studies*
RU/VH: Research university (very high research activity)	10	6	5
RU/H: Research university (high research activity	3	1	1
Master's L: Master's colleges and universities (larger programs)		2	
DRU: Doctoral/research university	2		
BAC/Diverse: Baccalaureate colleges–arts & science	1	1	
Assoc./Pub-R-L: Associate's–public rural-serving large	2	2	1
Total	18	12	7

For every topic, the percent importance was calculated, the equation used and worked examples for different topics in thermochemistry are found in Appendix B (B2). The percentage was determined by summing up the number of experts who rated a topic as either important or slightly important and then dividing this value by the total number

of experts who responded. In addition to determining percent importance, expert comments were compiled for each topic and the final comment section, where topics could be entered that were not included in the survey. For examples of how expert comments correlate to importance scores, see Appendix B (B2).

Think-Aloud Student Interviews

The target population for the TCI were students who were currently enrolled in first-semester general chemistry and had already covered thermochemistry material and second-semester general chemistry students (who had not yet covered thermodynamics). Given that the TCI intentionally does not cover any thermodynamic-specific topics (e.g., free energy, entropy, etc.), it was important that students had not covered these topics at the college-level.

Students were recruited via an announcement by the researcher during lecture in the spring 2010 semester. As an incentive for students to participate in the research study, 30 minutes of tutoring was offered (the majority of students did not take advantage of this offer). The think-aloud interview sample contained 12 students (8 female, 4 male) having varying majors (7 STEM and 5 non-STEM) from two universities (see Table 3).

Table 3. Student Sample for Think-Aloud Interviews.

Institutional Carnegie Classification	*Interviews*	
	Think-Aloud	*Novice Response Process Validity*
DRU: Doctoral/Research university	8	8
RU/VH: Research university (very high research activity)	4	5
BAC/Diverse: Baccalaureate colleges–arts & science		3
Total	12	16

A semi-structured think-aloud interview approach with probing questions was used for student interviews. This interview approach[174] has been successfully used in other chemical education studies to explore students' cognitive process when solving chemistry problems.[79] Students were asked to verbalize their thought process while solving thermochemistry problems developed by the researchers. These open-ended questions were designed to focus on the most important topics identified in the expert topic survey. Because obtaining accurate student conceptions is a goal of this study, probing questions were utilized in this interview approach.[175, 176] Students were given a non-chemistry warm-up problem to become familiarized with the interview approach[176] before being given 3 to 4 open-ended thermochemistry questions. Five open-ended interview questions were created for the student interviews. Two questions were given to all students; while up to two additional questions were given to maximize content coverage. The interview sessions included a 5-minute informed consent and interview

overview, a 5-minute warm up problem, and on average, a 45-minute session to answer the open-ended thermochemistry questions. Following the formal interview, students participated in a recap session where each question was reviewed by the researcher.

All interviews were video recorded (only the student workspace, with consent) and transcribed, including the recap session. Immediately following each interview, the researcher compiled field notes about the overall impression of the student and the interview, including any important aspects regarding conceptual understanding that might require specific evaluation in the coding process. Interview transcripts, video files, student work, and researcher field notes were imported into the qualitative analysis software package NVivo 8.[180] Each interview was coded through an iterative process. Initial coding placed excerpts of interview transcripts and video clips containing errors in bins based on a given thermochemistry topic (e.g., enthalpy, bond dissociation energy, work, etc.). These coded excerpts were then grouped into common themes (sign convention errors, definition errors, understanding errors, anomalous errors). From these themes, conceptual misunderstandings were identified and checked against those reported in the literature. A Fleiss Kappa of 0.67 was obtained using five independent coders who evaluated half of the transcripts coded in NVivo; this value represents fair to good agreement beyond chance.[203] Interview transcripts, student-generated work, and interview video footage were used in this evaluation. For each excerpt, the following was determined: (1) Does the excerpt contain an error that is clearly articulated by the student? (2) Is the student using a conceptual misunderstanding to explain the error? (3) What is the conceptual misunderstanding being used? and (4) Does the student consistently use this conceptual misunderstanding throughout the interview?

Novice Response Process
Validity Interviews

Using data from the expert topic survey and the student think-aloud interviews, a series of multiple-choice items were created. During the fall 2010 and spring 2011 semesters, these items were evaluated for response process validity. The target population and recruitment effort were the same as described above.

A total of 16 students (6 female, 10 male) were interviewed. Students' majors (9 STEM and 7 non-STEM) and year in college varied, as well as the lapse of time since the students were taught thermochemistry (4.2 months average; standard deviation of 3.5 months and a range from 0.5 to 12 months).

Novice response process validity interviews were designed to simulate a testing environment, such that students used authentic test-taking mentality. Therefore, a retrospective semi-structured think-aloud interview approach using probing questions was used. Retrospective interviews, occurring immediately after students answering testing items, are most appropriate for evaluation of test items in a test-taking environment.[181] Student participants were given a stack of ordered multiple-choice items on separate pieces of paper and instructed to answer items, in order, for approximately 15 minutes, with emphasis on thoughtfully answering items rather than completing all items. To ensure equal item coverage in interviews, item order was changed for each interview. During the interview, students were asked to read the item stem and explain their response process in answering each item. This included interpretation of item stems, item figures, and item responses. Probing questions by the researcher focused on identifying what conceptions students were using to choose their answer, or to eliminate responses.

Plausibility and independence of item responses were also evaluated based on student responses.

Iterative review, revision, and retesting of TCI items took place during the two semesters of interviews. Concerns with item wording, item option plausibility, item option ordering, and conceptions students were using to answer item options were the driving force for item revisions. Evaluation of revised items focused on the concerns found during initial testing and if these concerns were still present in follow-up interviews with students.

Expert Response Process Validity
Thermochemistry Concept
Inventory Survey

To collect evidence for expert response process validity, an online survey was developed, distributed, and responses collected using the Qualtrics®[183] web-based research suite. The survey was sent to the same 27 faculty initially contacted for the topics survey, in addition to 14 new faculty. The survey had four sections: (1) IRB consent, (2) demographics, (3) determination of topics significantly covered in the thermochemistry section, and (4) TCI items. Demographic information (section 2) included years teaching, number of students in general chemistry lecture section, and primary text book for the course. To assess what topics are actually covered in a typical thermochemistry section, experts were asked to indicate what concepts are significantly covered (section 3). Significantly covered was defined as "the majority of this topic is covered in the theromchemistry section, opposed to other sections", and experts could choose two options for each topic: significantly covered or not significantly covered. TCI items (section 4) were presented in the same format given to students, but the item

responses had radio buttons instead of letters (see Appendix B, B4). Each item also had a comment box where experts could provide feedback and voice any concerns about wording, content representation, or interpretation of response options.

A total of 9 experts (22%) completed the survey initially. All but two experts provided feedback on items in the form of comments. One item had enough comments and concerns that a follow-up survey regarding just that item was sent to the entire sample (41 experts) for clarification on ways to improve the item. Feedback was gathered from 12 experts (29%) during this follow-up; of these 12 experts, 9 completed the initial survey.

Of the 9 experts who completed the initial survey, 7 had taught for over 10 years. The typical lecture size was larger than 100 students. The text books used[15, 172, 173] were consistent with those used to choose thermochemistry topics. All expert survey responses were compiled into a spreadsheet by the Qualtrics® software for analysis. Item analysis included examining if there was consensus on the correct answer and if there were concerns about item wording or content representation. Only completed surveys were used in this analysis.

RESULTS AND DISCUSSION

Evidence for Content Validity:
Topic Percent Importance

Initial development of the TCI involved collecting evidence for content validity; specifically, determining the boundaries of what content is taught in the thermochemistry section of most general chemistry courses, and which topics are most important, and therefore, most emphasized in instruction. Established evidence for content validity is

essential, given that there are concepts covered in thermochemistry that may only be introduced but not emphasized. In addition, evidence for content validity can be used to argue against threats to construct validity. With the goal of creating a concept inventory that is short enough for formative assessment, the number of items and, therefore, concepts was limited, such that targeting the most important concepts is essential. Results from the survey used to determine the most important topics can be seen in Table 4.

The topics from the survey are grouped into three categories: definitions, sign conventions, and understanding. A general trend can be seen in Table 4, where the topics with the highest percent importance score are those in the definition-based category. Sign convention topics were generally the next most important, followed by those focused on understanding. Topics with low percent importance scores fell into two categories: topics that were not covered by many experts and those that lacked consensus among the experts surveyed.

Examples of expert ratings, explanations for those ratings, and corresponding percent importance score can be found in the Appendix B (B2). Explanations for not covering material generally involved being covered in second-semester general chemistry (e.g., understanding-Le Châtelier's principle), but did include comments on certain topics being too difficult or abstract for students (e.g., definition-state function). The input from these general chemistry instructors regarding the most important topics taught in thermochemistry was crucial for this research project; it served as the basis for what topics in thermochemistry was probed in student interviews to identify conceptual misunderstandings.

Student Conceptual Misunderstandings in Thermochemistry

An extensive body of literature on student conceptual understanding of thermochemical topics preceded this study (see Table 1 for a detailed summary). The topics in Table 4 were used to categorize previously published student conceptual misunderstandings and determine if there were thermochemical concepts with high percent importance scores but that lacked published conceptual misunderstandings, as shown in Table 5.

Student interviews addressed two important questions related to construct validity: (1) Does a lack of published conceptual misunderstandings relate to a lack of student conceptual misunderstandings or to a lack of research to identify these conceptions? and (2) Are conceptual misunderstandings published for students outside this study's target population (e.g., secondary students, student populations outside of the United States, etc.) found in our target population?

To identify conceptual misunderstandings in our target population, rich and descriptive data needed to be collected using cognitive student interviews. Open-ended interview questions were designed to target specific thermochemical concepts. These concepts included definition of system and surroundings, understanding Hess's Law, understanding heat capacity/specific heat capacity, and understanding heat of formation: These topics all had high percent importance rankings but little or no published conceptual misunderstandings.

Table 4. Percent Importance of Thermochemistry Topics as Rated by General Chemistry Instructors.

Topic	*Not Covered*[a] *%*	*Importance*[b] *%*
Endothermic and Exothermic		
Definition—Endothermic and Exothermic	0	100
Sign—Endothermic and Exothermic	0	100
Understanding—Endothermic and Exothermic	0	94
Thermal Energy and Temperature		
Definition—Heat	0	100
Understanding—Difference between q and T	0	100
Definition—Temperature	0	100
Understanding—Specific Heat	3	89
Understanding—Heat Capacity	3	89
Understanding—Conditions for Thermal Energy Transfer	33	72
Enthalpy		
Definition—Enthalphy	0	100
Sign—Enthalpy	0	100
Understanding—Heat of Reaction	0	100
Understanding—Heat of Formation	0	89
Understanding—Bond Disassociation Energy	0	89
Understanding—Hess's Law	0	72
First Law of Thermodynamics		
Definition—First Law of Thermodynamics	0	94
Understanding—Mathematical Form of First Law	17	50
Calorimetry		
Definition—System and Surroundings	0	89
Understanding—Constant Pressure Calorimetry	0	83
Understanding—Constant Volume Calorimetry	22	39
Work		
Definition—Work	6	61
Sign—Work	6	56
Uncategorized Topics		
Understanding—Phase Change	11	78
Understanding—Le Châtelier's Principle	22	72
Understanding—Forms of Energy	0	72
Sign—Internal Energy	33	39
Defintion—State Function	39	39

[a] Percent of experts who did not cover the topic and did not give it an importance score.
[b]Sample calculation can be found in supporting information (SI 4).

Table 5. Conceptual Misunderstandings in Thermochemistry.

% Importance	*Topic*	*Literature Reference*	*Student Population*[a]	*This Study*
100	Definition--Endothermic and Exothermic			
	•Exothermic/endothermic processes defined by energy loss/gain without distinction of thermal energy.			x
	•If work is being done on the system, the process is exothermic; if work is being done on the surroundings, the process is endothermic.			x
100	Definition--Heat			
	•Heat is a substance (noun).	94; 147	S, Int	x
	•Heat and enthalpy are the same thing.	73	PS	x
	•Heat and temperature are the same thing.	75; 85	S, Int	x
	•Heat and thermal energy are the same thing.			x
	•Heat is energy that is added to something.	93	PS	
	•Heat can be quantified as a specific amount of energy possessed by a body with temperature being a measure of that quantity.	85, 87	S, Int; PS	
	•Heat is a state function.	87	PS	
	•Heat is a substance that resides within objects and can pass from one to another.	94; 78	S, Int; PS, Int	
	•Heat is an extensive property.	85	S, Int	
100	Definition–Entalphy			
	•Entalpy is the change in heat.	73	PS	x
	•Entalphy is a measure of heat contained within a system.	73	PS	
100	Sign–Endothermic and Exothermic			
100	Sign–Enthalpy			
	The sign of ΔH_{rxn} cannot be determined without using tabulated values for enthalph.	93	PS	
100	Understanding–Heat of Reaction (ΔH_{rxn})			
	•ΔH_{rxn} is a scalar value.			x
	•ΔH_{rxn} of zero indicates how far a reaction will go towards completion.			x
	•ΔH_{rxn} can be found by subtracting the bond dissociation energy of the reactants from the bond dissociation of the products.			x
	•ΔH_{rxn} of zero indicates that no reaction will occur.			x
	•Enthalpy change is the same as internal energy change.	93		

Table 5. (continued)

% Importance	*Topic*	*Literature Reference*	*Student Population*[a]	*This Study*
94	Definition–Temperature			
	•Temperature is an extensive property of a substance or body.	90	PS, Int	x
	•Temperature is an accurate measure of heat.	85; 87	S, Int; PS	
	•Addition of thermal energy to a system will always result in an increase in temperature.	85; 148	S, Int	
	•Bodies with the same temperature will have the same energy	85; 94	S, Int	
	•Temperature of an object can be accurately determined by touch.	94	S, Int	
	•Temperature is a property of the substance from which the body is made.	94	S, Int	
94	Definition–First Law of Thermodynamics			
94	Understanding–Endothermic and Exothermic			
	•In an exothermic process, heat enters the system; in an endothermic process, heat exits the system.	90	PS, Int	x
	•Endothermic processes require energy to occur.	76; 74	PS, Int; S	
	•Exothermic reactions occur faster than endothermic reactions.	149; 90	SI, Int; PS, Int	
94	Understanding–Difference between q and T			
	Heat and temperature are the same thing.	75	S, Int	
89	Definition–System and Surroundings			
	•System includes everything you are studying, and surroundings is everything else.	150	PS	x
	•System can change during a reaction to accommodate products being formed.			x
	•System is contained within the surroundings, because surroundings is essentially everything, even the universe.			x
	•System always gives off heat.			x
	•Consideration of surroundings is not required when evaluating energy transfer process, only the system.	151	PS, Int	
89	Understanding–Specific Heat			
89	Understanding–Heat of Formation (ΔH_f)			
	•$\Delta H°_f$ is always exothermic.	90	PS, Int	
	•ΔH_f is the final enthalpy.			

Table 5. (continued)

% Importance	*Topic*	*Literature Reference*	*Student Population*[a]	*This Study*
89	Understanding–Bond Disassociation Energy			
	•Heat is released when a bond is broken.	74; 152; 90	S; PS; PS, Int	x
	•Chemical bonds are structures that require energy because they need to build.	74	S	
83	Understanding–Constant Pressure Calorimetry			
78	Understanding–Phase Change			
76	Understanding–Heat Capacity			
	Thermal energy can be transferred both from a hot to cold body and from a cold to hot body, analogous to reaching chemical equilibrium.			x
72	Understanding–Hess's Law			
72	Understanding–Forms of Energy			
72	Le Châtelier's Principle			
61	Definition–Work			
	•When work is done on a system, heat is added to the system.	87	PS	x
	•Work is a state function.	87	PS	
56	Understanding–Conditions for Thermal Energy Transfer			
	•Once thermal equilibrium is reached, small differences in temperature between two bodies can exist.	85	S, Int	x
	•Other factors can affect thermal energy transfer besides temperature differences between two bodies.	85	S, Int	
	•Transfer of thermal energy is synonymous with energy conversion.	83	PS	
56	Sign–Work			
	If work is done on the system, the sign for work is negative.			x
50	Understanding–Mathematical Form of First Law			
39	Definition–State Function			
39	Sign–Internal Energy			
39	Understanding–Constant Volume Calorimetry			

Note. S = secondary, PS = post secondary, SI = secondary instructor, PSI = post-secondary instructor, Int = international (outside United States)

Five open-ended questions focusing on these topics were created and used for student think-aloud interviews. Two of the open-ended questions (Figure 13) on Calorimetry, blocks, and bond dissociation energy (BDE) are used to illustrate how evidence for validity was collected through the item development process. The BDE item targeted student conceptual understanding of bond dissociation energy, definition of exothermic and endothermic reactions, and reaction enthalpy. The blocks item targeted student conceptual understanding of thermal energy transfer between a system and surroundings. Incorrect responses were identified for all five open-ended items and then coded by topic (e.g., enthalpy, thermal energy and temperature, etc.). For example, the following is an excerpt from Student 1 responding to the BDE item: "Exothermic is when you . . . wait . . . when it doesn't require energy to heat, I think. And then endothermic—I know energy is required, that's like heat coming into the system." This excerpt was coded in the Understanding–Endothermic and Exothermic category and represented the conceptual misunderstanding that only endothermic processes require energy to occur. In another example from the BDE item, Student 2 uses the well-documented conceptual misunderstanding that heat is released when a bond is broken: "Well the heat's being released when you're breaking the bonds." Student responses to the blocks item bring to light conceptual misunderstandings focused on defining the system and surroundings and conditions of thermal energy transfer. When asked to define the system and surroundings, Student 3 responded: "The system, I believe, is anything you're studying. The surrounding would be everything else." This excerpt provides evidence for the conceptual misunderstanding that the system includes everything you are studying and the surroundings is everything else. Using this logic, students typically struggle to explicitly distin-

guish what actually constitutes the system. For example, in the blocks item, several students qualify both the block and the water itself as the system. Therefore, correctly interpreting observed temperature changes (in the surroundings) and relating them to the change in the system becomes problematic.

In addition, evidence for the conceptual misunderstanding that thermal energy can be transferred both from a hot to cold body, and from a cold to hot body (analogous to chemical equilibrium) is exhibited in the following dialogue between the researcher and Student 4):

> Researcher: Can you draw a rectangle inside the coffee cup representing block T_3 and describe the heat flow with respect to the block?
>
> Student 4: So it would be like that [student draws arrows both coming and going from block].
>
> Researcher: Just so I understand your drawing, you have the heat entering the block and also leaving the block?
>
> Student 4: Yes.

BDE Open-Ended Item

Use the chemical equation and structural formula given below to estimate if the reaction enthalpy ($\Delta H_{(rxn)}$) will be exothermic, endothermic, or zero. Bond dissociation energies (D) are given in the table on the right.

$$O{=}O \longrightarrow O + O$$

$$O_{2(g)} \rightarrow O_{(g)} + O_{(g)}$$

Bond	D (kJ/mol)
O=O	498
O-O	142

Blocks Open-Ended Item

You want to DECREASE the temperature of the water in the Styrofoam cup shown below. Below there are three metal blocks of the same metal at three different temperatures. Which block(s), when placed in the cup, would you expect to see a decrease in the water temperature?

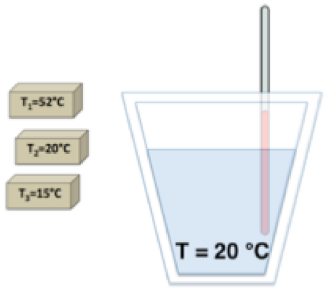


Figure 13. Two open-ended items used in student think-aloud interviews.

Many incorrect responses related to incorrect sign convention for DH_{rxn}, work, or heat were observed. Students often lacked any type of rationale for the sign convention used, relying simply on what they had memorized. In other incorrect responses, students misclassified numerical values of enthalpy and temperature. For example, students treated DH_{rxn} as a scalar quantity (three students) or temperature as an extensive quantity (six students). Students had difficulties distinguishing thermal energy from heat or enthalpy (eight students), and all but one student referred to heat as a substance. This was not surprising, as most text books treat heat as a substance or lack explicit distinctions between heat and thermal energy or enthalpy.[75, 85, 94] Students also had difficulty defining the system and surroundings for specified processes and reactions. Many struggled with the ambiguity of defining the system, especially for chemical reactions where a physical object (e.g., a metal block in a coffee-cup calorimeter) was not present (five students).

Many conceptual misunderstandings dealing with explanations for observed phenomena identified in secondary school students were not observed in our sample. Such conceptual misunderstandings included temperature as a measure of a body's heat,[85, 87] conductors transfer heat more slowly than insulators,[85] and the temperature of an object can accurately be determined by touch.[94] Given the prevalence of these conceptual misunderstandings in secondary students, some were still included as distracters in TCI items used in novice response process validity (RPV) interviews. Students in the novice RPV sample also did not find these distracters probable, providing evidence that our small interview sample may be representative of our target population. Quantitative analysis using a much large-student samples were also provided evidence if distracters using these alternative conceptions are unattractive to most students.

A pool of 12 multiple-choice items were created based on the open-ended items used in student think-aloud interviews as well as new items that have stems that would allow for inclusion of multiple independent conceptual misunderstandings as distracters. Wording from student interviews was incorporated into the wording of distracters to make them believable to students.[204]

Evidence for Novice Response Process Validity

One of the critical propositions of the use of TCI testing data is: If a student chooses an item distracter, he or she is most likely using a singular conceptual misunderstanding that has been defined by the test developer. Specifically, item distracters correlate to a singular conceptual misunderstanding, and students use the intended conceptual misunderstanding as a rationale when choosing a specific distracter. Novice RPV interviews were designed to collect evidence for these propositions. Evidence included having students verbalize what conception they used in choosing their answer and what conceptions they used to eliminate any responses. Categorization of evidence gathered from student interviews can include (1) no useable evidence, (2) evidence for response process validity, and (3) evidence for a need to revise item, as shown in Table 6. Based on the multiple-choice versions of the BDE and blocks items (Figures 14 and 15, respectively), examples for each category in Table 6 can be found in the Appendix B (B5). Both Figures 14 and 15 give further examples of evidence for item revision stemming from the multiple-choice version of items.

Table 6. Categorization of Types of Evidence Observed During Novice Response Process Validity Interviews.

Category	*Type of Evidence*
No Useable Evidence	•Did not attempt or did not answer item.
	•Guess; answered item without using rationale related to item content.
Evidence for Response Process Validity	•Chose correct option; explanation included correct conception.
	•Chose incorrect distracter; explanation included intended conceptual misunderstanding.
	•Eliminated correct option; explanation included intended conceptual misunderstanding.
	•Eliminated incorrect distracter; explanation included correct conception.
Evidence for a Need to Revise Item	•Chose correct answer; explanation included conception other than correct conception.
	•Chose incorrect distracter; explanation included a conceptual misunderstanding other than what was used in the design of the distracter.
	•Extraneous information used to either choose or eliminate item option.
	•Eliminated incorrect distracter because it did not seem plausible and/or reasonable.

BDE Item: Novice Response Process Validity

The dissociation of water into hydrogen and oxygen is given below with the associated standard reaction enthalpy. Choose the most accurate answer below.

$$2H_2O(g) \rightarrow 2H_2(g) + O_2(g) \qquad \Delta H^{\circ}_{rxn} = 468.5\ kJ/mol$$

	Bond Enthalpy
O-H	460.0 kJ/mol
H-H	436.4 kJ/mol
O=O	498.7 kJ/mol
O-O	142.0 kJ/mol

(A) If 1 mole of O_2 is produced, 498.7 kJ of energy will be released

(B) The bond enthalpy of the products is larger than the bond enthalpy of the reactant

(C) The ΔH°_f for H_2O is equal to -468.5 kJ/mol

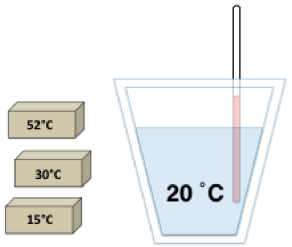
(D) For the reaction to occur, 468.5 kJ/mol needs to be added

Student	Response to Option A
Student 5	I thought option A was incorrect because if one mole of O_2 is made [pointing to the words in option A], 498.6 kJ of energy will be released. And I guess [point to the two different values for the bond enthalpies for single and double-bonded oxygen] that there are two values, here is a double bond [pointing to O=O] between oxygen and here is a single bond between oxygen, so this answer really doesn't specify what it means by "O_2", because I would assume that both of them can be O_2.
Student 6	Yeah, this one was confusing. [student reads stem out loud] At first I started reading through the answer and try to pick one out. But then I noticed the table and thought maybe I should use the table somehow . . . I don't know I just felt like I had to use this table because it was here.

Figure 14. Novice response process validity interviews of the multiple-choice bond dissociation energy item. This revealed that students struggled using bond enthalpy values and ascertaining the correct structure of O_2. This provides evidence for construct-irrelevant difficulty, a threat to validity of the interpretations of response A, which is designed to be the correct answer.

Blocks Item: Novice Response Process Validity

A styrofoam coffee cup contains water at 20 ˚C. Three identical metal blocks at three different temperatures are shown to the left of the cup. Choose the most accurate response below.



(A) If the 52 ˚C block is put into the cup, the final temperature of the water would be 32 ˚C

(B) Thermal energy will flow back and forth between a block and the water until thermal equilibrium is reached

(C) The system would be defined as everything in the coffee cup and the surroundings would be everything else

(D) If the block at 15 ˚C is added to the water, the process can be described as an endothermic process with respect to the block

Student	Response to Option A
Student 7	A, if the 52 degree block is put in the cup. There is not enough information, in the data to come up with 36. We don't know the mass of this block that would, as one simple thing or how much water is there.
Student 8	I don't feel like I had enough information [pointing to the 32 degrees in option A] to do the calculation to figure out if 32 degrees was at all right or not. Just thinking about it, it sounds reasonable, but not knowing what the metals are, etcetera.
Student 9	I eliminated A. It says the if the 52 degree block is put in cup the final temperature of the water will be 32 degrees. I didn't think that would be right, because it's just taking 52 minus 20 and saying that's the answer [circling 32 degrees in option]. That might not happen.

Figure 15. Novice response process validity interviews of blocks item. This revealed that students at multiple institutions did not find option A plausible.

The multiple-choice version of the BDE item was created using a simple decomposition reaction with associated bond enthalpy values, as shown in Figure 14. However, students struggled with using the bond enthalpy values, specifically related to molecular oxygen. Others were simply intimidated by the values and thought calculations were needed to answer the item. Given that option A was designed to identify the conceptual misunderstanding that bond breaking is an exothermic process, and not testing whether students know the correct structure for O_2, this increased the difficulty of the item due to construct-irrelevant knowledge. To attenuate this construct-irrelevant difficulty, the revised version of the BDE item (see Figure 16) uses a generic reaction with only one bond being broken (A-B) to form one new bond (B-C). Option A in the revised version utilized the simplified reaction structure to probe the conceptual misunderstanding of bond breaking being an exothermic process, while option B in the revised version was changed to be the correct answer. Option C remained structurally unchanged and just reflected the updated reaction enthalpy value of the revised version of the BDE item. For the blocks item in Figure 15, option A was found improbable to many students in our RPV sample, across all three institutions. Interestingly, this was a conception that was identified in the think-aloud student interviews. This provides an example of how RPV interviews can provide further evidence for conceptual misunderstandings identified from student think-aloud interviews or evidence for lack of generalizability. This is critical when the number of student interviews is small. Based on this evidence, option A was replaced.

This process of item evaluation and revision took place for all items; revised items were retested after revisions to collect further evidence for response process validity.

Once this evidence had been established for the current version of all TCI items, an online survey to establish expert response process validity was developed. Of the 12 items evaluated in the novice RPV interviews, 10 moved on to the expert RPV survey stage of development.

Evidence for Expert Response Process Validity

Expert feedback on item content, wording, and consensus of the correct answer are all sources for evidence of expert response process validity. If the TCI is a true measure of thermochemical conceptual understanding, then the correct response should represent the correct conception and the incorrect responses should represent conceptual misunderstandings. If the correct answer does not represent the correct conception or an incorrect response represents a correct conception, then the validity of using the total score of the TCI as a measure of conceptual understanding is threatened. Expert responses to an online survey included demographic information, multiple-choice responses to 10 TCI items, and comments on TCI items. Responses to TCI items were used to evaluate if experts reached a consensus of the correct response, while the comments provided insight on lack of consensus and identified concerns of content usage, representation, and item wording. The initial online TCI expert RPV survey was completed by nine experts, while a follow-up survey addressing concerns from the expert RPV survey was answered by 12 experts, including all nine who took the initial expert RPV survey. Of the 12 experts who participated in expert RPV surveys, seven also participated in the initial thermochemistry topic survey.

To check the assumption that thermochemical topics with high percentage importance are taught primarily in the thermochemistry section of general chemistry, experts were asked to rate the thermochemical topics presented in Table 4 as significantly covered or not significantly covered. All thermochemical topics rated with a high percent importance (above 75%) were scored as significantly covered by the majority of experts (9/12, 75%), while only the concepts related to work were scored as not significantly covered by the majority of experts (7/12, 58%). Thus, distracters representing conceptual misunderstandings related to work were minimized in the final version of the TCI, to reflect the lack of consensus of experts sampled in this study.

The average expert TCI score was 8.9/10 with a standard deviation of 0.6 and a maximum score of 10/10 (1/9 experts) and minimum score of 8/10 (2/9 experts). The average item agreement (number of experts choosing the consensus answer) was 8/9 with a standard deviation of 1.9, having a maximum agreement of 9/9 (6/10 items) and minimum agreement of 3/9 (1/9 items). For three of the items that lacked a full consensus of experts (scoring 8/9 or 7/9), no comments were provided to raise concerns about multiple correct answers. It is unclear if the lack of consensus reflects an expert error or an actual disagreement with the other experts surveyed.

The BDE item (see Figure 16) had a clear lack of consensus among experts, where option C (incorrect) was chosen by 6/9 experts, while 3/9 chose the answer that represented the correct conception (option B) as designed by the researchers. Six of the experts provided comments on the item, of which, four chose the incorrect response.
A general concern was that both option B (correct) and option C (incorrect) were interpreted as representing a correct conception. The other concern was that the term bond

enthalpy might be less familiar to students than the term bond energy. A follow-up survey was created that addressed this lack in consensus, where the following statement was provided: The third response is testing if students know that the reaction enthalpy can only convey the difference in energy of the reactants and products, and can not definitively be used to address how much energy is needed for the reaction to occur. Do you believe this concept is important for students to know in your first-semester general chemistry classroom? Of the 12 experts who responded to this follow-up question, seven believed that option C was related to an important concept, while five did not. All five experts who did not believe this concept was important provided comments as to why. Expert comments could be classified into three categories: (1) suggestions centered around changing wording of option response to clarify incorrect conception (e.g., such as exactly 400 kJ/mol or net 400 kJ/mol), (2) belief that response was testing a concept related to kinetics, which is not covered in thermochemistry, (3) that both answers are correct and a fourth option should include a both B and C above.

$$A-B + C \rightarrow A + B-C \qquad \Delta H^{\circ}_{rxn} = 400\ kJ/mol$$

(A) The breaking of the A-B bond is exothermic and the making of the B-C bond is endothermic

(B) The bond enthalpy of the reactants is larger than the bond enthalpy of the products

(C) For the reaction to occur, 400 kJ/mol needs to be added

Classification of Response*	Expert Responses to Version 4.0
1	For the third response to be definitively incorrect, a better statement would be "If 400 kJ/mol is added, the reaction will occur." As currently stated, the third response is a necessary but not sufficient condition.
1	The ambiguity might be resolved by adding the word "exactly" immediately before "400 kJ/mol" in the third response
1	These choices do not seem clear to me. The first choice is clearly wrong. On the second choice, what is meant by "the bond enthalpy of the reactants (products)"? I would say the magnitude of the bond enthalpy of A-B is greater than the magnitude of the bond enthalpy of B-C. On the third choice, I would say, a net 400 kJ/mol is absorbed when the reaction occurs.
2	The reaction does not necessarily occur when 400 kJ/mol is added. If the activation energy, Ea, is greater than Delta H of the reaction than the reaction will not occur
2	This seems to be getting at the difference between thermodynamics and kinetics. My students do not learn kinetics until the second semester. I would expect first semester students to understand that a positive enthalpy of reaction represents the NET amount of energy that would need to be added for the reaction to occur.
2	This is a question of kinetics, which is covered in the second semester. Technically, the third answer can be considered correct
2	Activation energy isn't covered until the second semester
3	To fully see if the students get the concept, one of the choices should be "both b and c above."
3	I think both the second and third answers are correct.
3	The first answer is incorrect; the second and third answers seem correct.
***1** = Change wording to clarify concept; **2** = Actual concept being tested includes kinetics; **3** = Statement that both A and B are correct	

Figure 16. Expert comments regarding bond dissociation energy item. This is both from the original survey and to a follow-up survey focusing solely on this item. Expert responses could be classified into one of three categories.

To address these concerns, novice RPV interviews of the BDE item were analyzed focusing on the comments made by experts in the RPV surveys. In the student RPV interviews, students both chose the option (referring to the amount of energy that needed to be added) using the intended conceptual misunderstanding as well as eliminated this distracter using the correct conception. Of the 10 students who responded to the BDE item in the RPV interviews, not one discussed any concept of activation energy in either choosing or eliminating the option. This is most likely because all students interviewed had not yet had instruction on kinetics or thermodynamics. Given that the TCI is intended to be used during thermochemistry instruction or before students receive instruction on thermodynamics, the concerns expressed by some experts that option C requires understanding of activation energy is minimized. However, the concern that clarity might be an issue was addressed by simplifying the structure of the option to the following: The reaction requires 400 kJ/mol of energy to occur.

Student familiarity of the term bond enthalpy varied by institution and ability, where the term enthalpy intimidated some students who lacked a conceptual understanding of enthalpy. One expert comment summarized this concern: “Our book uses the term bond energy. Strictly speaking, ‘bond enthalpy’ is more correct, but nobody actually talks that way in my experience.” Therefore, a revision was made to option B (see Chapte V, Figure 22) that adds the term energy in parenthesis after bond enthalpy.

In summary, obtaining feedback on TCI items from the stakeholders of the TCI, chemistry students and general chemistry instructors, was used to collect evidence for response process validity as well as to make targeted revisions to TCI items. By collecting evidence from students and educators at a range of academic institutions, evidence related

to the generalizability of the intended uses and interpretations of TCI data were collected. This evidence obtained from student and expert samples is complimentary. The expert sample provided feedback that was technical and informative for item revisions. Evaluation of 10 TCI items showed expert consensus on all but one item, which was revised using feedback provided by expert comments. Student feedback on TCI items was a rich source of information on the utility of the TCI as a diagnostic instrument for identifying student alternative conceptions, as well as providing evidence for or against construct-irrelevant variance. Figure 14 provides an example of how evidence of construct-irrelevant difficulty was identified through student RPV interviews and used to make item revisions. Figure 15 provides evidence of a distracter that seen as implausible to multiple students in our sample and provided the basis for the removal of this distracter.

CONCLUSION AND FUTURE RESEARCH

The methodology presented in this study puts emphasis on obtaining feedback from the stakeholders of a thermochemistry concept inventory: general chemistry instructors and students, multiple times during the item development process. To address our first research question (What are the most important thermochemical topics taught in college-level general chemistry?), we obtained feedback from general chemistry instructors from a diverse range of colleges and universities. Thermochemical topics most often covered in the thermochemistry chapter of general chemistry textbooks were ranked by a percent importance score. These scores were substantiated by expert comments, which provided insight on topics that lack consensus by our expert sample. In addition, topics with the highest percent importance score were presented to another sample of experts

(39% of the original expert sample) at the beginning of the response process validity survey. All topics with a percent importance score of over 75% also were rated as significantly covered in the thermochemistry section of general chemistry by over 75% of experts surveyed.

To address our second research question (What evidence supports the intended uses and interpretations of student scores and item responses on the TCI?), evidence of content validity and response process validity were collected from multiple qualitative studies. Additional evidence for response process and structural validity were collected from large-scale quantitative studies, part of a future publication, that will include the final version of the TCI and relevant psychometric analysis. Two threats to validity were evaluated in this study: construct underrepresentation and construct-irrelevant variance. The determination of the most important topics taught in the thermochemistry section of general chemistry courses was used to determine what conceptual misunderstandings were used as distracters in TCI items. This provides evidence for content validity and against the threat of construct underrepresentation of thermochemistry topics in the theromchemistry concept inventory. Interviewing students who had taken TCI items allowed for evaluation of the response process that students are using to both choose their answer and eliminate other options. This is a source of evidence for response process validity and against construct-irrelevant variance, both construct-irrelevant difficulty and easiness. Lastly, use of an extensive body of literature of student alternative conceptions in thermochemistry across multiple fields (chemistry, physics, engineering) and several student populations (college, secondary, international) along with student interviews from our target population helped identify what conceptual misunderstandings should be

included as distracters in TCI items. This helped increase the resolution of the construct of thermochemistry conceptual understanding, as measured by the TCI. Not surprisingly, not all previously-published conceptual misunderstandings were found in our sample. This is most likely due to the limitations in sample size of qualitative studies and that our target population has a specific background that is represented by a specific sub-set of thermochemical conceptual misunderstandings. By including some popular conceptual misunderstandings found in other populations as TCI distracters, even if not identified in our sample, will help minimize threats to construct validity and the resolution of the TCI. Quantitative analysis of these distracters will provide evidence for or against inclusion in the final version of the TCI.

Quantitative evaluation of the psychometric properties of the 10 items resulting from the research presented in this manuscript provides some measure of the utility of using validity as the cornerstone for the design, development, and evaluation of concept inventory items. Specifically, whether the time invested in collecting evidence for validity multiple times during the item development process results in items with favorable psychometric properties. These properties include test-level metrics (reliability) and item-level metrics (difficulty, discrimination, targeting to student ability). In addition, evidence for the internal structure of the TCI is evaluated, assessing the expectation that the TCI is unidimensional (only measures one construct, thermochemical conceptual understanding). This research culminates in the final version of all 10 TCI items, including psychometric properties, to provide complimentary evidence for validity to what was found in this study.

CHAPTER V

PSYCHOMETRIC ANALYSIS OF THE THERMOCHEMISTRY CONCEPT INVENTORY

ABSTRACT

Psychometric analysis of the Thermochemistry Concept Inventory (TCI) was completed in three stages: beta testing of 15 TCI items ($N = 280$), pilot testing of a 12-item TCI ($N = 485$), and a large data collection using a 10-item TCI ($N = 1{,}331$). The TCI was used in both formative assessment (pilot study) and summative assessment (large data collection study). Evidence for generalizability and administration conditions, response process validity, and structural validity was collected during all three stages of quantitative analysis. Both Classical Test Theory (CTT) and Rasch model analysis were used to collect this psychometric evidence. The TCI was found to be unidimensional, both using confirmatory factor analysis and principal component analysis (Rasch). The quality of TCI items was evaluated using traditional CTT analysis (item difficulty and item discrimination) and Rasch model analysis (item fit, item targeting, item-option discrimination). Strong evidence for high item quality was collected for all items on the TCI, with the exception of item K when used in summative testing. However, item K performed well when the TCI was used in formative assessment and was retained in the final version of the TCI. Test-level analysis indicated that the TCI was well targeted to

the ability of students in our testing samples. Traditional reliability estimates were low, as expected for a short assessments like the TCI, but strong evidence for item-level reliability was collected using the Rasch model. Scores on the TCI by students enrolled in a selective honors section of general chemistry were higher and statistically different than student enrolled in general sections, providing evidence that the total score on the TCI can be a measure of student ability.

INTRODUCTION

The use of concept inventories in the evaluation of student conceptual understanding has led to a wide array of assessments. The development, use, and evaluation of many concept inventories have been published in the *Journal of Chemical Education*. Concept inventory length, format, and intended use can vary widely. For example, use of multiple-choice formats can include single-tier items, two-tier items, or a combination of both. In addition, sophistication in the psychometric analysis of existing concept inventories, as well as newly-developed concept inventories, is moving to include both parametric (CTT) and nonparametric (Rasch model) statistical analysis.[195, 205] Many of the standards used by the assessment community are becoming accepted and expected for new assessment instruments published by the chemistry education research community,[195] including concept inventories. Evidence for the validity of uses and interpretations of testing data is now being collected throughout the design, development, and evaluation stages of assessments.[112] This evidence allows for test users to evaluate what construct the assessment is testing, what population the test is targeted for, what interpretations and uses of testing data are appropriate, and what psychometric evidence is provided for the test structure and relation to other variables. Depending on the intended interpretations and

uses of testing data, sources of evidence for validity can be both qualitative and quantitative in nature.

This paper presents quantitative evidence for the intended uses and interpretations of data from the TCI. The TCI is a 10-item multiple-choice assessment that uses identified thermochemical conceptual misunderstandings of college-level general chemistry students as distracter options. The design, development, and qualitative evaluation of the TCI items have been detailed previously in the *Journal of Chemical Education*.[206] The final version of the TCI is presented along with a detailed psychometric evaluation of data at the test and item levels. Use of qualitative data are used to help understand and explain quantitative results, and provide a complete appraisal of evidence for the validity of interpretations and uses of TCI testing data.

Intended Uses and Interpretations for TCI Testing Data

The TCI is designed to be a diagnostic assessment to identify student conceptual misunderstandings of the most important topics predominantly covered in the thermochemistry section of college-level general chemistry. The TCI is intended for use during the learning of thermochemical topics (formative assessment) as well as a diagnostic for students who have completed the thermochemistry section in general chemistry. The use of the TCI for summative assessment of thermochemistry conceptual understanding can provide evidence for the effectiveness of a new instructional strategy to address conceptual misunderstandings in thermochemistry. Alternatively, the TCI can be used to assess the thermochemical conceptions students are using prior to instruction on thermodynamics. Unlike many traditional assessments, the diagnostic nature of the

TCI puts emphasis on using and interpreting item responses rather than the total score. The total score on the TCI has been designed to be a measure of student conceptual understanding of thermochemical topics. However, details of specific conceptual misunderstandings can be gained by interpreting what incorrect answers (distracters) students find attractive. For all TCI items, each distracter represents a specific conceptual misunderstanding. The challenge with making correlations between a student's incorrect item responses and use of the conceptual misunderstanding that the distracter was designed to represent is that there is additional error associated with interpretations at the item-response level than at the item-score or test-score levels.[112] Therefore, qualitative and quantitative evidence for these interpretations need to be collected from students in the target population. These studies provide evidence for response process validity. Given that most of the interpretations of TCI testing data are intended to be at the item level, this dissertation presents evidence primarily at the item and item-response levels.

Psychometric Evidence for Interpretations and Uses of Thermochemistry Concept Inventory Testing Data

Psychometric evidence was collected throughout the development process of the TCI.[206] Quantitative evaluation of TCI testing data in this study benefit from the ability to sample larger portions of students in our target population and provide evidence that can increase the generalizability of the TCI results. For example, psychometric analysis of TCI testing data is used as evidence that conceptual misunderstandings that were identified in qualitative studies accurately reflect those used by larger samples from the target population. Alternatively, evidence for the inclusion of published thermochemical alternative conceptions not identified in qualitative studies was collected from quantita-

tive studies presented in this paper. In addition, evidence for response process validity from student cognitive interviews is compared with evidence from item response analysis from quantitative studies. For example, an item distracter that seems implausible to students interviewed in qualitative studies and is not selected by many students in quantitative studies, provides complimentary evidence for revision or removal of this response. Lastly, evidence for structural validity, whether the theoretical structure of the TCI agrees with statistical evidence, is evaluated. Figure 1 provides an overview of how qualitative and quantitative evidence was collected for TCI items leading to the 10-item version presented in the supporting information (see Appendix C). A key feature of this design methodology is the use of qualitative data to inform quantitative analysis.

Theoretical Framework

The TCI is a psychological test designed to measure the psychological construct of thermochemical conceptual understanding. TCI item options are designed to measure conceptual understanding of specific concepts, which are hypothetical constructs. That is, conceptual understanding cannot be measured directly, only estimated by making inferences from observations of student responses to TCI items. The degree to which these inferences accurately reflect conceptual understanding corresponds to the validity of interpretations that can be made from these observations. Psychometrics is the "science concerned with evaluating the attributes of a psychological test", which includes testing scores, validity, and reliability, among others.[117] Just as people have psychological attributes, such as conceptual understanding, psychological tests also have theoretical constructs that need to be estimated, namely validity and reliability.[117]

For this research study, evidence for the validity of interpretation of uses of TCI testing data were collected and reported as psychometric attributes of the TCI. Specifically, psychometric analysis focused on collecting evidence for structural and response process validity.

METHODOLOGY

Student Participants, Thermochemistry Concept Inventory Administration, and Data Collection

The three phases of quantitative data collection are shown in Figure 17: beta testing of potential TCI items (15 items), pilot testing the TCI (12 items), and large-scale data collection with the final TCI (10 items). The pool of 15 potential TCI items was reduced through evaluation during all three phases of quantitative analysis.

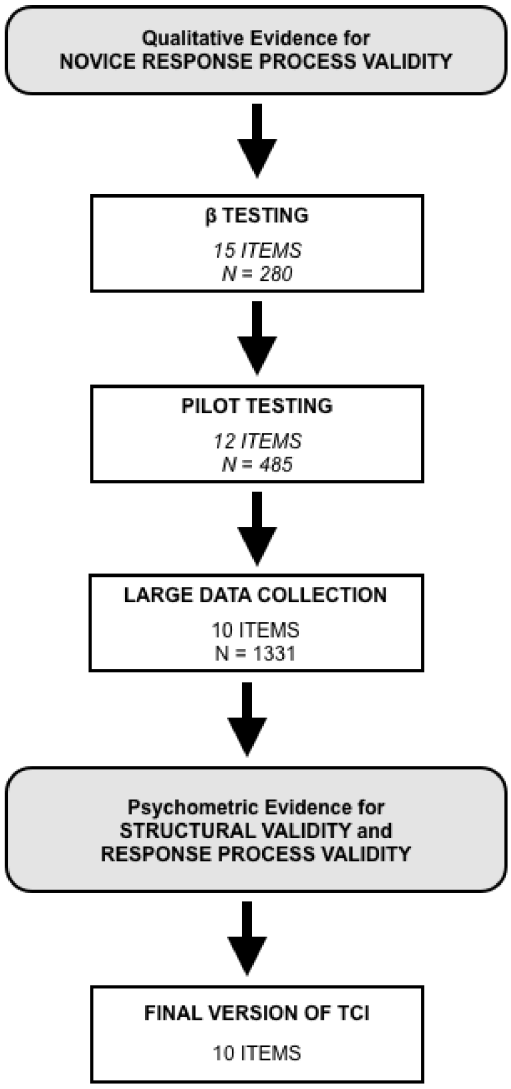


Figure 17. Development and evaluation of Thermochemistry Concept Inventory items using qualitative and quantitative (this study) evidence for validity, leading to final 10-item instrument.

Beta and pilot testing of Thermochemistry Concept Inventory items. The beta testing and pilot testing included 765 students from two different institutions (Carnegie Classification: β_1 and pilot = RU/VH, 571 students; β_2 = DRU, 194 students). During these administrations, the TCI was given in either lecture or during a required laboratory

recitation. Students in these administrations were informed that their score on the assessment was not part of their course or laboratory grade; therefore, these were low-stakes administrations. Beta testing took place in both first-semester general chemistry lecture after instruction on all topics in the thermochemistry section (β_2) and in second-semester general chemistry lecture before instruction on thermodynamics (β_1). Pilot testing of the TCI took place in second-semester general chemistry laboratory recitation and was administered by teaching assistants (TAs) before instruction on thermodynamics. The TCI was administered to all 18 TAs during the TA meeting prior to data collection to demonstrate the proper administration protocol and to help obtain TA buy-in on the utility of the TCI in formative assessment. All student identifiers included a TA code, such that misfitting students identified by Rasch person ability fit statistics could be traced back to a specific TA.

Large data collection. The large data collection utilized a large-enrollment general chemistry program located in the Pacific West region of the United States (Carnegie Classification: RU/VH). Four second-quarter general chemistry sections and one second-quarter honors section were administered the TCI (10 items) during lecture, after instruction on thermochemistry. The honors section was comprised of the top 2% of students from the first-quarter, based on course grade. The TCI was used in replacement of a quiz on thermochemistry, and student scores on the TCI were used for evaluation in the course. Based on the time required for the majority of students to complete the TCI during beta testing and the pilot test, a 30-minute block of time was use for the administration of the TCI for the large data collection. Most students required much less then 30 minutes, finishing around 15 minutes after receiving a paper copy of the TCI along with a

scantron. Institutional Review Board approval was obtained on the paper copy of the test form of the TCI, and consent was indicated by students on their scantron. Both the scantron and paper copy of the TCI were collected by instructors. After student scantrons were scanned by the instructors of each section, an independent party removed all data for students who did not provide consent. All student identifying information was removed before the first author obtained the data in a spreadsheet format.

Data Analysis

Classical Test Theory analysis. Both dichotomous and polytomous data sets were used for CTT analysis of TCI testing data. Item difficulty (p) and discrimination (D) were calculated using the top and bottom 27% of students based on TCI test scores, as the sample was greater than 200 students.[165] The proportion of students in the top 27% to answer an item correctly (p_{high}) and proportion of students in the bottom 27% to answer an item correctly (p_{low}) were calculated. Item difficulty represents the proportion of students in both the p_{high} and p_{low} who answered an item correctly, divided by the total number of responses. Item discrimination is the difference between p_{high} and p_{low} divided by the number of students in the high-scoring group. Thus, an item that can discriminate students of the high- and low-scoring groups would have a discrimination value close to 1. Items with both high item difficulties ($p < 0.25$) and low item difficulties ($p > 0.75$) generally yield low item discrimination values. The frequency that each item option was chosen by students was also determined using the polytomous data set, represented as a percentage.

Rasch model analysis. All analysis using the Rasch model utilized the polytomous data set, imported into the Winsteps program,[189] using the Partial Credit

Model.[185] Estimates for item difficulty, student ability, person reliability index and residuals for analysis of unidimensionality, and local independence were all obtained using Winsteps.

There are many benefits to including Rasch model analysis into the psychometric evaluation of assessment data. First, the Rasch Model is capable of transforming raw testing scores (ordinal scale) into ability measures (interval scale). Having estimates of student ability on an interval scale with invariant spacing is important when comparing differences in student abilities. The interval scale created by the Rasch model has units of log odds, more commonly referred to as logits. Second, the estimates of item difficulty from Rasch analysis are sample independent and can be compared between different samples. Third, item difficulty estimates made by the Rasch model are on the same interval scale as student ability estimates, such that item targeting can be evaluated. That is, how well item difficulties match student abilities of the target population. Lastly, as shown in equation 4, the probability of a student, *n*, answering an item, *i*, correctly, $P_{ni}(x_{ni} = 1)$, can be estimated using only student ability estimates (B_n) and item difficulty estimates (D_i).

$$P_{ni}\left(x_{ni} = 1\right) = \frac{\exp\left[B_n - D_i\right]}{1 + \exp\left[B_n - D_i\right]} \tag{4}$$

Thus, for an item of difficulty, D_i, a student with a higher ability is more likely to answer the item correctly, when compared to a student of lower ability. How well the model predicts student responses based on ability and difficulty estimates is given by two fit statistics, outfit and infit. Both fit statistics are chi-squared statistics divided by the

degrees of freedom, producing a mean-squared statistic (MNSQ) that has an expected value of 1.00.[189] Having fit statistics for both item difficulty and person ability estimates is unique to the Rasch model and is a powerful tool for evaluating item functioning for a given target population. Items that produce unexpected student responses and students who give unexpected answers can be identified using infit and outfit statistics. Lastly, Rasch model analysis can produce item option probability curves, which can be used to visualize which item options are most likely to be chosen by students of a given ability. This type of analysis can provide discrimination information at the item-option level, along with evidence for reliability of item option-level interpretations. This information is critical for assessments that are designed to make interpretations at the item-option level.

RESULTS AND DISCUSSION

Both CTT and Rasch model analysis were used to collect evidence for the intended uses and interpretations of TCI testing data. These two methods of analysis are complimentary. Both were used to determine if items provide informative diagnostic data, reflecting high-functioning items. An informative, high-functioning item can be described by the following characteristics: (1) all distracters are attractive to a certain proportion of the student sample, (2) item difficulty and discrimination indices are within the targeted range, (3) item difficulty targets student abilities found in the target population, and (4) item data fit the Rasch model. Evaluation of item functioning during beta testing informed the novice response process validity interviews obtained in the prior study[206] and further flagged specific items or item responses for revision or removal from the TCI. Revisions to items were then evaluated in the pilot study, resulting in the final version of the TCI used in the large data collection data set presented in this manuscript. Detailed

psychometric analysis of the final version of the TCI is presented in detail in this section, in addition to information gained from beta and pilot testing.

Evidence for Generalizability and Administration Conditions

The initial 15 TCI items (beta testing version) were developed to incorporate student alternative conceptions identified in student interviews as well as predominant alternative conceptions reported in the literature but not found in our qualitative studies.[206] To evaluate the generalizability of the TCI items, item option response frequency, as a percentage, was calculated. Options that were chosen by less than 10% of students in beta testing samples were flagged for evaluation in student novice response process interviews. Evaluation of these options in some cases led to revision or removal.[206] For pilot and large data collection studies, options with less than 10% response frequencies were evaluated using the Rasch model item option probability curves. If an item option was chosen by less than 10% of students, but was clearly discriminating students of different ability, no revisions were made. If, however, an option displayed no discrimination between students of different ability, it was removed. Examples of item option probability curves are given in later figures and in Appendix C. Many item options that focused on conceptual misunderstandings identified in secondary-school populations were not attractive to students based on quantitative data, nor were they reported as plausible based on qualitative student response process interview data.[206] These conceptual misunderstandings were included in TCI items used in beta testing, even though they were not identified in student interviews to confirm that the majority of students in our target population do not find these conceptual misunderstandings attrac-

tive. This provided evidence for the generalizability of qualitative data collected from a small sample within our target population. More quantitative data from schools of different sizes and geographic regions may provide additional evidence for the generalizability of the TCI.

Generalizability evidence may also include the robustness of an assessment to different administration conditions. Administration conditions include testing environment (e.g., lecture or lab), administrator (e.g., course instructor or teaching assistant), stakes (e.g., evaluation points assigned or voluntary formative assessment), testing instructions (e.g., verbal or written instructions), and time given to complete an assessment. Given that the TCI is intended to be used as a diagnostic formative assessment, beta and pilot testing were administered as a type of formative assessment, being voluntary and having no course evaluation associated with TCI scores. Within the formative assessment administration, both lecture and laboratory recitation testing environments were assessed. In addition, the TCI was used as a quiz for the large data collection sample after instruction on thermochemistry, reflecting a practical application of the TCI in general chemistry classrooms as a summative assessment. Lastly, the time span since students had received instruction on thermochemistry varied from one week (β_2 and large scale data collection) to greater than three months (β_1 and pilot testing). Students who were given the TCI as a quiz, and most likely studied thermochemistry prior to administration, did markedly better on certain items. Specifically, item K (presented in Figure 22) that simply addresses the sign convention of the enthalpy of reaction was answered correctly by 93% of the students when given as a quiz. However, when the TCI was given in formative assessment, item K was only answered correctly by

68% of the students in the pilot study. This difference is most likely due to the combination of students studying before taking the TCI, the time gap between instruction on thermochemistry, and difference in student ability in the two samples. However, most items performed similarly under the varying administration conditions, all having acceptable Rasch fit statistics. This provides evidence that the items on the TCI should perform well in both formative and summative assessment. However, this should be verified by researchers wanting to using the TCI, especially in populations significantly different than those used in the *Journal of Chemical Education*.[195]

Administration of the TCI by TAs in lab recitation is another testing administration variation. Increased error associated with difference in testing environment and administration instructions were a concern. However, there was no pattern of a specific TA having a disproportionate amount of misfitting students, providing evidence that there were not significant TA effects on student TCI performance.

Evidence for Structural Validity

Structural validity is the degree to which the actual test structure matches the designed theoretical structure based on the construct being measured.[116] If a test is designed to be unidimensional (only measuring one construct), it should be established that only one construct is being measured. The TCI was designed to measure one psychological construct, thermochemical conceptual understanding. Thus, the structure of the TCI should be unidimensional, with each item being a measure of thermochemical conceptual understanding. The basic questions in the analysis of test data dimensionality are (1) What is the difference between the observed outcome and the outcome predicted by the Rasch model? (2) If information explained by the Rasch model is removed, is there

any commonality among the remaining residuals? (3) What underlying traits might explain these residuals? Commonly, factor analysis, and specifically, confirmatory factor analysis, is used to verify unidimensionality of testing data.[166, 195] This analysis uses raw non-linear ordinal data, which can show a heavy dependence on the sample used to collect testing data.[187] In addition, there are no fit statistics generally assigned to factor loadings, which can limit the interpretability of this analysis.[166] For these reasons, confirmatory factor analysis was not used to obtain evidence for structural validity. Instead, principle component analysis (PCA) of standardized residuals (information not explained by the Rasch model) was used to address the questions presented above.

The PCA was run in the Winsteps program to analyze the correlated variance of the standardized residuals of items in the TCI not explained by the Rasch model. Items with factor loadings greater than ± 0.4 on a secondary contrast, with an associated eigenvalue greater than 2.00, would be flagged and provide evidence against unidimensionality.[189] Items D and G (see supporting information) both had loadings of 0.67 on the first contrast; however, the eigenvalue of the contrast was only 1.4, less than 2.00 cut-off. Therefore, these items were not seen as a threat to the unidimensionality or structural validity of the TCI. All other items had loading less than ± 0.4.

Additional evidence for structural validity can come from assessing the assumption made by the Rasch model that test items are locally independent. Local independence of items requires that the probability of getting one item correct is independent of the probability of getting another item correct.[189] To verify the local independence of the TCI items the inter-item correlation was evaluated using the same PCA analysis described above. If items display strong correlations ($R > 0.5$) among the standardized residuals,

especially between items without obvious content similarities, the assumption of local independence cannot be confirmed. No two items in the TCI had inter-item correlation values greater than 0.19. This provides evidence for the validity of the assumption of local independence and for structural validity.

Psychometric Evidence for Item Quality: Item Fit, Targeting, and Functioning

Psychometric analysis of TCI items utilizes both CTT and Rasch model analysis. Given that the TCI was shown to meet the criteria of unidimensionality and local independence, use of the Rasch model to evaluate item quality is appropriate.

Initial evaluation of Thermochemistry Concept Inventory items and item responses: Classical Test Theory Analysis. Raw TCI testing data were used to create a dichotomous data set that was used for CTT analysis. Item difficulty (*p*) and discrimination (*D*) estimates are shown in Figure 18.

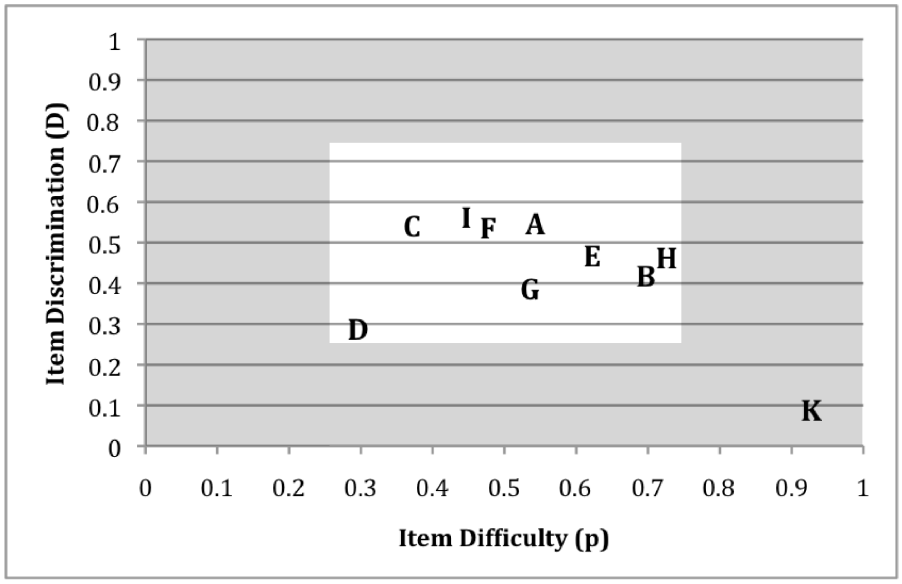


Figure 18. Classical Test Theory estimates of item discrimination and difficulty. Grey area indicates values outside of targeted range of 0.25-0.75.[207]

Targeted difficulty and discrimination values are between 0.25 and 0.75,[207] shown as the rectangular region in Figure 18. Of the 10 TCI items, nine fall within this region, while only one item is outside. The discrimination value shows dependence on the item difficulty value, such that items with poor difficulty estimates ($0.25 < p > 0.75$) generally do not discriminate students of different ability and have low discrimination estimates. Item K was by far the easiest item on the TCI ($p = 0.93$), which translates to a very poor discrimination estimate ($D = 0.09$). This item was added to the TCI after beta testing, as an easier item was needed to target students of the lowest ability. Item K has the simplest item structure of all TCI items and only provides information on students knowledge of sign convention of reaction enthalpy in relation to the definition of exothermic and endothermic reactions. Given that item K can be answered correctly simply by using memorized knowledge, it is not surprising that the majority of students taking the TCI as a quiz would answer this item correctly. However, for formative assessment, knowing if students are using the correct sign convention for reaction enthalpy is essential. Thus, item K is functioning as intended (being the easiest item), but does not provide much information when the TCI is used as a quiz.

Evidence for item fit to the Rasch model. Raw TCI testing data were used to make a polytomous data set that was analyzed using the Winsteps program. Item difficulty measure estimates calculated using the Rasch model have the same ordering as item difficulty estimates calculated by CTT, shown in Table 7. Rasch item difficulty measures are on a linear logit interval scale, which is useful for comparing to student ability measures that are on the same scale. In addition, each item difficulty measure has associated fit indices, used to evaluate how well student item responses fit the Rasch

model. Analysis of data fitting to the Rasch model focuses on identifying observations that are outliers to the data set and on unexpected response patterns in observations. Identifying outliers, using outfit statistics, was the first step in the analysis of TCI data, followed by identification of unexpected response patterns, using infit statistics.

Both outfit and infit are chi-squared statistics and are reported with associated Z-statistics to assess statistical significance.[166] Outfit is calculated by summing the square of standardized residuals for either all responses by an individual (student ability) or all responses to an item (item difficulty), and taking the average.[189]

Table 7. Item-level Psychometric Estimates for Both Clasical Test Theory and Rasch Model; Items Ordered from Hardest (item D) to Easiest (item K).

	Classical Test Theory		*Rasch model*		
Item	*Difficulty*[a]	*Discrimination*	*Difficulty Measure*[b]	*Infit MNSQ*[c]	*Outfit MNSQ*[c]
D	0.29	0.29	1.51	1.13	1.23
C	0.36	0.54	1.37	0.94	0.93
I	0.44	0.57	0.68	0.95	0.93
F	0.47	0.54	0.52	0.98	1.00
G	0.53	0.39	0.28	1.08	1.09
A	0.55	0.55	0.16	0.95	0.93
E	0.62	0.47	-0.22	1.00	0.97
B	0.69	0.42	-0.74	0.98	0.97
H	0.72	0.47	-0.96	0.93	0.84
K	0.93	0.09	-2.60	1.10	1.56

[a]The larger the value, the easier the item.
[b]The more negative the value, the easier the item.
[c]Acceptable range for MNSQ values is 1.00 ± 0.5.[189, 208]

When the average is divided by the degrees of freedom, the result is a MNSQ, which is reported by Winsteps.[189] The MNSQ values have an expect value of 1.00 and have a range from 0 to infinity. However, MNSQ values of 1.00 ± 0.5 are generally acceptable, and values of 1.00 ± 0.3 are used as more stringent evaluation criteria.[166, 189] Every MNSQ value has an associated Z-standardized statistic (ZSTD) to assess statistical significance. The MNSQ values with ZSTD values greater than ±2.00 represent statistically significant ($p > 0.05$) values.[189] However, note that for large data sets, ZSTD values increase due to increased statistical power and should be evaluated only after observations displaying MNSQ misfit have been identified.[189] Conceptually, outfit is sensitive to outliers, which is good for identifying outlying observations, but outfit is also easily skewed by these observations. Issues that are identified by poor outfit are generally easy to diagnose and easy to address; thus, outfit is normally the first fit statistic evaluated. For example, high outfit MNSQs (>1.5), can result from low-ability students correctly answering items above their ability level. One way students can correctly answer an item above their ability is by guessing the correct response. The outfit MNSQ statistics shown in Table 7 for all items, except item K (outfit MNSQ = 1.56) , are acceptable. No ZTSD statistics are given in Table 7, as the large sample size decreases the utility of this statistic for diagnostic purposes.

The infit MNSQ has reduced sensitivity to outliers displayed by the outfit statistic.[189] The infit statistic is calculated the same as outfit, but is weighted by the statistical information (model variance) of observations.[189] This model variance is larger for observations where the Rasch model should provide an accurate prediction (e.g., when a student's ability is close to an item's difficulty) and smaller for extreme observations

(e.g., when a student's ability is much more or less than an item's difficulty).[166] This makes infit sensitive to inlier observations that display an unexpected response pattern. Observations with misfitting infit statistics are more complex and more difficult to diagnose. High infit MNSQs (> 1.5) can result from items that are well-targeted to student ability, but poorly predict observed outcomes.[189] Determining why an item is misbehaving based on item infit values is much more difficult, because it could involve some component of the item construction or some part of a student's response process. These generally cannot be answered solely by Rasch analysis.[166] Items with poor infit statistics can be evidence against response process validity and should be evaluated using qualitative research methods. As shown in Table 7, all items on the TCI had infit statistics well within the acceptable range. This is strong statistical evidence for response process validity. Item K is not seen as problematic in terms of response process validity as its infit value was acceptable.

Evidence for item targeting. Student ability estimates can be calculated by transforming TCI raw scores onto a logit interval scale. Thus, every student with the same raw score on the TCI will have the same Rasch ability estimate. The difference being that the spacing between student ability measures are invariant and can be compared to item difficulty measures. An assessment is most informative when item difficulties are matched with student abilities. Given that a student population has a distribution of abilities, item difficulties should also vary to sample students at different ability levels. An easy way to evaluate item targeting for a sample is to plot Rasch student ability and item difficulty measures on the same logit axis, commonly known as a Wright Map. The Wright Map for the large data collection sample is shown in Figure 19.

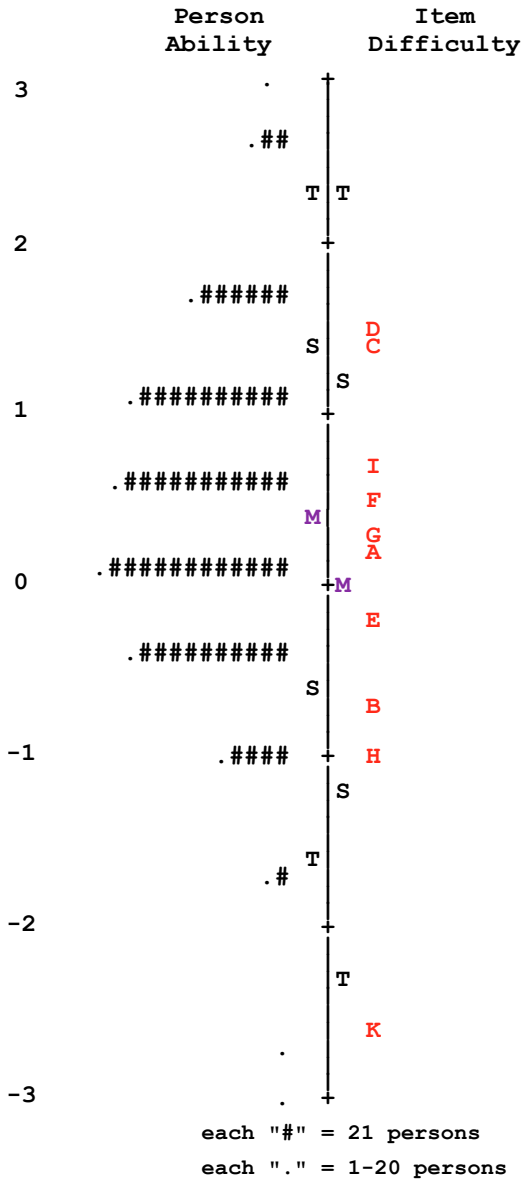


Figure 19. Wright map of item person ability and item difficulty plotted on a logit scale. M = mean, S = one standard deviation, T = two standard deviations.

The mean of the item difficulty measures is centered at 0 logits and can be compared to the mean of student ability measures. When the means of item difficulty and student ability are close to one another and the spread of item difficulties covers the range

of student abilities, this is an indication of good test targeting. The TCI items display excellent targeting to the population studied, where all items except for one (item K) are well matched with the majority of the student abilities. This provides evidence that thermochemistry content tested by TCI items varies in difficulty and can provide targeted information for a range of student abilities. For this sample, the average student ability is greater than zero logits, with a proportion of students with abilities above the item with the greatest item difficulty (item D). When there are no items above a student's ability measure, there can be a threat to reliability of the student ability estimate. However, in pilot testing, which used the TCI in formative assessment, the average student ability was just below zero logits and item D was well targeted to the students with the greatest ability measures. Perfect item targeting is difficult to obtain for samples with varying average abilities and different stakes of testing. Nevertheless, item targeting was satisfactory in both low (formative assessment) and higher-stakes (quiz) testing. Item targeting is also evidence for the item reliability. Items that have difficulty measures close to the average student ability have high item reliability estimates. Thus, items around the center of the TCI scale (e.g., items I, F, G, A, E) will inherently provide the most reliable measures, when compared to an item that is at the extreme of the TCI scale (e.g., item K). This can help test users place confidence in item-level interpretations of TCI testing data.

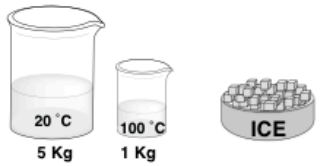
Evidence for item functioning: Rasch option probability curves. Distracter analysis of TCI items was used for both the item development process and in the evaluation of the final version of the TCI items used in the large data collection study. Item option probability curves (OPCs) provide a visual representation of the attractiveness (probability of choosing a given response) of item options based on student ability, as

shown in Figure 20. Rasch item OPCs can be used to evaluate researcher expectations of option attractiveness to students of certain abilities, which is critical for diagnostic distracter-driven concept inventories.[209] For example, the correct answer should have a low probability of being chosen by students of the lowest ability (as estimated by the Rasch model) and should increase in probability as student ability increases. If this is not demonstrated in an item OPC, then there might be a threat to response process validity. Likewise, certain distracters representing specific alternative conceptions that deviate from the correct conception should be more likely chosen by lower-ability students than higher-ability students. In addition, item OPCs can be used to evaluate how item options can discriminate students of different abilities. Item A (shown in Figure 20) is an example of an item with good item distracter functioning.

For item A, each response option targets as specific student-ability range. This is important for the evaluation of options B and C (item A), which were only selected by 8% and 11% of the student sample, respectively. Options that seem unattractive to students based simply on response frequency might actually be functioning well, if they are attractive to a small portion of the sample that has a specific ability range. Therefore, item OPCs can help address if low option response is simply error due to student guessing or is providing information about an conceptual misunderstanding in a portion of the student population.

Item A

Two beakers of differing volumes contain pure water at different temperatures. Ice is added to the water in each beaker. Choose the most accurate answer given below.



(A) Equal amounts of ice will melt in each beaker

(B) The water is considered the system because it is giving off heat

(C) The melting of the ice in either container is considered an exothermic process

(D) More ice will melt in the beaker with water at 100 °C

Item A			
Item Option	**Count**	**%**	**Rasch Average Ability**
A	706	55	0.82
B	101	8	-0.39
C	142	11	-0.29
D	343	27	0.02

CTT Difficulty	0.546
CTT Discrimination	0.551
Rasch Difficulty Measure	0.16

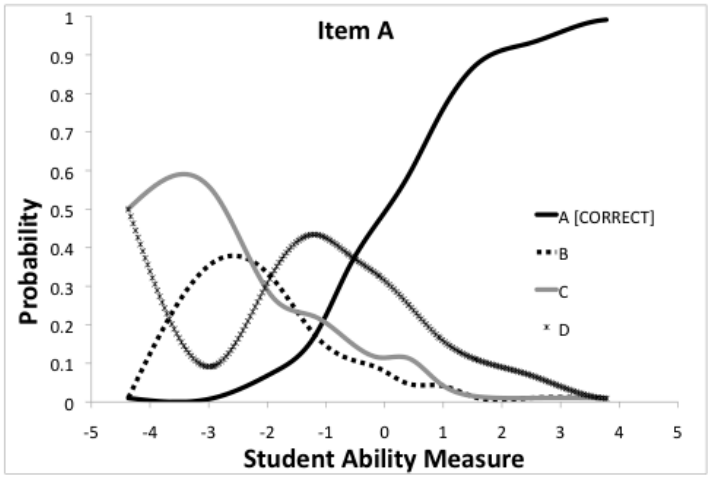


Figure 20. Psychometric information for item A for both item-level (Classical Test Theory difficulty and discrimination, Rasch difficulty measure) and item option-level (option count and frequency, Rasch average ability of students choosing option).

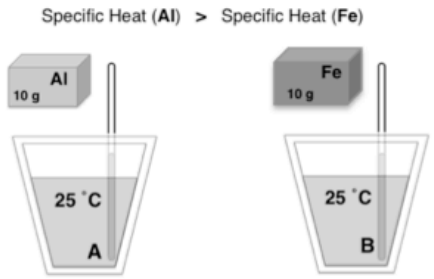
An additional key feature can be seen in the OPC of item A, that is, that as student ability increases, the probability of choosing any distracter decreases and the probability of choosing the correct option increases.

For item C (Figure 21), both item option response frequency and the item OPC demonstrate that option A is not attractive to students, based on the extremely low response frequency that displayed no discrimination of student abilities. The correct answer was the most probable answer for students of higher ability, as this item was the second most difficult item on the TCI. In contrast, analysis of the two other distracters (options B and C) demonstrate discrimination of students based on ability. Option B was the most probable answer for students of the lowest ability, and represents the conceptual misunderstanding that the rate of thermal energy transfer can be determined using the thermal properties of materials (e.g., specific heat capacity). Option C was the most probable for students of average ability and represents the conceptual misunderstanding that heat can be quantified as a specific amount of energy possessed by a body with temperature being a measure of that quantity.[85, 87]

Based on both CTT and Rasch analysis, item C has acceptable difficulty and discrimination values, acceptable Rasch item fit statistics, and informative OPC with the exception of option A. This provides evidence that option A should be removed from item C, but that the item should be included in the final version of the TCI.

Item C

A block of Aluminum (Al) and a block of Iron (Fe) each at 50 ˚C are simultaneously dropped into identical styrofoam cups containing the same amount of water at 25 ˚C water. Choose the most accurate answer given below.



(A) After adding either block to the water, the process can be described as an endothermic process, with respect to the block

(B) Thermal energy will be transferred faster between the Al block and the water than between the Fe block and the water

(C) The final temperature of the water in both A and B will be the same

(D) The water in A will have a higher final temperature than the water in B

Item C			
Item Option	**Count**	**%**	**Rasch Average Ability**
A	41	3	-0.11
B	343	27	-0.10
C	513	40	0.22
D	395	31	1.09
CTT Difficulty	0.364		
CTT Discrimination	0.54		
Rasch Difficulty Measure	1.37		

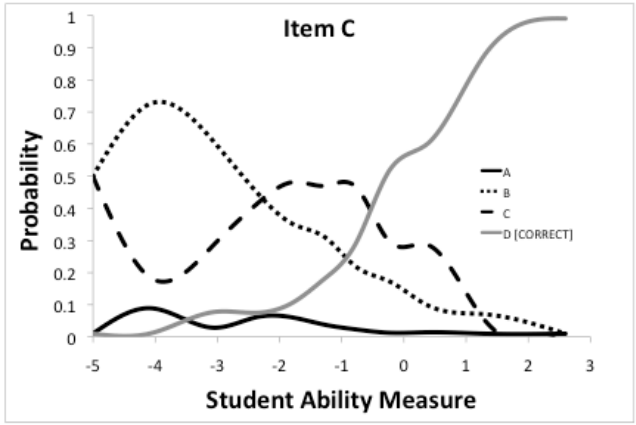


Figure 21. Item C and associated psychometric information demonstrates that option A is unattractive for students (3% option frequency) and does not discriminate among students based on ability (OPC).

Psychometric estimates and item OPCs for all items can be found in the supporting information (see Appendix C). In addition to the removal of option A from item C, option A from item H was also removed for the same reasons as presented above. In contrast, option C of item K also had a poor-performing option, as shown in Figure 22, however, it was retained in the final version of the TCI. Removing this option from item K could increase the threat to validity (construct-irrelevant easiness)[112, 206] by providing information that could be used by students to answer other items, specifically that the reaction enthalpy can be used to determine if a reaction is endothermic or exothermic. By keeping option C, this threat to validity can be minimized, even if this option itself does not provide much useful information.

Reliability of the Intended Interpretations and uses of Thermochemistry Concept Inventory Testing Data

Estimating reliability is situation specific.[210] Just as validity cannot be summarized by one coefficient, evidence for reliability can come from multiple sources and take multiple forms.[112] For the TCI, the reliability of student item-option responses used to diagnose the use of conceptual misunderstandings requires that students find the same option the most attractive under independent, identical administration conditions. Therefore, evidence for reliability for the intended interpretations and uses of TCI testing data are derived from item-level analysis of measurement error, rather than measurement error related to the test score (e.g., Cronbach's alpha). Item fit statistics calculated from the Rasch model are one source of evidence for reliability. Of the 10 TCI items, nine display good fit to the Rasch model, with the exception of item K. Interpretations of item K are not advisable when the TCI is used as a quiz, as the error associate with testing data

from this item are much greater than for the other TCI items. In addition to Rasch fit statistics, item OPCs provide evidence for reliability. Item options that can discriminate students of different abilities decrease item option-level error and increase precision.[112] Item OPCs demonstrate that the majority of TCI item options are attractive to a specific range of student ability; where for some items, each option is the most probable for a specific student ability range (see Figure 20).

The Rasch model also has a reliability index that can provide additional evidence for reliability: the person reliability index. The person reliability index indicates the ability of the instrument to distinguish high and low ability students. The person reliability index is equivalent to traditional test reliability, estimated by Cronbach's alpha values.[189] However, Rasch person reliability indexes generally do not include persons with extreme scores, which inflate this estimate. Traditional reliability estimates do use persons with extreme scores and generally have a higher reliability estimate when compared to the Rasch person reliability index.[189] Just as with traditional reliability measures, the person reliability index increases with an increase in the number of items.[166] The person reliability index for the TCI is 0.32, which is low and not surprising given the short length of the test. Less emphasis is placed on the person reliability index, as it reflects the error in student abilities as measured by the TCI. Given the emphasis on using item option responses rather that the total score on the TCI, it is less critical that the TCI has a high person reliability index.[112] However, this does place more responsibility on the researchers to ensure that all TCI items are high functioning and can be used to make valid interpretations of student conceptual misunderstandings.

Item K

If a reaction has a **positive** reaction enthalpy (ΔH_{rxn}), choose the most accurate response below.

(A) The reaction can be described as an **exothermic** process

(B) The reaction can be described as an **endothermic** process

(C) There is **NOT** enough information to determine if the reaction is an exothermic or endothermic process

Item K			
Item Option	**Count**	**%**	**Rasch Average Ability**
A	55	4	-0.48
B	1200	93	0.43
C	37	3	0.32
CTT Difficulty	0.928		
CTT Discrimination	0.089		
Rasch Difficulty Measure	-2.60		

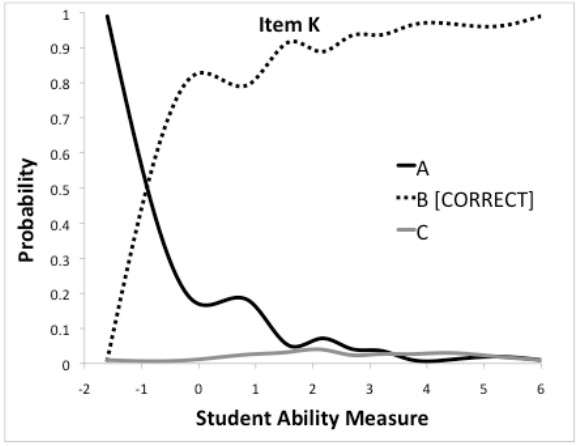


Figure 22. Item K psychometric information demonstrates that option C does not provide useful or reliable information.

Thermochemical Conceptual Understanding as a Measure of Student Ability

The Rasch model uses a student's total TCI score as an estimate of ability. To check this assumption, in addition to the four general chemistry sections (A, B, C & D), the TCI was also administered to an honors section of second-quarter general chemistry. Students in the top 2% of first-quarter general chemistry (by final grade in course) were automatically enrolled into the honors section of second-quarter general chemistry. This provides a unique, independent measure of student ability that can be used to evaluate the ability of the TCI to discriminate students of different abilities. Section-level data are shown in Table 8, where sections A, B, C, and D were combined for statistical comparison to the honors section.

Table 8. Large Scale Data Collection Sample Information.

Section	*N*	*M*	*SD*
A	330	5.55	1.89
B	315	5.55	1.64
C	338	5.71	1.77
D	310	5.55	1.77
Honors	37	7.35	1.45

There was a significant difference between the TCI total score for the general sections ($M = 5.58$, $SD = 1.79$) and the honors section ($M = 7.35$, $SD = 1.46$;

$t(1330) = -7.24$, $p < 0.0001$, two-sided). The magnitude of the difference in the means was small to moderate (eta squared = 0.04). Students in the honors section performed better on all 10 items of the TCI, with an average increase in item difficulty of 0.18 ($SD = 0.15$) and a range of 0.08 (item K) to 0.48 (item C). This provides evidence that the TCI total score can distinguish students with marked difference in ability.

In summary, evidence for the reliability of the intended interpretations and uses of TCI testing data demonstrates that all but one item on the TCI (item K) produce reliable student- and item-level data. In addition, the total score on the TCI can differentiate students of different abilities. This provides evidence that the construct measured by the TCI, conceptual understanding in thermochemistry, is a useful measure of student ability in general chemistry.

SUMMARY AND CONCLUSIONS

Evidence for different lines of validity based on the intended uses and interpretations of testing data is critical for evaluation of an assessment instrument. This evidence should be easily understood by the target test user, which for the TCI is general chemistry instructors. Given the complex nature of some of the statistical analysis used to collect evidence for structural and response process validity, care was taken to provide explanations to allow readers to evaluate this evidence, especially evidence derived from analysis using Rasch model analysis. The lines of evidence for validity are not mutually exclusive. A strong validity argument will demonstrate how different lines of validity are coherent and can inform one another.[204] For example, evidence for content validity should be used in the evaluation of evidence for structural validity.

The use of beta and pilot testing of the TCI items provided invaluable information that compliment evidence collected from qualitative studies. These studies together led to the final version of the TCI. Psychometric evaluation of the TCI using CTT and the Rasch model provided evidence for structural validity and response process validity. Of the 10 items on the TCI administered to the large data collection sample, only item K had unsatisfactory psychometric properties. However, this item was designed to be an easy item and is needed for proper item targeting of student abilities when the TCI is used in formative assessment and in samples that are of lower ability. In addition, two item options, option A of item C (see Figure 21) and option A for item H (see Appendix C), were unattractive to students of all abilities and were removed from the TCI. These are the only changes made to create the final version of the TCI.

Validity as a Compass for Assessment Development

The TCI was designed to be a short and informative diagnostic instrument to provide both students and instructors accurate information about student conceptual misunderstandings related to thermochemistry. Limiting the test length required an emphasis on getting information from each item option, rather than simply from each item. This required psychometric analysis of individual items and item responses, both in the development process and in the evaluation of a final version of the test. The use of the Rasch model allowed for high-resolution analysis of item-level testing data to evaluate the accuracy (validity) and precision (reliability) of interpretations and uses of testing data at the item-response level. The TCI total test score was used to estimate student ability and can differentiate students based on ability, with regard to thermochemical conceptual

understanding. The honors student section did significantly better on the TCI than the large-enrollment sections, providing evidence that the TCI total score can differentiate students with differences in ability determined by external criteria.

Potential Users of the Thermochemistry Concept Inventory

Chemistry instructors and chemical education researchers interested in using the finalized version of the TCI should contact the corresponding author for an electronic copy. In addition to the finalized version, a detailed answer key has been made to use in formative assessment to provide students detailed explanations why each distracter is incorrect and to explain which conceptual misunderstandings is associated with each incorrect answer.

CHAPTER VI

SUMMARY, CONCLUSIONS, AND FUTURE RESEARCH

SUMMARY: ADDRESSING RESEARCH QUESTIONS

The findings reported in Chapters IV and V are be summarized by answering the research questions proposed at the beginning of this study.

- Q1 What conceptual misunderstandings in thermochemistry are students using in college-level general chemistry classrooms?
  - Q1a What thermochemistry topics are taught in most general chemistry classrooms?
  - Q1b Of these topics, which are classified as most important by practicing general chemistry instructors?
  - Q1c What conceptual misunderstandings (from the important topics) are being used by students?

To answer Research Question Q1, each sub-question needed to be addressed first. Results from a thermochemistry topic survey given to general chemistry instructors (experts) presented in Table 4 address Research Question Q1b. Topics chosen to be included in the thermochemistry topic survey came from the thermochemistry chapters of the most popular texts used in general chemistry classrooms. To evaluate if the majority of topics, rated as important by experts in the topic survey, were primarily taught in the thermochemistry section of general chemistry, rather than simply introduced, a second survey was sent to a larger pool of experts. Surprisingly, topics like understanding–state

function lacked consensus by experts (see Appendix B, B2). Some instructors believed understanding the concept was critical for understanding enthalpy, while other instructors did not believe their students could grasp the concept until the second semester of general chemistry. All of the thermochemistry topics with a percent importance rating above 75% were significantly covered in the thermochemistry section by 75% of experts surveyed (39% from the original expert sample responded to this survey). This provided evidence that those topics with high percent importance ratings are primarily taught in the thermochemistry section, and those with low ratings (e.g., understanding–Work), are covered at a different point in the course, if at all.

A literature review revealed student conceptual misunderstandings that were identified in secondary, post-secondary, or instructor populations. A summary of reported conceptual misconceptions and those found in student think-aloud studies are described in Chapter IV. An important finding was that not all alternative conception reports for secondary students were found in the target population of this study. This included absence from student think-aloud interviews, lack of plausibility during novice response process validity interviews, and low response frequency and discrimination (using OPCs). This finding highlights the importance of collecting validity evidence for the target population in addition to utilizing previously published literature focusing on different student populations.

Q2 What development methodology is necessary to create the Thermochemistry Concept Inventory?

Q2a How many items are needed to cover the most important conceptual misunderstandings?

Q2b What is the best format for these items?

Q2c What are the most important criteria to evaluate these items in pilot tests?

A pool of 15 multiple-choice items were initially created to incorporate the most conceptual misunderstandings as item distracters. Some items used similar conceptual misunderstandings but different item stems and chemical context. This allowed for comparison of items using both qualitative and quantitative data, and eventually led to a decision of which items should be included in the final version of the TCI. Table 7 provides a summary of item development, including the number of items at each point in the development process. Both student feedback from novice response process validity interviews and expert feedback from an online survey (Chapter IV) provided evidence for response process validity. This was a critical criteria to evaluate items in the development process, specifically, if the item distracters could accurately predict what conceptual misunderstanding students were using. Item distracters that were not attractive to students (e.g., not plausible based on student interviews or low option response frequency) and/or had characteristics of poor item discrimination based on Rasch option probability curves were identified in the item development and evaluation process. These items or item options were revised and retested or removed from the TCI. All of the initial 15 multiple-choice items had four item options (one correct answer, three distracters) to maximize the number of thermochemical conceptual misunderstandings represented in the TCI. However, some of these distracters were removed; such that, of the final 10 items on the TCI, only four items have the four item options. The number items each option has does not need to be the same, as each item is independent and quantitative analysis techniques do not require items to have the same number of item options. This is not surprising, as a

meta-analysis focusing on the number of item options used in multiple-choice options found that most items function as three-option items, even if more options are present.[211] Thus, three-option multiple-choice items seem to be the best format for the majority of the TCI items.

Q3 How do items in the Thermochemistry Concept Inventory perform in pilot studies?

Q3a How will the performance of items be measured?

Q3b With what student population will item performance be evaluated?

Q3c How will these performance measurements be evaluated?

Q3d What changes to items will be implemented to improve performance?

Q3e How do these changes affect item performance?

Both qualitative (Chapter IV) and quantitative (Chapter V) methods were used to evaluate items, keeping with the intended uses and interpretations of TCI testing data in mind. Novice response process validity interviews provided evidence for generalizability of specific conceptual misunderstandings identified in think-aloud student interviews. Conceptions identified in think-aloud interviews and those published in the literature were not always attractive or plausible to students in the target population (college-level general chemistry students), as represented by TCI item options. Revisions to these items or item options in some cases increased item performance, but not in all cases. In addition, construct-irrelevant variance (construct-irrelevant difficulty or easiness) was determined using qualitative evidence collected in student response process validity. For example, the BDE item shown in Figure 14 required students to use knowledge outside the intended construct to answer the item correctly. This was a threat to validity and an

example of construct-irrelevant difficulty. A new item stem and item options were created (see Figure 16) to mitigate this threat to validity. In addition, expert feedback on TCI items using an online administration platform allowed for evaluation of item wording and consensus of experts on the correct answer. As illustrated in Figure 16, not all experts agreed on the correct answer of the BDE item, which was revised to address their concerns. The other 9 items in the TCI had complete or nearly complete expert consensus of the correct answer and only small concerns over stem or option wording as discussed in Chapter IV.

Quantitative evaluation of item performance included use of CTT measures (item difficulty and discrimination; Figure 18 and Table 7) and Rasch model analysis (item difficulty and item fit; Table 7, item targeting; Figure 19, option probability curves; Figures 20, 21, 22). Item K (Figure 22) demonstrated unsatisfactory CTT item discrimination and difficulty, as well as Rasch item fit (outfit MNSQ) and item discrimination based on option probability curves. However, in pilot studies where the TCI was used in formative assessment, item K performed satisfactorily and was needed to target students of the lowest ability. In contrast, when the TCI was used as a quiz, item K did not provide useful or reliable information. To provide the broadest scope of utility of the TCI in both formative and summative assessment, item K was included in the final version of the TCI. All other items performed very well, both in CTT and Rasch analysis.

Q4 What are the intended uses and interpretations to be made using data collected from the Thermochemistry Concept Inventory, and what evidence is there for the validity and reliability of these uses and interpretations?

Q4a What types of validity need to be established for the uses and interpretations of testing data collected using the Thermochemistry Concept Inventory?

Q4b How will these types of validity be established and with what evidence?

Q4c What threats to validity and reliability are expected?

Q4d How will these threats be mitigated or addressed?

The proposed intended uses and interpretations of TCI testing data are detailed in Chapter IV. Based on these intended uses and interpretations, content validity, response process validity, and structural validity were the most critical sources of evidence for construct validity needed to evaluate two major threats to validity of TCI items: construct-underrepresentation and construct-irrelevant variance. Evidence for content validity was obtained from the expert thermochemistry topic survey and the expert response process validity survey of TCI items (Chapter IV). High percent importance scores (> 75%) accurately reflected topics significantly covered by instructors in the thermochemistry section of most general chemistry courses. Percent importance scores prioritized what conceptual misunderstandings were used as the basis of distracters in the TCI. This provides evidence against the threat to validity of construct underrepresentation of thermochemical conceptual understanding as measured by the TCI.

Evidence for response process validity was collected in two qualitative studies: novice response process validity interviews and expert response process validity online survey. Both studies evaluated how primary stakeholders of the TCI understood, interpreted, and answered TCI multiple-choice items. In both studies, interpretations outside

those expected by the researcher were identified and addressed using item revisions. Some of these revisions were described in detail in Chapter IV.

Evidence for structural validity of the TCI was collected in the large data collection study. The TCI was found to be unidimensional using principle component analysis (PCA) of standardized residuals from the Rasch model. In addition, all TCI items were shown to have minimal inter-item correlation of the standardized residuals from the Rasch model based on PCA analysis, providing strong evidence for local independence of TCI items.

IMPLICATIONS OF THIS RESEARCH

The research described in this study used a collection of evidence for the validity of the interpretations and intended uses of TCI testing data as a central focus in the design, development, and evaluation of TCI items. Currently, there are no other published studies in the chemical education research community that has used such an extensive collection of validity evidence while creating a concept inventory. This research did not set out to be the first to use such a methodology, but it was required to create a short and informative assessment. The scope of thermochemistry topics taught in the college-level general chemistry series required focusing the boundaries of concepts covered in the TCI to only those deemed most important by general chemistry content experts. The use of a percent-importance score to rank topics was a unique way of helping establish boundaries of testing content, providing an accessible, easy to interpret and communicate scores for a broad range of content topics.

A major methodological question that this research hoped to answer was: Would the time commitments required to collect and analyze qualitative data from several qualitative studies yield an assessment with highly informative and high-functioning items? Specifically, could a small initial pool of items (15 items) be revised and culled to 10 to 12 items that maintained the required content coverage and desired psychometric properties? As demonstrated from the results in Chapter V, the final 10 items on the TCI performed extremely well when given as a quiz, with exception of one item (item K). The poor performance of item K was not surprising, as this item was specifically designed to be informative when the TCI is used in formative assessment. The item may provide useful information for samples with a lower average ability when given at a quiz; however, this will need to be evaluated in a future research study. Overall, the instrument methodology used in this research study clearly demonstrated that obtaining feedback from the primary stakeholders of the TCI (general chemistry instructors and students) multiple times during the development process can yield items with psychometric properties that reflect high-functioning and informative items.

The use of the Rasch model during the development of assessment instruments is new to the chemistry education research field. Specifically, Rasch option probability curves (OPCs) have only previously been reported in the science education field for very large data sets.[209] The TCI benefitted from having a relatively small number of items and being administered to relatively large samples of students. In addition, the TCI testing data met the assumptions of the Rasch model (e.g., unidimensionality and local independence), which is not always the case for instrument testing data. However, given the inferential and explanatory power of Rasch analysis and prevalence in the greater

assessment community, this type of analysis may be more common in future CER as more researchers become familiar with the use of this type of probabilistic analysis.

Lastly, the power of using qualitative research to inform quantitative research and vice versa was demonstrated in this study. Limitations in the sample size of participants in qualitative studies were compensated with large sample sizes of quantitative studies. Generalizability of TCI items, specifically whether conceptual misunderstandings incorporated as TCI distracters are attractive to students sampled from our target population, could be evaluated quantitatively. Use of the Rasch model OPCs provided an unmatched level of quantitative resolution, which was extremely helpful in making decisions on revision or removal of item options answered by less than 10% of students. When qualitative student interview data and quantitative item response-level data are congruent, this provides strong evidence for construct validity, specifically response process validity.

FUTURE RESEARCH USING THE THERMOCHEMISTRY CONCEPT INVENTORY

This research study has provided potential users of the TCI with the necessary information required to make decisions on how TCI testing data may be used and interpreted and for what population. The TCI items have acceptable psychometric properties when used in formative assessment (pilot study) and summative assessment (large data collection study) at large research-intensive universities. Using the TCI as a diagnostic instrument to identify students' use of thermochemical conceptual misunderstandings is supported by qualitative and quantitative evidence presented in Chapters IV and V. However, more research should be conducted at smaller, liberal art universities,

which serve students in our target population but who were not a part of either qualitative or quantitative studies of this research.

Given that the TCI has been designed and evaluated for use in either formative or summative assessment, researchers or instructors interested in either identifying student conceptual misunderstandings in thermochemistry (formative assessment) or measuring student conceptual understanding of thermochemical topics (summative assessment) now have an informative tool. For use in formative assessment, the TCI may be used after instruction on thermochemical topics or before instruction of thermodynamics in the general chemistry series. For use in summative assessment, the TCI can be used as a pre-test and a post-test for courses that look to evaluate the effectiveness of new instructional techniques focusing on improving thermochemical conceptual understanding. The TCI can also be used as a general measure of thermochemical conceptual understanding for general chemistry, which may be used to collect evidence for validity with respect to association to other variables (e.g., concurrent, convergent, predictive, or divergent validity).

REFERENCES

[1]Cooper, C.; Pearson, P. A. Genetically Optimized Predictive System for Success in General Chemistry Using a Diagnostic Algebra Test. *J. Sci. Educ. Technol.* **2012,** *21*, 197–205.

[2]McFate, C.; Olmsted, J. III. Assessing Student Preparation through Placement Tests. *J. Chem. Educ.* **1999,** *76* (4), 562–565.

[3]Rixse, J.; Pickering, M. Freshman Chemistry as a Predictor of Future Academic Success. *J. Chem. Educ.* **1985,** *62* (4), 313–315.

[4]Russell, A. A Rationally Designed General Chemistry Diagnostic Test. *J. Chem. Educ.* **1994,** *71* (4), 314–317.

[5]Bird, L. Logical Reasoning Ability and Student Performance in General Chemistry. *J. Chem. Educ.* **2010,** *87* (5), 541–546.

[6]Lewis, S.; Lewis, J. Predicting at-risk students in general chemistry: comparing formal thought to a general achievement measure. *Chem. Educ. Res. Pract.* **2007,** *8* (1), 32–51.

[7]Xu, X.; Lewis, J. Refinement of a Chemistry Attitude Measure for College Students. *J. Chem. Educ.* **2011,** 88, 561–568.

[8]Lewis, S.; Shaw, J.; Heitz, J. Attitude Counts: Self-Concept and Success in General Chemistry. *J. Chem. Educ.* **2009,** *86* (6), 744–749.

[9]Akbas, A.; Kan, A. Affective Factors That Influence Chemistry Achievement (Motivation and Anxiety) and the Power of These Factors to Predict Chemistry Achievement-II. *Journal of Turkish Science Education* **2007,** *4* (1), 10–19.

[10]Clark, G.; Riley, W. The Connection between Success in a Freshman Chemistry Class and a Student's Jungian Personality Type. *J. Chem. Educ.* **2001,** *78* (10), 1406–1411.

[11]Kennepohl, D.; Guay, M.; Thomas, V. Using an Online, Self-Diagnostic Test for Introductory General Chemistry at an Open University. *J. Chem. Educ.* **2010,** *87* (11), 1273–1277.

[12]Shields, S.; Hogrebe, M.; Spees, W.; Handlin, L.; Noelken, G.; Riley, J.; Frey, R. A Transition Program for Underprepared Students in General Chemistry: Diagnosis, Implementation, and Evaluation. *J. Chem. Educ.* **2012,** *89*, 995–1000.

[13]Easter, D. Factors Influencing Student Prerequisite Preparation for and Subsequent Performance in College Chemistry Two: A Statistical Investigation. *J. Chem. Educ.* **2010,** *87* (5), 535–540.

[14]Gabel, D.; Bunce, D. Research on Problem Solving: Chemistry. In Handbook of Research on Science Teaching and Learning; Gabel, D., Ed. MacMillian: New York, 1994.

[15]Zumdahl, S.; Zumdahl, S. *Chemistry*. 8th ed.; Cengage Learning: Independence, KY, 2010.

[16]Jasien, P., What Do You Mean That "Strong" Doesn't Mean "Powerful"? *J. Chem. Educ.* **2011,** *88*, 1247–1249.

[17]Johnstone, A. H. Why is Science Difficult to Learn? Things are Seldom What They Seem. *Journal of Computer Assisted Learning* **1991,** *7*, 75–83.

[18]Kelly, R.; Jones, L. Investigating students' ability to transfer ideas learned from molecular animations of the dissolution process. *J. Chem. Educ.* **2008,** *85*, 303–309.

[19]Corradi, D.; Elen, J.; Clarebout, G. Understanding and Enhancing the Use of Multiple External Representations in Chemistry Education. *J. Sci. Educ. Technol.* **2012,** *DOI:10.1007/s10956-012-9366-z*

[20]Olympiou, G.; Zacharias, Z.; deJong, T. Making the Invisible Visible: Enhancing Students' Conceptual Understanding by Introducing Representations of Abstract Objects in a Simulation. *Instructional Science* **2012,** *DOI:10.1007/s11251-012-9245-2*

[21]Dixon, J.; Emery, A. Jr. Semantics, Operationalism, and the Molecutlar-Statisitical Model in Thermodynamics. *Am. Sci.* 1965, *53*, 428–436.

[22]Feltovich, P.; Spiro, R.; Coulson, R. Learning, Teaching, and Testing for Complex Conceptual Understanding. In *Test Theory for a New Generation of Tests*; Frederiksen, N., Mislevy, R., Bejar, I., Eds.; Lawrence Erlbaum: Hillsdale, NJ, 1993.

[23]Wandersee, J., Mintzes, J.; Novak, J. Research on Alternative Conceptions in Science. In *Handbook of Research on Science Teaching and Learning*; Gabel, D., Ed.; MacMillan: New York, 1994.

[24]Lloyd, B. A Review of Curricular Changes in the General Chemistry Course during the Twentieth Century. *J. Chem. Educ.* **1992,** *69* (8), 633–636.

[25]Spencer, J. N. General Chemistry Course Content. *J. Chem. Educ.* **1992,** *69* (3), 182–186.

[26]Lloyd, B.; Spencer, J. New Directions for General Chemistry. *J. Chem. Educ.* **1994,** *71* (3), 206–209.

[27]Murphy, K., Holme, T.; Zenisky, A.; Caruthers, H.; Knaus, K. Building the ACS Exams Anchoring Concept Content Map for Undergraduate Chemistry. *J. Chem. Educ.* **2012,** *89*, 715–720.

[28]Cooper, M., The Case for Reform of the Undergraduate General Chemistry Curriculum. *J. Chem. Educ.* **2010,** *87* (3), 2231–2232.

[29]Spencer, J. N. New Approaches to Chemistry Teaching. *J. Chem. Educ.* **2006,** *83* (4), 528–533.

[30]Busch, D.; Nameroff, T. Exploring the Molecular Vision: Report from a SOCED Invitational Conference. *J. Chem. Educ.* **2004,** *81* (2), 177–179.

[31]Spotts, T. Discriminating factors in faculty use of instructional technology in higer education. *Educational Technology & Society* **1999,** *2* (4), 92–99.

[32]Pienta, N. Bring Your Own Containers: Educating an Open Mind. *J. Chem. Educ.* **2011,** *88*, 1447–1448.

[33]Cook, E., Kinnetz, P.; Owens-Misner, N. Faculty Perceptions of Job Rewards and Instructional Development Activities. *Innovative Higher Education* **1990,** *14* (2), 123–130.

[34]Jabker, E.; Halinski, R. Instructional Development and Faculty Rewards. *Journal of Higher Education* **1978,** *49* (4), 316–328.

[35]Anderson, T. R. Bridging the Educational Research-Teaching Practice Gap: The Importance of Bridging the Gap between Science Education Research and its Application in Biochemistry Teaching and Learning: Barriers and Strategies. *Biochem. Mol. Biol. Educ.* **2007,** *35* (6), 465–470.

[36]Henderson, C. Promoting instructional change in new faculty: An evaluation of the Physics and Astronomy New Faculty Workshop. *Am. J. Phys.* **2008,** *76* (2), 179–187.

[37]Silverthorn, D.; Thorn, P.; Svinicki, M. It's difficult to change the way we teach: Lessons from the Integrative Themes in Physiology Curriculum Module Project. *Advances in Physiology Education* **2006,** *30*, 204–214.

[38]Bauer, C. Beyond "Student Attitudes": Chemistry Self-Concept Inventory for Assessment of the Affective Component of Student Learning. *J. Chem. Educ.* **2005,** *82* (12), 1864–1870.

[39]Bauer, C. Attitude towards Chemistry: A Semantic Differential Instrument for Assessing Curriculum Impacts. *J. Chem. Educ.* **2008,** *85* (10), 1440–1445.

[40]Sanger, M.; Phelps, A. What Are Students Thinking When They Pick Their Answers? A Content Analysis of Students' Explanation of Gas Properties. *J. Chem. Educ.* **2007,** *84* (5), 870–874.

[41]Metz, P.; Smith, K. J. Evaluating Student Understanding of Solution Chemistry through Microscopic Representations. *J. Chem. Educ.* **1996,** *73* (3), 233–235.

[42]Naah, B.; Sanger, M. Student misconceptions in writing balanced equations for dissolving ionic compounds in water. *Chem. Educ. Res. Pract.* **2012,** *13*, 186–194.

[43]Wagner, E.; Sasser, H.; DiBiase, W. Predicting Students at Risk in General Chemistry Using Pre-semester Assessments and Demographic Information. *J. Chem. Educ.* **2002,** *79* (6), 749–755.

[44]Ausubel, D., *Educational Psychology: A Cognitive View*. Holt, Rinehart and Winston, Inc.: New York, 1968.

[45]Hovey, N.; Krohn, A. An evaluation of the Toledo chemistry placement examination. *J. Chem. Educ.* **1963,** *40* (7), 370–372.

[46]Nurrenbern, S.; Pickering, M. Concept Learning versus Problem Solving: Is There a Difference. *J. Chem. Educ.* **1987,** *64* (6), 508–510.

[47]Cornog, J.; Stoddard, G. Predicting Performance in Chemistry. *J. Chem. Educ.* **1925,** *2* (8), 701–708.

[48]Pickering, M. Helping the High Risk Freshman Chemist. *J. Chem. Educ.* **1975,** *52* (8), 512–514.

[49]Spencer, H. Mathematical SAT Test Scores and College Chemistry Grades. *J. Chem. Educ.* **1996,** *73* (12), 1150–1153.

[50]Ozsogomonyan, A.; Loftus, D. Predictors of general chemistry grades. *J. Chem. Educ.* **1979,** *56* (3), 173–175.

[51]Johnstone, A. H. You Can't Get There from Here. *J. Chem. Educ.* **2009,** *87* (1), 22–29.

[52]Cassel, J. R. T.; Johnstone, A. The Effect of Language on Student Performance on Multiple Choice Tests in Chemistry. *J. Chem. Educ.* **1984,** *61* (7), 613–615.

[53]Johnstone, A. Chemical education research in Glasgow in perspective. *Chem. Educ. Res. Pract.* **2006,** *7* (2), 49–63.

[54]Johnstone, A.; Selepeng, A. Language Problem Revisited. *Chem. Educ. Res. Pract.* **2001,** *2* (1), 19–29.

[55]Sachs, J. S. Recognition memory for syntactic and semantic aspects of connected discourse. *Perception and Psychophysics* **1967,** *2* (9), 437–442.

[56]Miller, G. The Magical Number Seven, Plus or Minus Two: Some Limits on Capacity for Processing Information. *The Psychological Review* **1956,** *63*, 81–97.

[57]Johnstone, A. H. Topic difficulties in chemistry. *Educ. Chem.* **1971,** *36*, 45–48.

[58]Johnstone, A. H.; Sleet, R. J.; Vianna, J. F. An Information Processing Model of Learning: Its Application to an Undergraduate Laboratory Course in Chemistry. *Studies in Higher Education* **1994,** *19* (1), 77–87.

[59]Johnstone, A. H. Chemistry Teaching–Science or Alchemy? *J. Chem. Educ.* **1997,** *74* (3), 262–268.

[60]Reid, N. A Scientific Approach to the Teaching of Chemistry. *Chem. Educ. Res. Pract.* **2008,** *9*, 51–59.

[61]National Research Council. *How People Learn: Brain, Mind, Experience, and School*, Expanded ed.; Committee on Developments in the Science of Learning; John, B., Brown, A., Cocking, R., Eds.; Commission on Behavioral and Social Sciences and Education, National Academy Press: Washington, DC, 2000.

[62]Cooper, M.; Sandi-Urena, S. Design and Validation of an Instrument To Assess Metacognitive Skillfulness in Chemistry Problem Solving. *J. Chem. Educ.* **2009,** *86* (2), 240–245.

[63]Rickey, D.; Stacy, A. The Role of Metacognition in Learning Chemistry. *J. Chem. Educ.* **2000,** *77* (7), 915–920.

[64]National Research Council. *Knowing What Students Know: The Science and Design of Educational Assessment*. Committee on the Foundations of Assessment; James, P., Chudowsky, N., Glaser, R., Eds.; Board on Testing and Assessment, Center for Education. Division of the Behavioral and Social Sciences and Education, National Academy Press: Washington, DC, 2001.

[65]Piaget, J. *The Principles of Genetic Epistemology*. Basic Books: New York, 1972.

[66]Bodner, G. Constructivism: A Theory of Knowledge. *J. Chem. Educ.* **1986,** *63* (10), 873–878.

[67]Bodner, G. Twenty Years of Learing: How to Do Research in Chemical Education. *J. Chem. Educ.* **2004,** *81* (5).

[68]Bretz, S. Nova's Theory of Education: Human Constructivism and Meaningful Learning. *J. Chem. Educ.* **2001,** *78* (8), 1107–1117.

[69]Anderson, T. R.; Schonborn, K. J. Bridging the Educational Research–Teaching Practice Gap: Conceptual Understanding, Part 1: The Multifaceted Nature of Expert Knowledge. *Biochem. Mol. Biol. Educ.* **2008,** *36* (4), 309–315.

[70]White, R.; Gunstone, R. *Probing Understanding*. The Falmer Press: New York, 1992.

[71]Ralya, L.; Ralya, L. Some Misconceptions in Science Held By Prospective Elementary Teachers. *Sci. Educ.* **1938,** *22* (5), 244–251.

[72]Azizoglu, N.; Alkan, M.; Geban, O. Undergraduate Pre-Service Teachers' Understandings and Misconceptions of Phase Equilibrium. *J. Chem. Educ.* **2006,** *83* (6), 947–959.

[73]Beall, H. Probing Student Misconceptions in Thermodynamics with In-Class Writing. *J. Chem. Educ.* **1994,** *71* (12), 1056–1057.

[74]Boo, H. K. Students' Understanding of Chemical Bonds and the Energetics of Chemical Reactions. *J. Res. Sci. Teach.* **1998,** *35* (5), 569–581.

[75]Brook, A.; Briggs, H.; Bell, B.; et al. *Aspects of Secondary Students' Understanding of Heat: Full Report*; The University of Leeds, Centere for Studies in Science Education and Mathematics Education: Leeds, UK, 1988.

[76]Cakmakci, G. Identifying Alternative Conceptions of Chemical Kinetics amoung Secondary School and Undergraduate Students in Turkey. *J. Chem. Educ.* **2010,** *87* (4).

[77]Cakmakci, G.; Aydogdu, C. Designing and evaluating an evidence-informed instruction in chemical kinetics. *Chem. Educ. Res. Pract.* **2011,** *12*, 15–28.

[78]Carson, E. M.; Watson, J. R. Undergraduate Students' Understanding of Enthalpy Change. *U. Chem. Educ.* **1999,** *3* (2), 46–51.

[79]Cheung, D. Using Think-aloud Protocols to Investigate Secondary School Chemistry Teachers' Misconceptions about Chemical Equilibrium. *Chem. Educ. Res. Pract.* **2009,** *10*, 97–108.

[80]Clark, D. Longitudinal Conceptual Change in Students' Understanding of Thermal Equilibrium: An Examination of the Process of Conceptual Restructuring. *Cognition and Instruction* **2006,** *24* (4), 467–563.

[81]Cooper, M., Grove, N.; Underwood, S.; Klymkowsky, M. Lost in Lewis Structures: An Investigation of Student Difficulties in Developing Representational Competence. *J. Chem. Educ.* **2010,** *87* (8), 869–874.

[82]Erickson, G.; Tiberghien, A. Heat and Temperature. In *Children'' Ideas in Science*, Driver, R., Guesne, E., Tiberghien, A., Eds.; Open University Press: Milton Keynes, London, 1985.

[83]Hadfield, L.; Wieman, C. Students Interpretations of Equations Related to the First Law of Thermodynamics. *J. Chem. Educ.* **2010,** *87* (7), 750–755.

[84]Harrison, A.; Grayson, D.; Treagust, D. Investigating a Grade 11 Student's Evolving Conceptions of Heat and Temperature. *J. Res. Sci. Teach.* **1999,** *36* (1), 55–87.

[85]Kesidou, S.; Duit, R. Students' Conceptions of the Second Law of Thermodynamics—An Interpretive Study. *J. Res. Sci. Teach.* **1993,** *30* (1), 85–106.

[86]Lewis, E.; Lin, M. Heat Energy and Temperature Concepts of Adolescents, Adults, and Experts: Implications for Curricular Improvements. *J. Res. Sci. Teach.* **2003,** *40*, S155–S175.

[87]Meltzer, D., Investigation of Students' Reasoning Regarding Heat, Work, and the First Law of Thermodynamics in an Introductory Calculus-Based General Physics Course. *Am. J. Phys.* **2004,** *72* (11).

[88]Mulford, D., Robinson, W., An Inventory for Alternate Conceptions among First-Semester General Chemistry Students. *J. Chem. Educ.* **2002,** *79* (6).

[89]Sanger, M., Campbell, E.; Felker, J.; Spencer, C. "Concept Learning versus Problem Solving": Does Particle Motion Have an Effect? *J. Chem. Educ.* **2007,** *84* (5), 875–879.

[90]Sozbilir, M.; Bennett, J. Turkish Prospective Chemistry Teachers' Misunderstandings of Enthalpy and Spontaneity. *Chem. Educator* **2006,** *11*, 355–363.

[91]Taber, K.; Watts, K. Learners' Explanation for Chemical Phenomena. *Chem. Educ. Res. Pract.* **2000,** *1* (3), 329–353.

[92]Taber, K. *Chemical Misconceptions—Prevention, Diagnosis and Cure*. Royal Society of Chemistry: London, 2002; Vol. Volume 1: Theoretical background.

[93]Thomas, P.; Schwenz, R. College Physical Chemistry Students' Conceptions of Equilibrium and Fundamental Thermodynamics. *J. Res. Sci. Teach.* **1998,** *35* (10), 1151–1160.

[94]Thomaz, M.; Malaquias, I.; Valente, M.; et al. An Attempt to Overcome Alternative Conceptions Related to Heat and Temperature. *Phys. Educ.* **1995,** *30* (19), 19–26.

[95]Towns, M.; Kraft, A. *Review and synthesis of research in Chemical Education from 2000-2010*; The National Academies National Research Council Board of Science Education: West Lafayette, IN, 2011.

[96]Posner, G.; Strike, K.; Hewson, P.; et al., Accommodation of a Scientific Conception: Toward a Theory of Conceptual Change. *Sci. Educ.* **1982,** *66* (2), 211–227.

[97]Chinn, C.; Brewer, W. An Empirical Test of a Taxonomy of Responses to Anomalous Data in Science. *J. Res. Sci. Teach.* **1998,** *35* (6), 623–654.

[98]Hestenes, D.; Wells, M.; Swackhamer, G. Force Concept Inventory. *Phys. Teach.* **1992,** *30*, 141–158.

[99]Limbarkin, J. *Concept Inventories in Higher Education Science*; Washington, DC, 2008.

[100]Barbera, J.; VandenPlas, J. All Assessment Material Are Not Created Equal: The Myths about Instrument Development, Validity, and Reliability. In *In Investigating Classroom Myths through Research on Teaching and Learning*, Bunce, D., Ed. American Chemical Society: Washington, DC, 2011.

[101]Schonborn, K.; Anderson, T. Bridging the Educational Research—Teaching Practice Gap: Conceptual Understanding, Part 2: Assessing and Developing Student Knowledge. *The International Union of Biochemistry and Molecular Biology* **2008,** *36* (5), 372–379.

[102]Leite, L. Heat and Temperature: An Analysis of How These Concepts are Dealt with in Textbooks. *Eur. J. Teach. Educ.* **1999,** *22* (1), 75–88.

[103]National Research Council. *Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering;* Singer, S., Nielsen, N., Schweingruber, H., Eds.; Committee on the Status, Contributions, and Future Directions of Discipline-Based Education Research, Board on Science Education, Division of Behavioral and Social Science and Education; National Academy Press: Washington, DC, 2012.

[104]Streveler, R.; Miller, R.; Santiago-Roman, A., et al. Rigorous Methodology for Concept Inventory Development: Using the 'Assessment Triangle' to Develop and Test the Thermal and Transport Science Concept Inventory (TTCI). *Int. J. Eng. Educ.* **2011,** *27* (5), 968–984.

[105]Taber, K. Finding the Optimum Level of Simplification: The Case of Teaching about Heat and Temperature. *Phys. Educ.* **2000,** *35* (5).

[106]Johnstone, A. Chemical Education Research: Where from Here? *U. Chem. Educ.* **2000,** (4), 1.

[107]Committee on Undergraduate Science Education; Board on Science Education; Division of Behavioral and Social Sciences and Education; National Research Council. *Science Teaching Reconsidered: A Handbook;* National Academy Press: Washington, DC, 1997.

[108]Bodner, G. Why Changing The Curriculum May Not Be Enough. *J. Chem. Educ.* **1992,** *69* (3), 186–190.

[109]Libarkin, J. Concept Inventories in Higher Education Science. In *Promising Practices in Undergraduate STEM Education Workshop 2*; National Research Council: Washington, DC, 2008.

[110]Mestre, J. *Learning Goals in Undergraduate STEM Education and Evidence for Achieving Them;* White Paper for Board of Science Education, National Research Council, National Academies of Science: Washington, DC, 2008.

[111]Hake, R., Interactive-Engagement versus Traditional Methods: A Six-Thousand-Student Survey of Mechanics Test Data for Introductory Physics Courses. *Am. J. Phys.* **1998,** *66* (1), 64–74.

[112]American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing;*. American Educational Research Association: Washington, DC, 1999.

[113]Haladyna, T. *Development and Validating Multiple-Choice Test Items*. 3rd ed.; Lawrence Erlbaum: Mahwah, NJ, 2004.

[114]Crotty, M. *The Foundations of Social Research;* Sage: London, 2003.

[115]American Educational Research Association, A. P. A., & National Council on Measurement in Education. *Technical recommendations for psychological tests and diagnostic techniques*; American Educational Research Association: Washington, DC, 1954.

[116]Loevinger, J. Objective Tests As Instruments of Psychological Theory. *Psychol. Rep.* **1957,** *3* (Monograph Supplement 9), 635–694.

[117]Furr, R. M.; Bacharach, V. *Psychometrics: An Introduction*; Sage: Thousand Oaks, CA, 2008.

[118]Cronbach, L.; Meehl, P. Construct Validity in Psychological Test. *Psychological Bulletin* **1955,** *52* (4), 281–302.

[119]Vonsiadou, S. Conceptual Change Research: An Introduction. In *International Handbook of Research on Conceptual Change*; Routledge: New York, 2008.

[120]Chi, M. Three Types of Conceptual Change: Belief Revision, Mental Model Transformation, and Categorical Shift. In *International Handbook of Research on Conceptual Change*, Vonsiadou, S., Ed.; Routledge: New York, 2008; pp 61–82.

[121]diSessa, A. Bird's-Eye View of the "Pieces" vs. "Coherence" Controversy (From the "Pieces" Side of the Fence). In *International Handbook of Research on Conceptual Change;* Vosniadou, S., Ed.; Routledge: New York, 2008.

[122]Vonsiadou, S.; Vamvakoussi, X.; Skopeliti, I. The Framework Theory Approach to the Problem of Conceptual Change. In *International Handbook of Research on Conceptual Change;* Vosniadou, S., Ed.; Routledge: New York, 2008.

[123]Herron, D. Piaget for chemists. Explaining What "good" Students Cannot Understand. *J. Chem. Educ.* **1975,** *52* (3), 146–150.

[124]Bodner, G.; Klobuchar, M. The Many Forms of Constructivism. *J. Chem. Educ.* **2001,** *78*.

[125]Bunce, D. Does Piaget Still Have Anything to Say to Chemists? *J. Chem. Educ.* **2001,** *78* (8), 1107.

[126]Driver, R. Pupils' Alternative Frameworks in Science. *Eur. J. Sci. Educ.* **1981,** *3* (1), 93–101.

[127]Mayer, R. Rote versus Meaningful Learning. *Theory into Practice* **2002,** *41* (4), 226–232.

[128]Guilford, J. P. Three Faces of Intellect. *American Psychologist* **1959,** *14* (8), 469–479.

[129]Bloom, B. Engelhart, M. *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook 1: Cognitive Domain;* Longmans: New York, 1956.

[130]Krathwohl, D. A Revision of Bloom's Taxonomy: An Overview. *Theory into Practice* **2001,** *41* (4), 212–218.

[131]Anderson, L.; Krathwohl, D. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives;* Longman: New York, 2001.

[132]Holme, T.; Murphy, K. Assessing Conceptual and Algorithmic Knowledge in General Chemistry with ACS Exams. *J. Chem. Educ.* **2011,** *88*, 1217–1222.

[133]Ignaz, C.; Proksa, M. Conceptual Questions and Lack of Formal Reasoning: Are They Mutually Exclusive? *J. Chem. Educ.* **2012,** *89* (10), 1243–1248.

[134]Nakhleh, M. Are Our Students Conceptual Thinkers or Algorithmic Problem Solvers? *J. Chem. Educ.* **1993,** *70* (1), 52–55.

[135]Sanger, M. Evaluating Students' Conceptual Understanding of Balanced Equations and Stoichiometric Ratios Using a Particulate Drawing. *J. Chem. Educ.* **2005,** *82* (1), 131–134.

[136]Thagard, P. *Conceptual Revolutions*. Princeton University Press: Princeton, NJ, 1992.

[137]Treagust, D. Development and Use of Diagnostic Tests to Evaluate Students' Misconceptions in Science. *Int. J. Sci. Educ.* **1988,** *10* (2), 159–169.

[138]Vygotsky, L. *Thought and Language*; Wiley: New York, 1962.

[139]Pushkin, D. Introductory Students, Conceptual Understanding, and Algorithmic Success. *J. Chem. Educ.* **1998,** *75* (7), 809-810.

[140]Abimola, I. O. The Problem of Terminology in the Study of Student Conceptions in Science. *Sci. Educ.* **1988,** *72* (2), 175–184.

[141]Nakhleh, M. Why Some Students Don't Learn Chemistry. *J. Chem. Educ.* **1992,** *69* (3).

[142]Taber, K. *Chemical Misconceptions—Prevention, Diagnosis and Cure, Volume 1: Theoretical Background*; Royal Society of Chemistry: London, 2002.

[143]Villafane, S.; Loertscher, J.; Minderhout, V., et al. Uncovering Students' Incorrect Ideas about Foundational Concepts for Biochemistry. *Chem. Educ. Res. Pract.* **2011,** *12*, 210–218.

[144]Abimola, I. O.; Baba, S. Misconceptions & Alternative Conceptions in Science Textbooks: The Role of Teachers as Filters. *The American Biology Teacher* **1996,** *58* (1), 14–19.

[145]Blosser, P. Science Misconceptions Research and Some Implications for the Teaching of Science to Elementary School Students. *ERIC/SMEAC Science Education Digest* **1987,** *1*, 3–4.

[146]Driver, R.; Easley, J. Pupils and Paradigms: A Review of Literature Related to Concept Development in Adolescent Science Students. *Studies in Science Education* **1978,** *5*, 61–84.

[147]Clough, E.; Driver, R. Secondary Students' Conceptions of the Conduction of Heat: Bringing Together Scientific and Personal Views. *Phys. Educ.* **1985,** *20*, 176–182.

[148]Nachmias, R. A Microcomputer-based Diagnostic System for Identifying Students' Conception of Heat and Temperature. *Int. J. Sci. Educ.* **1990,** *12* (2), 123–132.

[149]Sozbilir, M.; Pinarbasi, T.; Canpolat, N. Prospective Chemistry Teachers' Conceptions of Chemical Thermodynamics and Kinetics. *Eurasia Journal of Mathematics, Science & Technology Education* **2010,** *6* (2), 111–121.

[150]Greenbowe, T.; Meltzer, D. Student Learning of Thermochemical Concepts in the Context of Solution Calorimetry. *Int. J. Sci. Educ.* **2003,** *25* (7), 779–800.

[151]Carson, E. M.; Watson, J. R. Undergraduate Students' Understandings of Entropy and Gibbs Free Energy. *U. Chem. Educ.* **2002,** *6*, 4–12.

[152]Teichert, M.; Stacy, A. Promoting Understanding of Chemical Bonding and Spontaneity through Student Explanation and Integration of Ideas. *J. Res. Sci. Teach.* **2002,** *39* (6), 464–496.

[153]Kuhn, T. *The Structure of Scientific Revolutions;*. University of Chicago Press: Chicago, 1962.

[154]diSessa, A. A History of Conceptual Change Research: Threads and Fault Lines. In *The Cambridge Handbook of: The Learning Sciences*; Sawyer, K., Ed.; Cambridge University Press: New York, 2006; pp 265–281.

[155]Duit, R.; Treagust, D.; Widodo, A. Teaching Science for Conceptual Change: Theory and Practice. In *International Handbook of Research on Conceptual Change;* Vosniadou, S., Ed.; Routledge: New York, 2008.

[156]White, R.; Gunstone, R. The Conceptual Change Approach and the Teaching of Science. In *International Handbook of Research on Conceptual Change;* Vosniadou, S., Ed.; Routledge: New York, 2008.

[157]Driver, R.; Erickson, G. Theories-In-Action: Some Theoretical and Empirical Issues in the Study of Students' Conceptual Framework in Science. *Studies in Science Education* **1983,** *5*, 61–84.

[158]Jonassen, D. Model Building for Conceptual Change. In *International Handbook of Research on Conceptual Change;* Vosniadou, S., Ed.; Routledge: New York, 2008.

[159]Klymkowsky, M.; Garvin-Doxas, K. Recognizing Student Misconceptions through Ed's Tools and the Biology Concept Inventory. *PLoS Biology* **2008,** *6* (1), 14–17.

[160]American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. *Standards for Educational and Psychological Tests and Manuals*; American Educational Research Association: Washington, DC, 1966.

[161]Shaw, E.; Korbin, J.; Patterson, B.; Mattern, K. *The Validity of the SAT for Predicting Cumulative Grade Point Average by College Major*; College Board: New York, 2012.

[162]Thorndike, R. The Analysis and Selection of Test Items. In *Personnel Selection; Test and Measurement Techniques;* John Wiley: Oxford, England, 1949.

[163]Crocker, L.; Algina, J. *Introduction to Classical and Modern Test Theory*. Holt; Rinehart and Winston: New York, 1986.

[164]Gadermann, A.; Guhn, M.; Zumbo, B. Estimating Ordinal Reliability for Likert-Type and Ordinal Item Response Data: A Conceptual, Empirical, and Practical Guide. *Practical Assessment, Research & Evaluation* **2012,** *17* (3), 1–13.

[165]Kelly, T. The Selection of Upper and Lower Groups for the Validation of Items. *Education Psychology* **1939,** *30*, 17–24.

[166]Bond, T.; Fox, C. *Applying the Rasch Model*. 2nd ed.; Routledge: New York, 2007.

[167]Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danmarks Paedagogiske Institut: Copenhagen, Denmark, 1960.

[168]Wright, B. Logits? *Rasch Measurement Transactions* **1993,** *7* (2), 288.

[169]Masters, G.; Wright, B. The Essential Process in a Family of Measurement Models. *Psychometrika* **1984,** *49* (4), 529–544.

[170]Brown, T.; LeMay, E.; Bursten, B., et al. *Chemistry: The Central Science*. 12th ed.; Prentice Hall: Upper Saddle River, NJ, 2012.

[171]McMurry, J.; Fay, R. *Chemistry*, 5th ed.; Pearson: Upper Saddle River, NJ, 2010.

[172]Silberberg, M. *Chemistry: The Molecular Nature of Matter and Change,* 6 ed.; McGraw-Hill: New York. 2012.

[173]Tro, N. *Chemistry: A Molecular Approach*, 1st ed.; Pearson: Upper Saddle River, NJ, 2010.

[174]Bowen, C. Think-Aloud Methods in Chemistry Education. *J. Chem. Educ.* **1994,** *71* (3), 184–190.

[175]Beatty, P.; Willis, G. Research Synthesis: The Practice of Cognitive Interviewing. *Public Opin. Q.* **2007**, 1–25.

[176]van Someren, M.; Barnard, Y.; Sandberg, J. *The Think Aloud Method: a Practical Guide to Modelling Cognitive Processes*; Academic Press: London, 1994.

[177]Schoenfeld, A. Making Sense of "Out Loud" Problem-Solving Protocols. *The Journal of Mathematical Behavior* **1985,** *4*, 171–191.

[178]Griffard, P. the Two-Tier Instrument on Photosynthesis: What Does it Diagnose? *Int. J. Sci. Educ.* **2001,** *23* (10), 1039–1052.

[179]Merriam, S., *Qualitative Research*; Jossey-Bass: San Francisco, 2009.

[180]QSR International. *NVivo qualitative data analysis software*, version 8; QST: Burlington, MA, 2008.

[181]Willis, G. *Cognitive Interviewing: A "How To" Guide. Reducing Survey Error Through Research on the Cognitive and Decision Processes in Surveys*. Presented at the Meeting of the American Statistical Association; Research Triangle Institute; Triangle Park, NC, 1999.

[182]Nitko, A., Developing Multiple-Choice Items: Basic Principles. In *Educational Tests and Measurement: An Introduction*; Harcourt Brace Jovanovich: New York, 1983; pp 189-214.

[183]Qualtrics. *Research Suite;* Qualtrics: Provo, UT, 2005.

[184]SPSS, Inc. *PASW Statistics 18;* SPSS: Chicago, 2009.

[185]Masters, G. A Rasch Model for Partial Credit Scoring. *Psychometrika* **1982,** *47* (2), 149–174.

[186]Linacre, J. Comparing "Partial Credit" and "Rating Scale" Models. *Rasch Measurement Transactions* **2000,** *14* (3), 768.

[187]Wright, B. Conventional Factor Analysis vs. Rasch Residual Factor Analysis. *Rasch Measurement Transactions* **2000,** *14* (2), 753.

[188]Linacre, J. Detecting Multidimensionality: Which Residual Data-Type Works Best? *Journal of Outcome Measurement* **1998,** *2* (3), 266–283.

[189]Linacre, J. M. *Winsteps (Version 3.70.0) [Software]*, Beaverton, OR, 2010.

[190]Linacre, J. Unidimentional Models in a Multidimensional World. *Rasch Measurement Transactions* **2009,** *23* (2), 1209.

[191]Wright, B. Comparing Factor Analysis and Rasch Measurement. *Rasch Measurement Transactions* **1994,** *8* (1), 350.

[192]Wright, B. Anchoring and Standard-Errors. *Rasch Measurement Transactions* **1993,** *6* (4), 259.

[193]Bunce, D. Survey Development. *J. Chem. Educ.* **2008,** *85* (10), 1439.

[194]Bretz, S. Qualitative Research Designs in Chemistry Education Research. In *Nuts and Bolts of Chemical Education Research*; Bunce, D., Cole, Renee, Eds.; American Chemical Society: Washington, DC, 2008.

[195]Arjoon, J.; Xu, X.; Lewis, J. Understanding the State of the Art for Measurement in Chemistry Education Research: Examining the Psychometric Evidence. *J. Chem. Educ.* **2013,** *DOI:10.1021/ed3002013*

[196]Rich, S. What do pupils know of chemistry when they begin to study it? *J. Chem. Educ.* **1925,** *2* (8).

[197]Yeo, S.; Zadnik, M. Introductory Thermal Concept Evaluation: Assessing Students' Understanding. *Phys. Teach.* **2001,** *39*.

[198]Chu, H.-E.; Treagust, D.; Yeo, S., et al. Evaluation of Students' Understanding of Thermal Concepts in Everday Contexts. *Int. J. Sci. Educ.* **2012,** *34* (10), 1509–1534.

[199]Sreenivasulu, B.; Subramaniam, R. University Students' Understanding of Chemical Thermodynamics. *Int. J. Sci. Educ.* **2012,** *DOI:10.1080/09500693.2012.683460*

[200]Prince, M.; Vigeant, M.; Nottis, K. Development of the Heat and Energy Concept Inventory: Preliminary Results on the Prevalence and Persistence of Engineering Students' Misconceptions. *Journal of Engineering Education* **2012,** *101* (3), 412–438.

[201]Messick, S. Validity of Psychological Assessment: Validation of Inferences From Persons' Responses and Performance as Scientific Inquiry Into Score Meaning. *American Psychologist* **1995,** *50* (9), 741–749.

[202]Messick, S. Meaning and Values in Test Validation: The Science of Ethics and Assessment. *Educational Researcher* **1989,** *18* (5), 5–11.

[203]Fleiss, J. *Statistical Methods for Rates and Proportions: The Measurement of Interrater Agreement*; John Wiley & Sons: New York, 1981.

[204]Wilson, M. *Constructing Measures: An Item Response Modeling Approach.* Lawrence Erlbaum: Mahwah, NJ, 2005.

[205]Barbera, J. A Psychometric Analysis of the Chemical Concept Inventory. *J. Chem. Educ.* **2013,** *DOI:10.1021.ed3004353*

[206]Wren, D., Barbera, J. Methodology for the Design, Development and Qualitative Evaluation of Thermochemistry Concept Inventory Items. *J. Chem. Educ.* **2013,** *submitted*.

[207]Kline, T. *Psychological Testing: A Practical Approach to Design and Evaluation;* Dage: Thousand Oaks, CA, 2005.

[208]Linacre, J. What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Measurement Transactions* **2002,** *16* (2), 878.

[209]Herrmann-Abell, C.; DeBoer, G. Using Distractor-driven Standards-based Multiple-Choice Assessments and Rasch Modeling to Investigate Hierarchies of Chemistry Misconceptions and Detect Structural Problems with Individual Items. *Chem. Educ. Res. Pract.* **2011,** *12*, 184–192.

[210]Meyer, J. P. *Reliability;* Oxford University Press: New York, 2010.

[211]Rodriguez, M. Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research. *Educational Measurement: Issues and Practice* **2005,** *24* (2), 3–13.

APPENDIX A

INSTITUTIONAL REVIEW BOARD APPROVAL

UNIVERSITY of NORTHERN COLORADO
Institutional Review Board (IRB)

December 15, 2010

TO: Megan Babkes Stellino
School of Sport and Exercise Science

FROM: The Office of Sponsored Programs

RE: Exempt Review of *Novice Face Validity Interviews for Thermochemistry Concept Test Items,* submitted by David A. Wren (Research Advisor: Jack Barbera)

The above proposal is being submitted to you for exemption review. When approved, return the proposal to Sherry May in the Office of Sponsored Programs.

I recommend approval.

MStellino 1/4/2011
Signature of Co-Chair Date

The above referenced prospectus has been reviewed for compliance with HHS guidelines for ethical principles in human subjects research. The decision of the Institutional Review Board is that the project is exempt from further review.

IT IS THE ADVISOR'S RESPONSIBILITY TO NOTIFY THE STUDENT OF THIS STATUS.

Comments:
- add videotape to announcement & attached revisions
- minor revisions to consent
- email 1/21/11

25 Kepner Hall ~ Campus Box #143
Greeley, Colorado 80639
Ph: 970.351.1907 ~ Fax: 970.351.1934

UNIVERSITY of NORTHERN COLORADO
Institutional Review Board (IRB)

May 13, 2011

TO: Maria Lahman
Applied Statistics and Research Methods

FROM: The Office of Sponsored Programs

RE: Exempt Review of *Thermochemistry Concept Inventory Assessment,* submitted by David A. Wren (Research Advisor: Jack Barbera)

The above proposal is being submitted to you for exemption review. When approved, return the proposal to Sherry May in the Office of Sponsored Programs.

I recommend approval.

M-K. K — June 7, 2011
Signature of Co-Chair — Date

The above referenced prospectus has been reviewed for compliance with HHS guidelines for ethical principles in human subjects research. The decision of the Institutional Review Board is that the project is exempt from further review.

IT IS THE ADVISOR'S RESPONSIBILITY TO NOTIFY THE STUDENT OF THIS STATUS.

Comments: 5-23-11 emailed David MKK

25 Kepner Hall ~ Campus Box #143
Greeley, Colorado 80639
Ph: 970.351.1907 ~ Fax: 970.351.1934

UNIVERSITY of NORTHERN COLORADO
Institutional Review Board (IRB)

January 6, 2012

TO: Gary Heise
School of Sport and Exercise Science

FROM: The Office of Sponsored Programs

RE: Exempt Review of *Expert Face Validity Online Survey for a Thermochemistry Concept Inventory,* submitted by David Wren (Research Advisor: Jack Barbera)

The above proposal is being submitted to you for exemption review. When approved, return the proposal to Sherry May in the Office of Sponsored Programs.

I recommend approval.

Gary D. Heise 26 Jan 2012
Signature of Co-Chair Date

The above referenced prospectus has been reviewed for compliance with HHS guidelines for ethical principles in human subjects research. The decision of the Institutional Review Board is that the project is exempt from further review.

IT IS THE ADVISOR'S RESPONSIBILITY TO NOTIFY THE STUDENT OF THIS STATUS.

Comments: emailed approval; (gave verbal vote)

25 Kepner Hall ~ Campus Box #143
Greeley, Colorado 80639
Ph: 970.351.1907 ~ Fax: 970.351.1934

UNIVERSITY of NORTHERN COLORADO
Institutional Review Board (IRB)

IRB CONTINUATION REVIEW

Project Title: *Identification of Student Conceptions in Thermochemistry for First Semester General Chemistry*
Name of Researcher: David A. Wren (Research Advisor: Jack Barbera)

1. Check one:

___ The stages of this research involving data gathering from, or other contact with, human subjects are complete (or will be completed by the first of next month).

___ The stages of this research involving data gathering from, or other contact with, human subjects were not completed and will not be continued.

✓ I request an additional one year approval period for the data-collection phase of the project.

2. How many subjects participated (or have participated to this point)?
572

3. How many subjects, after providing consent, chose not to participate or dropped out during their participation? Ø

4. Describe any adverse events or unanticipated problems involving risks to subjects or others.
None.

5. Describe any complaints you may have received, or concerns that were expressed, about the research.
None.

6. Please summarize any recent information that has come to your attention regarding risks associated with the type of research you are conducting.
None.

7. Signed informed consent forms must be retained on campus for three years and must be available for IRB review. Where on campus are the informed consent forms for this study being stored?
Ross 3695

UNIVERSITY of NORTHERN COLORADO
Institutional Review Board (IRB)

8. Please submit with this continuation review a copy of the informed consent form used in this research. (If you must duplicate an actual form signed by a subject, please block out that subject's name.) Did the form you used deviate from that which you submitted as part of your original IRB proposal? If so, how?

No.

9. If your original proposal called for debriefing of subjects, was (is) such debriefing completed for all subjects? If not, please explain why.

N/A

10. If your subjects were (are) from 12 to 18 years old, was (is) informed consent obtained from parents/guardians as well as from the participants? If no, please explain why.

N/A

11. If your subjects were (are) from 7 to 11 years old, was (is) assent obtained from the subjects in addition to informed consent from parents/guardians? If no, please explain why.

N/A

12. To request an additional one year approval, submit to Sherry May in the Office of Sponsored Programs (OSP):

* A new IRB proposal if there are substantial changes from the original, e.g., additional variables are to be assessed, there is a large increase in the subjects' participation time, another measurement is being added or there is some other significant change in methodology.
* Only an addendum to the original proposal if there are to be only minor changes, e.g., a few more subjects than anticipated, additional researchers will have access to the data etc.
* If you wish an additional one year approval and there are no changes in your proposal, return only this form.

Researcher's or Research Advisor's signature — Date: 4-1-12

Email address: david.wren@unco.edu

Approved for 1 year — IRB Co-Chair's Signature — Date: 4/22/2012

UNIVERSITY of NORTHERN COLORADO

CONSENT FORM FOR HUMAN PARTICIPANTS IN RESEARCH
UNIVERSITY OF NORTHERN COLORADO

Project Title: Thermochemistry Concept Inventory Assessment
Researcher: David Wren, doctoral student in the chemistry education program
Phone number: (303) 399-1883 Email: wren9464@bears.unco.edu
Research Advisor: Dr. Jack Barbera, Assistant Professor, Department of Chemistry and Biochemistry
Phone Number: (970) 351-2545 Email: jack.barbera@unco.edu

The primary goal of this research project is to create a thermochemistry concept inventory composed of items that produce valid data for first semester general chemistry students. Items for this research are multiple choice questions probing student's conceptual understanding of thermochemistry. The validity of the conclusions we can draw from student's answers on items in our concept inventory is dependent on how much our questions are measuring student conceptual understanding. Thus, test items need to probe commonly held student conceptions using easy to understand and clearly worded stems and multiple choice responses. To obtain the most information from each test item, each response needs to seem attractive to some student population. Additionally, no item should be too easy or too difficult and each multiple choice response within an item should have a probability of being chosen by students.

Any risk associated from participating in this study will be no different than what you may experience in a normal testing situation in a chemistry course. You may feel anxious or frustrated by taking quizzes or tests, but we hope to minimize these feelings because the outcome of taking this concept inventory has no connection with your evaluation in your general chemistry course or your final grade. The benefits to you include reviewing some concepts in thermochemistry and knowing that you are helping to make a very power instructional tool for future general chemistry students. If you decide to let your survey responses be used in this research, your participation will be anonymous and will not affect your grade in the course, either positively or negatively.

Confidentiality will be maintained during the course of data collection and analysis. Signed consent forms will be stored separately from the data so that names cannot be linked to the information collected. Each participant shall have a random eight digit code assigned to them for confidentiality and data analysis purposes.

I understand that by signing this consent form I am allowing my responses to this concept inventory to be used in this research study.

Questions: If you have any questions about the design or results of this study, or about the nature of your participation, you may contact the researchers at any time by contacting the researchers using the phone numbers indicated at the top of this form.

Participation is voluntary. You may decide NOT to participate in this study and if you do begin participation you may still decide to stop and withdraw at any time. Having read the above and having had an opportunity to ask any questions, please sign below if you would like to participate in this research. A copy of this form will be given to you to retain for future reference. If you have any concerns about your selection or treatment as a research participant, please contact the Office of Sponsored Programs, Kepner Hall, University of Northern Colorado Greeley, CO 80639; 970-351-2161.

Print name _____

_____ _____
Participant's Signature Date

_____ _____
Researcher's Signature Date

Announcement by Instructors before Administration of Thermochemistry Concept Inventory

Because the researcher will not be able to administer all concept inventories, instructors will need to make a statement to inform students about the research study and their options in regards to participation. The following statement will be provided for each instructor to read to ensure students are adequately informed before administration of the concept inventory.

Instructor Statement

Researchers at the University of Northern Colorado have created a concept inventory to assess general chemistry student's conceptual understanding of key topics in thermochemistry. A concept inventory is a multiple choice test that is used to identify common alternate or misconceptions that students might have in a particular topic. Identification of misconceptions can help instructors design lectures specific for a particular course or class, or to help evaluate improvement in student conceptual understanding after a particular subject is taught.

Today you will be taking the thermochemistry concept inventory, which is 10 questions long and will take about 20 minutes to complete. Everyone will take the concept inventory, but your responses will only be used in this research study if you sign the consent form on the first page of the inventory. Your responses will help create the final version of this useful instrument and are essential in evaluating the quality of each question and the inventory as a whole. Your participation in the research will be anonymous and your performance will not be used in any type of evaluation for this course.

It is important that you answer each question to the best of your ability, as this will provide the most useful information to the researchers conducting this study. In addition, we will go over the answers for the inventory questions, so you can see if you hold any misconceptions in thermochemistry.

I will now pass out the test and you will have 20 minutes to complete the 10 questions.

APPENDIX B

SUPPORTING INFORMATION FOR QUALITATIVE STUDENTS

B1. One page of the online thermochemistry topic survey administered to experts.

Introduction

Below you will find a list of topics regarding **THERMOCHEMISTRY**.

For topics that you typically teach during the first semester of the two semester sequence, please rank the importance of the topics using the following scale:

1. Important
2. Slightly Important
3. Neutral
4. Slightly Unimportant
5. Unimportant

A comment box is provided after each question, feel free to insert notes to the researchers regarding any of the topics or subtopics.

An additional comment box is provided at the end of the survey to gather any thoughts about other topics you think we should add.

A working understanding (definition) of the following terms:

State Functions

NOT Covered Important 1 2 3 4 5 Unimportant Comments?

Temperature

NOT Covered Important 1 2 3 4 5 Unimportant Comments?

B2. Expert sample participation and institutional classification.

Institutional Carnegie Classification	Content Topic Survey Expert Participation (n)	Expert Response Process Validity Expert Participation (n)	Expert Participation in Both Studies (n)
RU/VH: Research University (very high research activity)	10	6	5
RU/H: Research University (high research activity)	3	1	1
Master's L: Master's Colleges and Universities (larger programs)		2	
DRU: Doctoral/Research University	2		
Bac/Diverse: Baccalaureate Colleges— Arts & Science	1	1	
Assoc/Pub-R-L: Associate's— Public Rural-serving Large	2	2	1
Total	**18**	**12**	**7**

B3. Two examples of expert ratings of importance and associated comments. The definition of a state function was not covered by a minority of the sample and had a lack of consensus for those who did cover this topic. This resulted in a low percent importance score. The topic of Le Châtelier's principle was not covered by a minority of experts but was ranked as important or slightly important by the majority of experts. This resulted in a moderate percent importance score. Reasons for not covering this topic mainly focused on covered in second semester general chemistry.

$$\%\ \text{Importance} = \frac{\overbrace{n_1 + n_2}^{\text{Number Important}}}{\underbrace{n_0 + n_1 + n_2 + n_3 + n_4 + n_5}_{\text{Number of Experts in Study}}} \times 100$$

n_0 = not covered
n_1 = important
n_2 = slightly important
n_3 = neutral
n_4 = slightly unimportant
n_5 = unimportant

Definition-State Function

Importance	Respondents	Comment
0 (*Not Covered*)	**7**	I mention what state functions are but don't really go into too much depth
1 (*Important*)	**4**	Critical. Enthalpy cannot be understood without the context of state functions
2	**3**	
3	**2**	I just use brief descriptions so the kids can read the text
4	**1**	
5 (*Unimportant*)	**1**	Most of my students are unable to grasp State Functions, therefore they are not formally covered until the second semester
Total	18	
	% Not Covered = 39%	**% Importance = 39%**

$$\%\ \text{Importance} = \frac{4+3}{7+4+3+2+1+1} \times 100 = 39\%$$

Le Châtelier's Principle

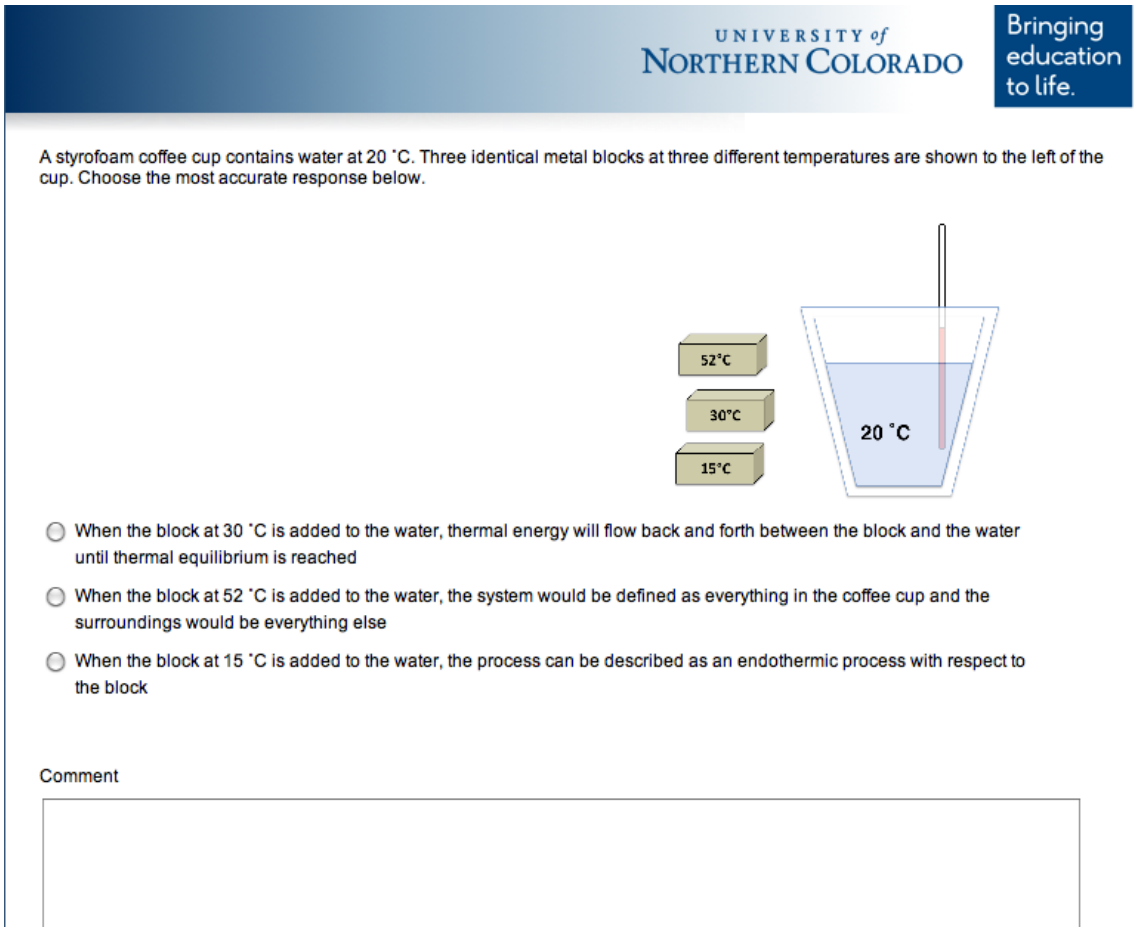
Importance	Respondents	Comment
0 (*Not Covered*)	**4**	(1) Included in equilibrium chapters, but not in thermochemistry chapter (2) Covered 2nd semester
1 (*Important*)	**11**	
2	**2**	Yes, along with an experiment to assist in comprehension
3	**0**	
4	**1**	Casual mentioning of the concept and saved for equilibrium in later chapters
5 (*Unimportant*)	**0**	
Total	18	
	% Not Covered = 22%	**% Importance = 72%**

$$\%\ \text{Importance} = \frac{11+2}{4+11+2+0+1+0} \times 100 = 72\%$$

B4. Student sample participation and institutional classification**.**

Institutional Carnegie Classification	Think-Aloud Interviews (n)	Novice Response Process Validity Interviews (n)
DRU: Doctoral/Research University	8	8
RU/VH: Research University (very high research activity)	4	5
Bac/Diverse: Baccalaureate Colleges—Arts & Science		3
Total	12	16

B5. Example of the item format for the blocks item as seen by experts using the online-interface provide by the Qualtrics survey.



B6. Evidence for response process validity and need for item revision.

Item	Option	Option Response
Blocks	D	If the block at 15 ˚C is added to the water, the process can be described as an endothermic process with respect to the block

Chose correct option; explanation included correct conception

1211_1: [reads option B out loud] I thought that this was a pretty good answer, because it says 15 degrees Celsius (pointing to the block) and 20 degrees Celsius [pointing to the water]. The water is warmer and I'm assuming the warmer temperature would heat up the T3 block. Because it would heat up the block, so thermal energy would be entering the block. And like I said earlier, retaining energy would be endothermic. So I would describe that as an endothermic process with respect to the block.

29311: [reads stem out loud] If the 15 degree block is dropped in the water the process is described as endothermic with respect to the block, because it will take energy into the block to raise the block's temperature up to something less than 20 degree level, depending on how much water is in there [points to the calorimeter].

Eliminated correct option; explanation included intended alternative conception

19411: [reads option D] I thought endothermic was the wrong word to describe . . . I felt like it was trying to talk about a chemical change, rather then just the physical change of temperature between the two [block and water].

Item	Option	Option Response
Blocks	B	Thermal energy will flow back and forth between a block and the water until thermal equilibrium is reached

Chose incorrect distracter; explanation included intended alternative conception

81110: But B, ultimately there is going to be some sort of energy transfer in between the block and the water. And even if they were identical temperatures there's still going to be energy going back and fourth. In the case with B, that's what is going on. Until there is an equilibrium reached, more energy is going from the block to the water, and visa versa, but with the way things work, there is energy going back and fourth, the reactions just going in one particular direction.

18211: [explains item stem correctly] I picked option B.

Researcher: Can you explain your reasoning?

18211: I remembered some time down the road that somebody saying something about they have to reach equilibrium [student gesticulates hands back and forth]. This is why I picked option B, because it made the most sense.

Item	Option	Option Response
BDE	D	For the reaction to occur, 400 kJ/mol needs to be added

Eliminated incorrect distracter; explanation included correct conception

1211_2: [reads option D out loud] No, that's what comes out of the reaction [circling the 468.5 kJ/mol], it doesn't mean what you add to the reaction to make it occur. It's the difference. The change is the difference.

31110: For the reaction to occur, 468.5 kJ/mol need to be added, that isn't necessarily true. Like I said the H_{rxn} is the difference. In order to disassociate water you would need to add however much energy [pointing to the bond enthalpies] on this side of the equation [pointing to the reactant] in order to get water to break up into hydrogen and oxygen.

Chose incorrect distracter; explanation included intended alternative conception

81110: [reads stem out loud] So it's saying this is the reaction [pointing to the balanced chemical equation], water going to create two gases, and in this reaction this much energy is being released [pointing to the H_{rxn}]. No, this much energy is not being released, it is what is required to make this reaction occur. If it were a negative number, it would be how much energy is being released.

81110: I'm going to jump to option D, the option I chose to be correct. [student reads the entire option out loud] That's summarizing what this equation is saying [pointing to balanced chemical equation]. That's just the cleanest answer.

Evidence for a Need to Revise Item

Item	Option	Option Response
BDE	A	For the reaction to occur, 468.5 kJ/mol needs to be added

Extraneous information used to either choose or eliminate item option

1211_2: Yeah, this one was confusing. [student reads stem out loud] At first I started reading through the answer and try to pick one out. But then I noticed the table and thought maybe I should use the table somehow . . . I don't know I just felt like I had to use this table because it was here.

31110: This question wants to know, it tells you that the delta H of reaction for water disassociating to hydrogen and oxygen is given here [points to H_{rxn}], and then choose the answer that best explains the reaction. [Pause] I thought option A was incorrect because if one mole of O2 is made [pointing to the words in option A], 498.6 kJ of energy will be released. And I guess [point to the two different values for the bond enthalpies for single and double-bonded oxygen] that there are two values, here is a double bond [pointing to O=O] between oxygen and here is a single bond between oxygen, so this answer really doesn't specify what it means by "O_2", because I would assume that both of them can be O_2. Just that they have different bond enthalpies.

[8311 used very similar explanation to eliminate option A]

29311: [pointing to option A and the balanced chemical equation] There is one mole produced, and that is not the right number [498.7 kJ], if I draw that out [draws the structural formula for H_2O], it should be 920 kJ if I were to use these bond energies [pointing to BDE for O-H, 460.0 kJ/mol], so I knew option A could not be right.

[1311_2 & 1211_2 both has some miscalculation using BDE from table that was used to eliminate option A]

Did not attempt or answer item

8311: For D, honestly I just didn't feel like finishing the problem. What I thought I had to do was, it says for the reaction to occur blah blah blah needs to be added. I just didn't feel like adding all these bond enthalpies for H_2O, which is what I thought I had to do. I'm not sure if I had to do that or not.

Revision: Replace stem with a generic reaction where no specific bond disassociate energies will be needed to answer the item.

Item	Option	Option Response
Blocks	A	If the 52 ˚C block is put into the cup, the final temperature of the water would be 32 ˚C

Eliminated incorrect distracter because it did not seem plausible and/or reasonable

81110: A, if the 52 degree block is put in the cup. There is not enough information, in the data to come up with 36 degrees. We don't know the mass of this block that would, as one simple thing or how much water is there.

18211: I eliminated A. It says the if the 52 degree block is put in cup the final temperature of the water will be 32 degrees. I didn't think that would be right, because it's just taking 52 minus 20 and saying that's the answer [circling 32 degrees in option]. That might not happen.

Researcher: Is there a reason that might not happen?

18211: I can't say for sure. It could happen, it couldn't happen. You need more information.

1211_1: [correctly explains stem, reads option A out loud] I don't know how it calculated the 32 degrees for the answer. I did know how about to calculating it.

19411: I don't feel like I had enough information [pointing to the 32 degrees in option A] to do the calculation to figure out if 32 degrees was at all right or not. Just thinking about it, it sounds reasonable, but not knowing what the metals are, etc.

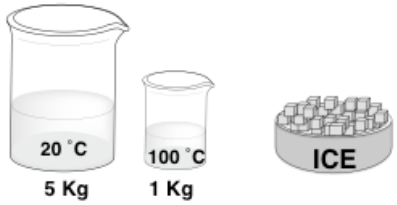
APPENDIX C

SUPPORTING INFORMATION FOR QUANTITATIVE STUDIES

C1. TCI Item Psychometric Summaries

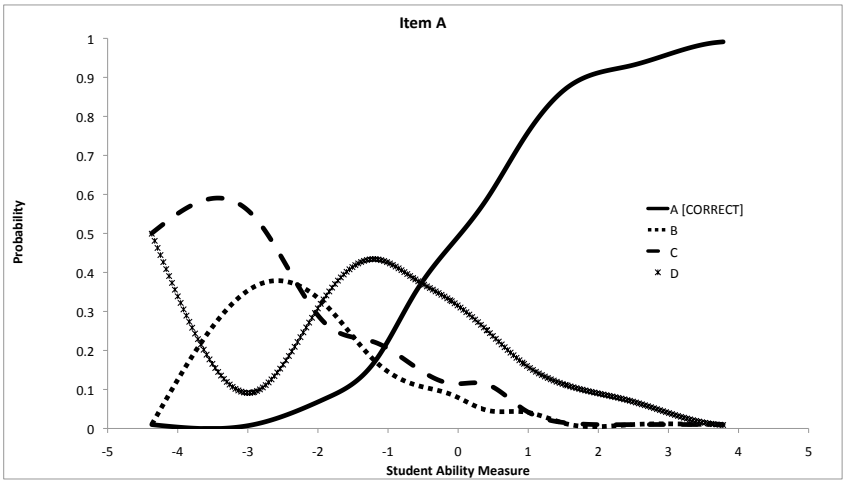
Item A

Two beakers of differing volumes contain pure water at different temperatures. Ice is added to the water in each beaker. Choose the most accurate answer given below.



(A) Equal amounts of ice will melt in each beaker
(B) The water is considered the system because it is giving off heat
(C) The melting of the ice in either container is considered an exothermic process
(D) More ice will melt in the beaker with water at 100 °C

Item A			
Item Option	**Count**	**%**	**Rasch Average Ability**
A	706	55	0.82
B	101	8	-0.39
C	142	11	-0.29
D	343	27	0.02
CTT Difficulty	0.546		
CTT Discrimination	0.551		
Rasch Difficulty Measure	0.16		



Item K

If a reaction has a **positive** reaction enthalpy (ΔH_{rxn}), choose the most accurate response below.

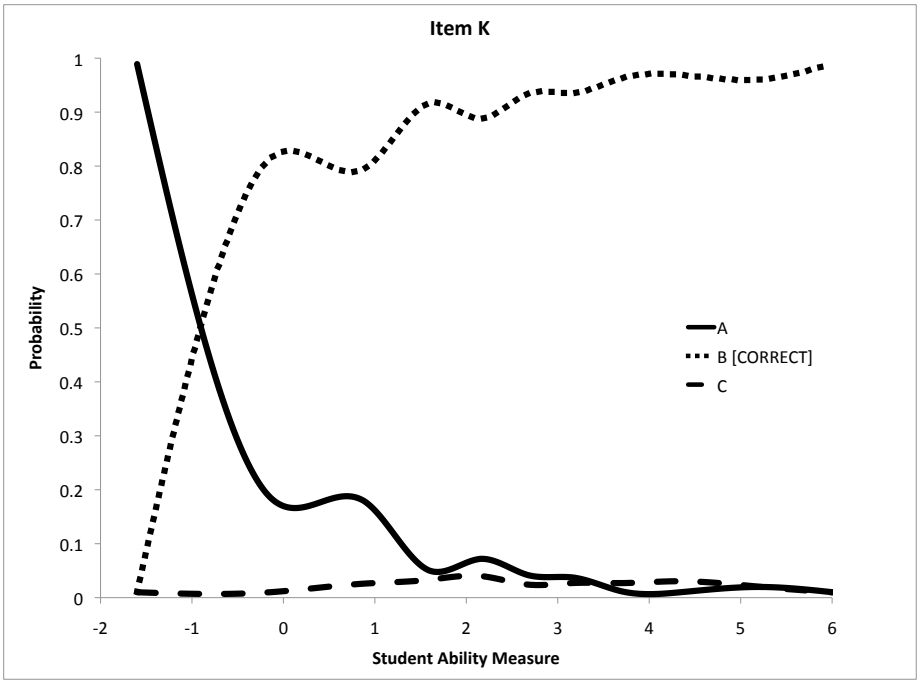
(2) The reaction can be described as an **exothermic** process

(3) The reaction can be described as an **endothermic** process

(4) There is **NOT** enough information to determine if the reaction is an exothermic or endothermic process

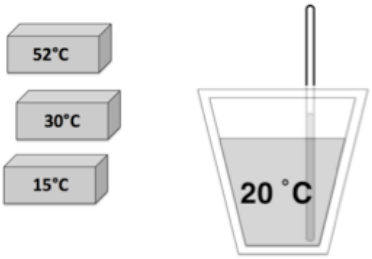
Item K			
Item Option	**Count**	**%**	**Rasch Average Ability**
A	55	4	-0.48
B	1200	93	0.43
C	37	3	0.32

CTT Difficulty	0.928
CTT Discrimination	0.089
Rasch Difficulty Measure	-2.60



Item B

A styrofoam coffee cup contains water at 20 ˚C. Three identical metal blocks at three different temperatures are shown to the left of the cup. Choose the most accurate response below.

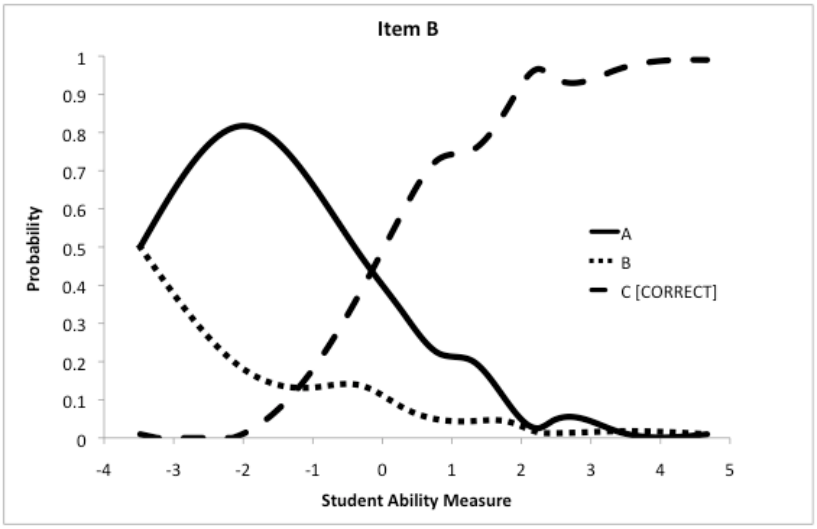


(A) When the block at 30 ˚C is added to the water, thermal energy will flow back and forth between the block and the water until thermal equilibrium is reached

(B) When the block at 52 ˚C is added to the water, the system would be defined as everything in the coffee cup and the surroundings would be everything else

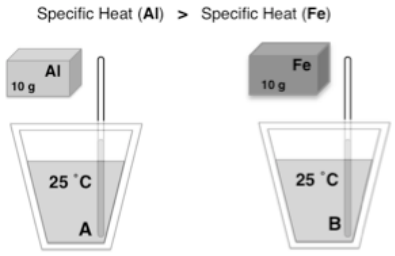
(C) When the block at 15 ˚C is added to the water, the process can be described as an endothermic process with respect to the block

Item B			
Item Option	**Count**	**%**	**Rasch Average Ability**
A	286	22	-0.28
B	75	6	-0.20
C	930	72	0.64
CTT Difficulty	0.689		
CTT Discrimination	0.423		
Rasch Difficulty Measure	-0.74		



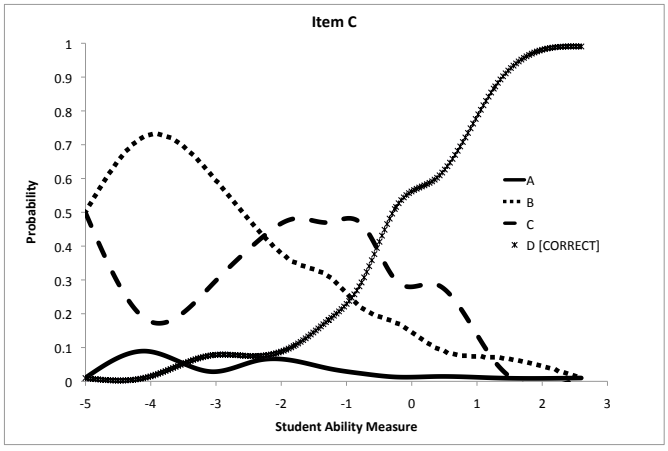
Item C

A block of Aluminum (Al) and a block of Iron (Fe) each at 50 ˚C are simultaneously dropped into identical styrofoam cups containing the same amount of water at 25 ˚C water. Choose the most accurate answer given below.



(A) After adding either block to the water, the process can be described as an endothermic process, with respect to the block
(B) Thermal energy will be transferred faster between the Al block and the water than between the Fe block and the water
(C) The final temperature of the water in both A and B will be the same
(D) The water in A will have a higher final temperature than the water in B

Item C			
Item Option	**Count**	**%**	**Rasch Average Ability**
A	41	3	-0.11
B	343	27	-0.10
C	513	40	0.22
D	395	31	1.09
CTT Difficulty	0.364		
CTT Discrimination	0.54		
Rasch Difficulty Measure	1.37		



Item D

Use the following reaction and the associated standard reaction enthalpy to choose the most accurate answer below.

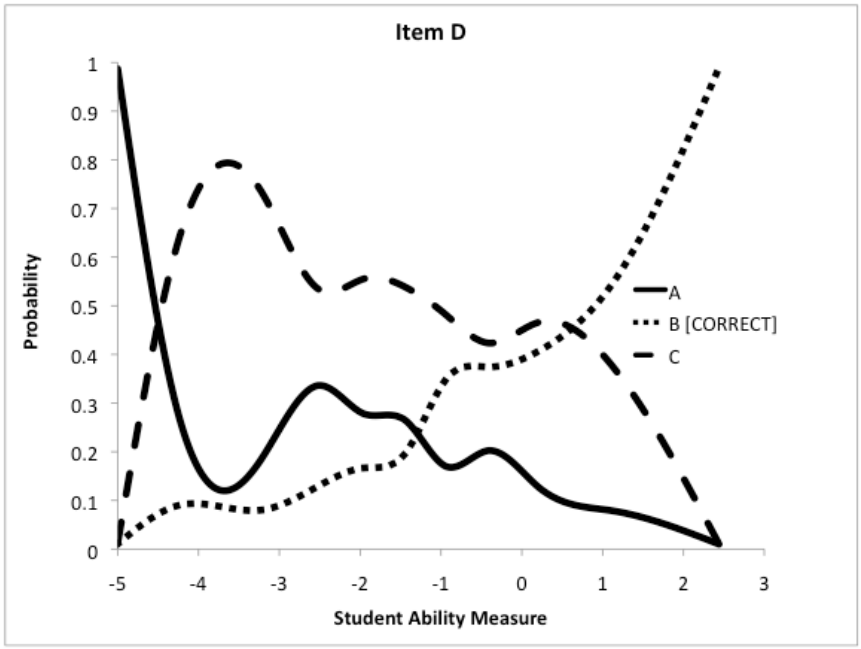
$$A-B + C \rightarrow A + B-C \qquad \Delta H^{\circ}_{rxn} = 400\ kJ/mol$$

(A) The breaking of the A-B bond is exothermic and the making of the B-C bond is endothermic

(B) The bond enthalpy (energy) of the reactants is larger than the bond enthalpy (energy) of the products

(C) The reaction requires 400 kJ/mol of energy to occur

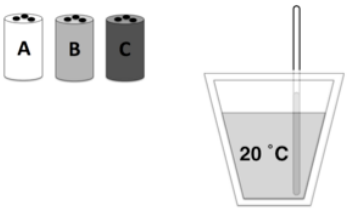
Item D			
Item Option	**Count**	**%**	**Rasch Average Ability**
A	282	22	0.11
B	365	28	0.84
C	644	50	0.26
CTT Difficulty	0.29		
CTT Discrimination	0.294		
Rasch Difficulty Measure	1.51		



Item E

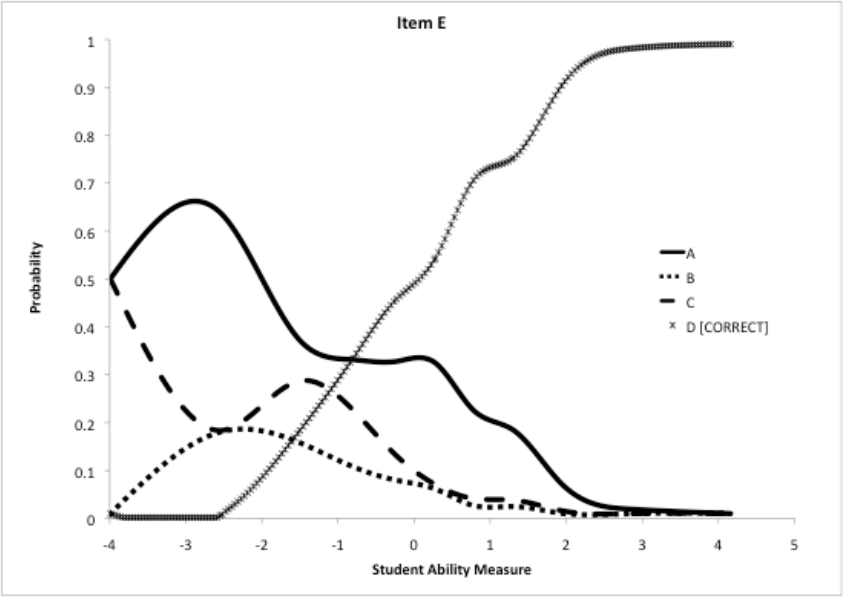
A styrofoam coffee cup contains water at 20 ˚C. Three salt shakers are shown below, containing salts A, B & C. Use the reaction enthalpies given to choose the answer that most accurately describes what would happen when equivalent moles of salt are added to the water.

Reaction	$\Delta H^{\circ}_{dissolution}$
A (s) --> A (aq)	-100 kJ/mol
B (s) --> B (aq)	50 kJ/mol
C (s) --> C (aq)	0 kJ/mol



(A) When salt A is added to the water, heat is created
(B) When salt C is added to the water, it will not dissolve
(C) The temperature of the water in the cup will increase when salt B is added
(D) Adding salt A will result in the largest change in temperature

Item E			
Item Option	**Count**	**%**	**Rasch Average Ability**
A	319	25	-0.03
B	65	5	-0.32
C	103	8	-0.41
D	804	62	0.71
CTT Difficulty	0.617		
CTT Discrimination	0.474		
Rasch Difficulty Measure	-0.22		



Item F

The production of carbon dioxide from elemental carbon and oxygen is shown in the reaction below. For this reaction, choose the most accurate statement below.

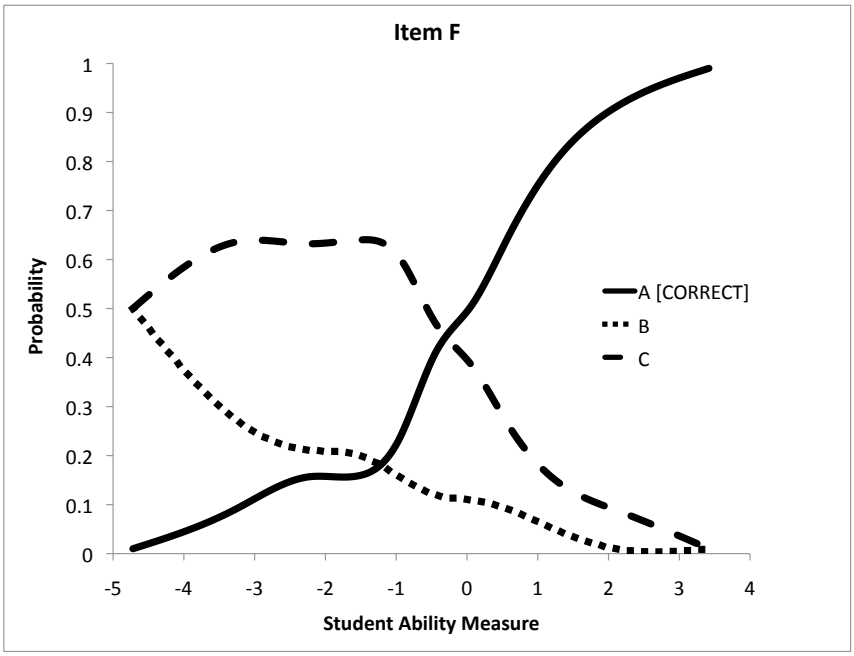
$$\mathbf{C\,(s) + O_2\,(g) \;\text{-->}\; CO_2\,(g)} \qquad \Delta H^{\circ}_{rxn} < 0$$

(A) The product is more energetically stable than the reactants

(B) The production of CO_2 (g) is an endothermic process

(C) The change in enthalpy of the reaction depends on the amount of heat contained in the reactants and product

Item F			
Item Option	**Count**	**%**	**Rasch Average Ability**
A	609	47	0.85
B	153	12	-0.05
C	529	41	-0.02
CTT Difficulty	0.471		
CTT Discrimination	0.543		
Rasch Difficulty Measure	0.52		



Item G

Use the chemical equations below to choose the most accurate response. Each chemical equation represents the formation of a molecule from elements in their standard state.

Reaction	Equation	ΔH°_{rxn}
[1]	A (g) + B (g) --> AB (g)	-100 kJ/mol
[2]	C (g) + D (g) --> CD (g)	-500 kJ/mol

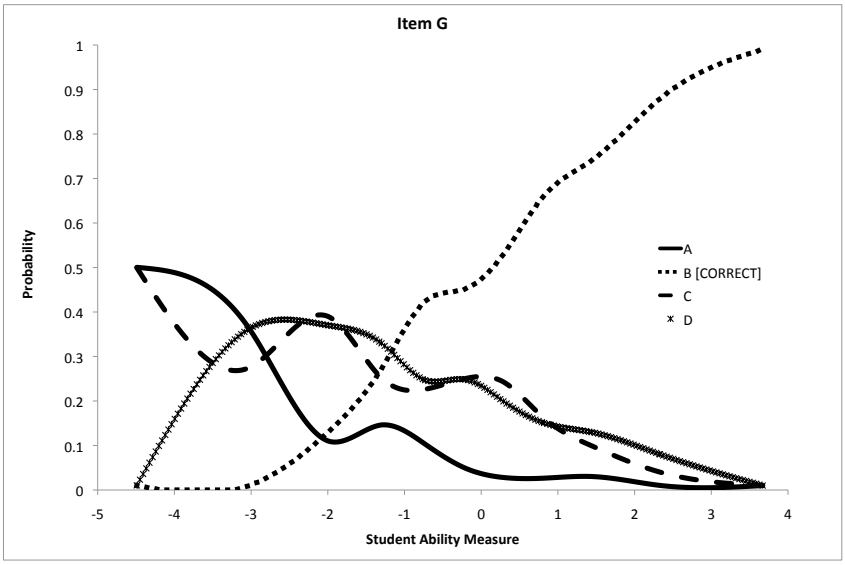
(A) Reaction [2] will reach completion faster than reaction [1]

(B) The bond energy for AB (g) is less than the bond energy for CD (g)

(C) Based on the $\Delta H^{\circ rxn}$ values, neither reaction requires energy to occur

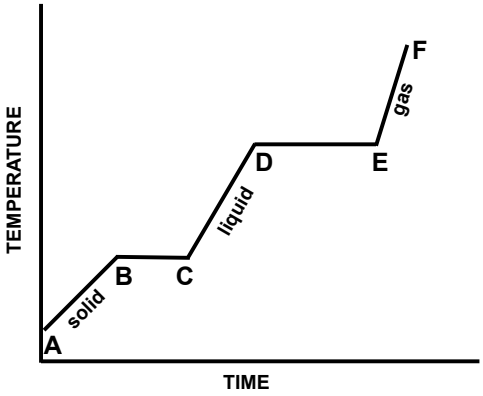
(D) Reaction [1] is more endothermic than reaction [2]

Item G			
Item Option	**Count**	**%**	**Rasch Average Ability**
A	75	6	-0.37
B	674	52	0.72
C	270	21	0.08
D	273	21	0.07
CTT Difficulty	0.533		
CTT Discrimination	0.391		
Rasch Difficulty Measure	0.28		



Item H

Using the heating curve for water provided, select the most accurate answer.



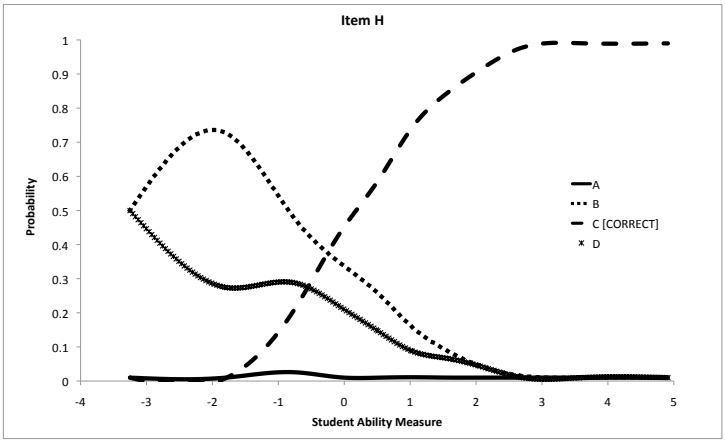
(A) The Y-axis of this graph could also be labeled as *heat*, because temperature and heat are the same

(B) Moving from point **D** to **E** temperature is constant, therefore no thermal energy is added

(C) The freezing of water, represented by moving from **C** to **B**, is an exothermic process

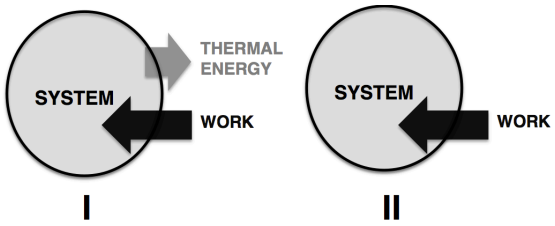
(D) The water at point **C** is a liquid, therefore the temperature cannot be 0 ˚C

Item H			
Item Option	**Count**	**%**	**Rasch Average Ability**
A	6	0	-0.40
B	192	15	-0.44
C	976	76	0.65
D	118	9	-0.36
CTT Difficulty	0.72		
CTT Discrimination	0.469		
Rasch Difficulty Measure	-0.96		



Item I

Two identical systems I and II are shown below. The direction and magnitude of thermal energy transfer and work are represented by arrows. Use this information to choose the most accurate response below.



(A) The total energy (internal energy) for system I will increase

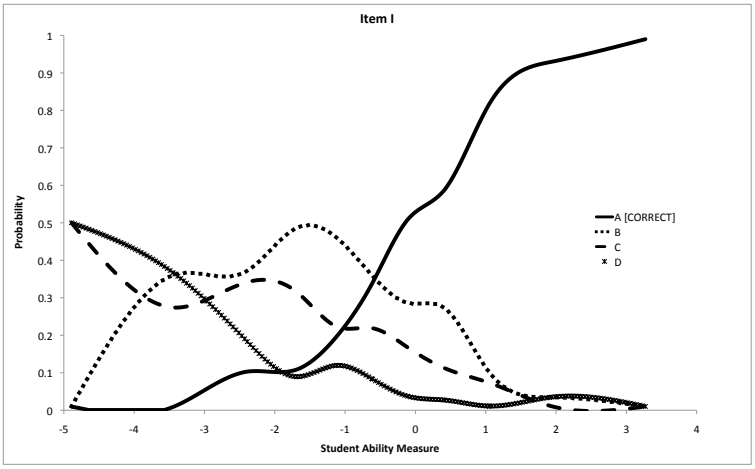
(B) The temperature of system I will decrease

(C) The process shown in system II can be described as endothermic

(D) The sign of the work with respect to system II is negative

Item I			
Item Option	**Count**	**%**	**Rasch Average Ability**
A	568	44	0.93
B	408	32	0.04
C	228	18	-0.07
D	87	7	-0.30

CTT Difficulty	0.441
CTT Discrimination	0.569
Rasch Difficulty Measure	0.68



Item K

If a reaction has a **positive** reaction enthalpy (ΔH_{rxn}), choose the most accurate response below.

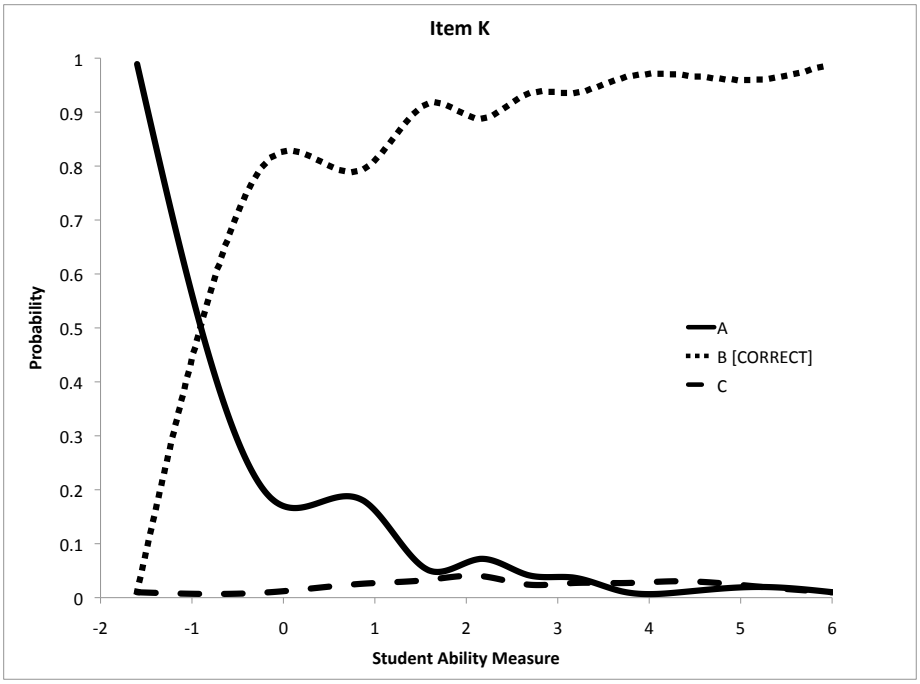
(2) The reaction can be described as an **exothermic** process

(3) The reaction can be described as an **endothermic** process

(4) There is **NOT** enough information to determine if the reaction is an exothermic or endothermic process

Item K			
Item Option	**Count**	**%**	**Rasch Average Ability**
A	55	4	-0.48
B	1200	93	0.43
C	37	3	0.32

CTT Difficulty	0.928
CTT Discrimination	0.089
Rasch Difficulty Measure	-2.60



C2. Final Version for the Thermochemistry Concept Inventory

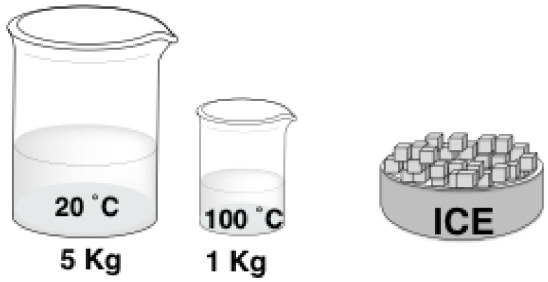
1. If a reaction has a **positive** reaction enthalpy (ΔH_{rxn}), choose the most accurate response below.

(A) The reaction can be described as an **exothermic** process

(B) The reaction can be described as an **endothermic** process

(C) There is **NOT** enough information to determine if the reaction is an exothermic or endothermic process

2. Two beakers of differing volumes contain pure water at different temperatures. Ice is added to the water in each beaker. Choose the most accurate answer given below.



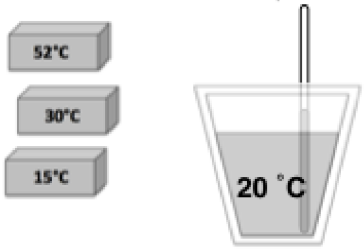
(A) Equal amounts of ice will melt in each beaker

(B) The water is considered the system because it is giving off heat

(C) The melting of the ice in either container is considered an exothermic process

(D) More ice will melt in the beaker with water at 100 °C

3. A styrofoam coffee cup contains water at 20 ˚C. Three identical metal blocks at three different temperatures are shown to the left of the cup. Choose the most accurate response below.



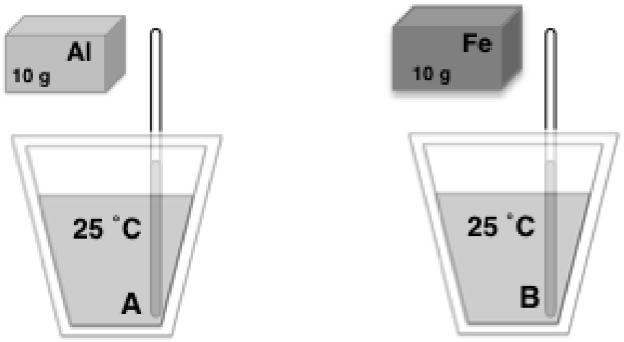
(A) When the block at 30 ˚C is added to the water, thermal energy will flow back and forth between the block and the water until thermal equilibrium is reached

(B) When the block at 52 ˚C is added to the water, the system would be defined as everything in the coffee cup and the surroundings would be everything else

(C) When the block at 15 ˚C is added to the water, the process can be described as an endothermic process with respect to the block

4. A block of Aluminum (Al) and a block of Iron (Fe) each at 50 ˚C are simultaneously dropped into identical styrofoam cups containing the same amount of water at 25 ˚C water. Choose the most accurate answer given below.

Specific Heat (**Al**) > Specific Heat (**Fe**)



(A) After adding either block to the water, the process can be described as an endothermic process, with respect to the block

(B) Thermal energy will be transferred faster between the Al block and the water than between the Fe block and the water

(C) The final temperature of the water in both A and B will be the same

(D) The water in A will have a higher final temperature than the water in B

5. Use the following reaction and the associated standard reaction enthalpy to choose the most accurate answer below.

$$A-B + C \rightarrow A + B-C \qquad \Delta H^{\circ}_{rxn} = 400\ kJ/mol$$

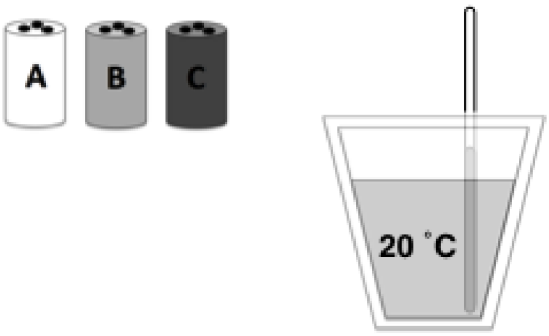
(A) The breaking of the A-B bond is exothermic and the making of the B-C bond is endothermic

(B) The bond enthalpy (energy) of the reactants is larger than the bond enthalpy (energy) of the products

(C) The reaction requires 400 kJ/mol of energy to occur

6. A styrofoam coffee cup contains water at 20 ˚C. Three salt shakers are shown below, containing salts A, B & C. Use the reaction enthalpies given to choose the answer that most accurately describes what would happen when equivalent moles of salt are added to the water.

Reaction	$\Delta H^{\circ}_{dissolution}$
A (s) --> A (aq)	-100 kJ/mol
B (s) --> B (aq)	50 kJ/mol
C (s) --> C (aq)	0 kJ/mol



(A) When salt A is added to the water, heat is created

(B) When salt C is added to the water, it will not dissolve

(C) The temperature of the water in the cup will increase when salt B is added

(D) Adding salt A will result in the largest change in temperature

7. The production of carbon dioxide from elemental carbon and oxygen is shown in the reaction below. For this reaction, choose the most accurate statement below.

$$\mathbf{C\,(s) + O_2\,(g) \;\text{-->}\; CO_2\,(g)} \qquad \Delta H^{\circ}_{rxn} < 0$$

(A) The product is more energetically stable than the reactants

(B) The production of CO_2 (g) is an endothermic process

(C) The change in enthalpy of the reaction depends on the amount of heat contained in the reactants and product

8. Use the chemical equations below to choose the most accurate response. Each chemical equation represents the formation of a molecule from elements in their standard state.

Reaction	Equation	ΔH°_{rxn}
[1]	A (g) + B (g) --> AB (g)	-100 kJ/mol
[2]	C (g) + D (g) --> CD (g)	-500 kJ/mol

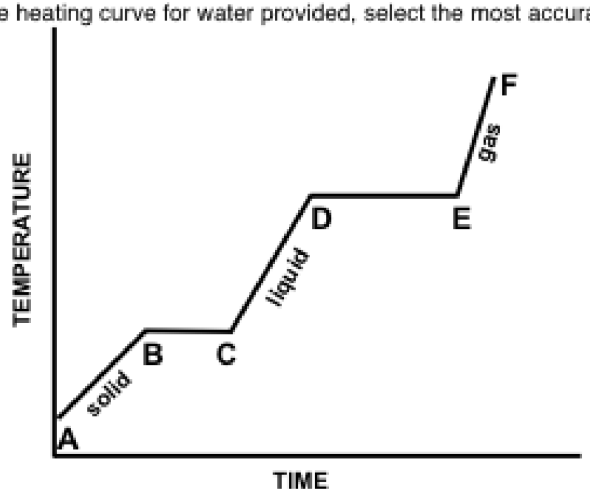
(A) Reaction [2] will reach completion faster than reaction [1]

(B) The bond energy for AB (g) is less than the bond energy for CD (g)

(C) Based on the ΔH°_{rxn} values, neither reaction requires energy to occur

(D) Reaction [1] is more endothermic than reaction [2]

9. Using the heating curve for water provided, select the most accurate answer.

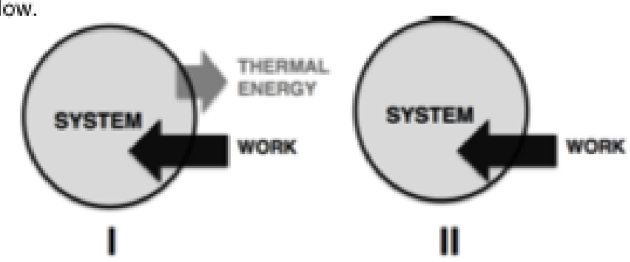


(A) Moving from point **D** to **E** temperature is constant, therefore no thermal energy is added

(B) The freezing of water, represented by moving from **C** to **B**, is an exothermic process

(C) The water at point **C** is a liquid, therefore the temperature cannot be 0 ˚C

10. Two identical systems I and II are shown below. The direction and magnitude of thermal energy transfer and work are represented by arrows. Use this information to choose the most accurate response below.



(A) The total energy (internal energy) for system I will increase

(B) The temperature of system I will decrease

(C) The process shown in system II can be described as endothermic

(D) The sign of the work with respect to system II is negative

C3. Formative Assessment Key for Thermochemistry Concept Inventory

Thermochemistry Concept Inventory Key

- All correct answers are indicated with **BOLD** face type
- Explanation for all responses, correct and incorrect, are given below each item
- Distracters that correspond to similar alternative conceptions are given in brackets at end of explanations

---

1. If a reaction has a **positive** reaction enthalpy (ΔH_{rxn}), choose the most accurate response below.

(A) The reaction can be described as an **exothermic** process

(B) The reaction can be described as an endothermic process

(C) There is **NOT** enough information to determine if the reaction is an exothermic or endothermic process

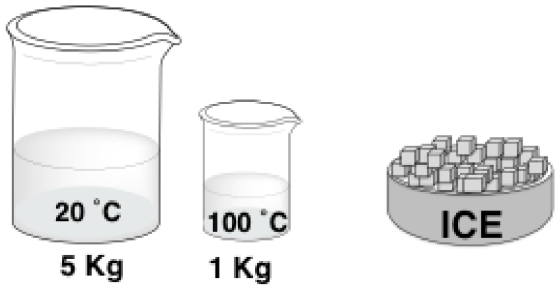
*A: The definition of an **exo**thermic reaction is a reaction that has a **negative** ΔH_{rxn}. [see 3C, 4A, 6C, 7B, 9B, 10C]*

***B**: The definition for an **endo**thermic reaction is a reaction that has a **positive** ΔH_{rxn}, which is stated in the question.*

C: The question does provide enough information to determine that the reaction is endothermic, based on the sign of the ΔH_{rxn}.

Thermochemistry Concept Inventory Key

2. Two beakers of differing volumes contain pure water at different temperatures. Ice is added to the water in each beaker. Choose the most accurate answer given below.



(A) Equal amounts of ice will melt in each beaker

(B) The water is considered the system because it is giving off heat

(C) The melting of the ice in either container is considered an exothermic process

(D) More ice will melt in the beaker with water at 100 °C

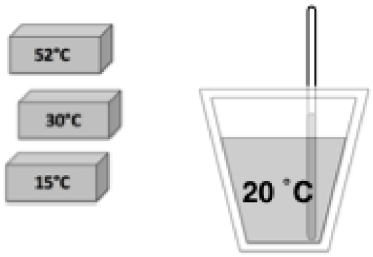
A: *The total amount of thermal energy in both beakers are equal, therefore equal amounts of ice will melt in each beaker. This can be seen by using $q=ms\Delta T$, where the specific heat (s) for the water in each beaker is the same, but the beaker at 20 ˚C has 5 times the mass (m) but 1/5 the expected temperature change (ΔT). Thus, the total amount thermal energy available to melt the ice will be equivalent.*

B: Defining of the system and surroundings are important in order to understand a process, in this case, the melting of ice in a water bath. The system in not always defined as the body that thermal energy is leaving. For this problem, the ice can be considered the system and the water/beaker being the surroundings. The temperature of the water can be measured, such that information about the nature of the ice melting (exothermic or endothermic process) can be addressed. [see 3B]

C: If the ice cubes are defined as our system, for the ice to melt, thermal energy is transferred from the warmer water to the colder ice. This would describe an endothermic process, as thermal energy is entering the system (ice). [see 1A, 3C, 4A, 6C, 7B, 9B, 10C]

D: For the same reason given for the correct answer, A, even though the beaker with water at 100 ˚C seems like it could melt more ice due to the high temperature, the amount (mass) of the water needs to be accounted for assess the thermal energy available to melt the ice. [see 4C & 9A]

3. A styrofoam coffee cup contains water at 20 ˚C. Three identical metal blocks at three different temperatures are shown to the left of the cup. Choose the most accurate response below.



(A) When the block at 30 ˚C is added to the water, thermal energy will flow back and forth between the block and the water until thermal equilibrium is reached

(B) When the block at 52 ˚C is added to the water, the system would be defined as everything in the coffee cup and the surroundings would be everything else

(C) When the block at 15 ˚C is added to the water, the process can be described as an endothermic process with respect to the block

A: Unlike chemical equilibrium, where concentrations of products and reactants fluctuate until a fixed ratio is reached, thermal equilibrium simply refers to the point where two thermal bodies that are in contact (can transfer thermal energy) have reached a final temperature. The "zeroth" law of thermodynamics states that thermal energy flows from a hot body to a cold body.

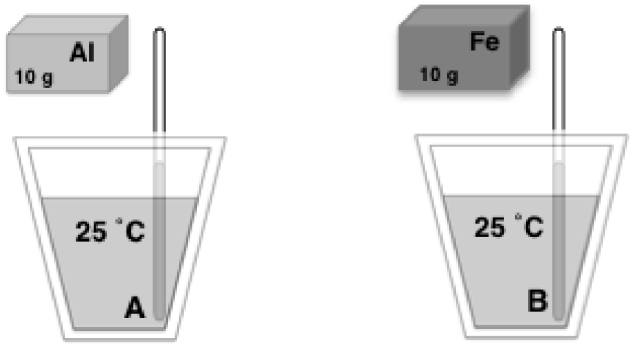
B: Defining of the system and surroundings are important in order to understand a process, in this case, the transfer of thermal energy between blocks of metal and a water bath. If the system was defined as everything in the coffee cup (water and metal block), there would be no way to assess if the process was endothermic or exothermic, since the thermometer is measuring the temperature change of the water, which would be defined as part of the system. Therefore, to obtain information about this process, the water should be defined as the surroundings and the block defined as the system. In calorimetry experiments such as this, we typically measure temperature changes of the surroundings. [see 2B]

***C:** The block at 15 ˚C can be considered the system, and the water at 20 ˚C as the surroundings. The 15 ˚C block, when placed in the water, will be at a lower temperature than the surrounds, so thermal energy will be transferred to block (the system), which can be described as an endothermic process with respect to the block. [see 1A, 2C, 4A, 6C, 7B, 9B, 10C]*

Thermochemistry Concept Inventory Key

4. A block of Aluminum (Al) and a block of Iron (Fe) at 50 ˚C are simultaneously dropped into identical styrofoam cups containing the same amount of water at 25 ˚C water. Choose the most accurate answer given below.

Specific Heat (**Al**) **>** Specific Heat (**Fe**)



(A) After adding either block to the water, the process can be described as an endothermic process, with respect to the block

(B) Thermal energy will be transferred faster between the Al block and the water than between the Fe block and the water

(C) The final temperature of the water in both A and B will be the same

(D) The water in A will have a higher final temperature than the water in B

A: An endothermic process can be defined as a process where thermal energy enters a system, and an exothermic process can be defined as a process where thermal energy exits a system. In both cases, if the metal blocks are defined as the system, thermal energy will be transferred from the blocks at a warmer temperature to the water at a cooler temperature. This would represent an exothermic process. [see 1A, 2C, 3C, 6C, 7B, 9B, 10C]

B: The rate of thermal energy transfer cannot be determined from the specific heat of materials. [see 6B & 8A]

C: Al has a higher specific heat compared with Fe, such that Al will have higher total energy compared to Fe. Even though both blocks have the same temperature, the amount of total energy is different.This will result in the water in cup A having a higher temperature than in cup B. [see 2D & 9A]

***D**: Using $q=ms\Delta T$, the only difference in the metal blocks is the specific heat (s). Because Al has a larger value for s, it will have more total thermal energy to transfer to the water in cup A, allowing it to reach a higher final temperature.*

Thermochemistry Concept Inventory Key

5. Use the following reaction and the associated standard reaction enthalpy to choose the most accurate answer below.

$$A-B + C \rightarrow A + B-C \qquad \Delta H^{\circ}_{rxn} = 400\ kJ/mol$$

(A) The breaking of the A-B bond is exothermic and the making of the B-C bond is endothermic

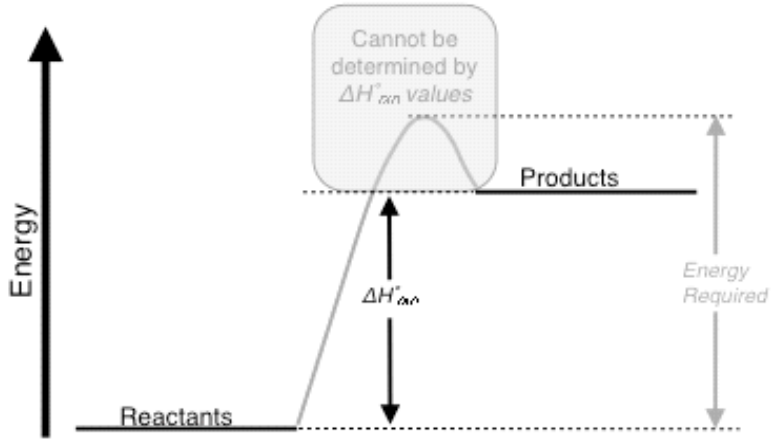
(B) The bond enthalpy (energy) of the reactants is larger than the bond enthalpy (energy) of the products

(C) The reaction requires 400 kJ/mol of energy to occur

A: Bond breaking requires an input of energy and bond making releases energy. Therefore, bond breaking will always be an endothermic process and bond making will always be an exothermic process. The ΔH°_{rxn} simply relates the difference in energies between these two processes in a chemical reaction. [see 5B, 7A, 8B]

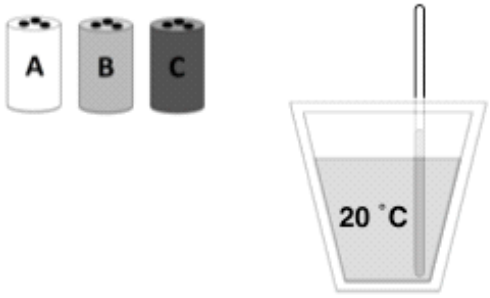
***B**: Looking at the reaction, the A-B bond is broken and the B-C bond is formed. Qualitatively, one can recognize that the value for ΔH°_{rxn} is positive, meaning the energy required to break the A-B bond is more the the energy gained from making the B-C bond. Mathematically, ΔH°_{rxn} can be calculation by the following equation: ΔH°_{rxn} = [Bond Enthalpy of Bonds **Broken**] - [Bond Enthalpy of Bonds **Formed**], a positive ΔH°_{rxn} reveals the bond enthalpy of A-B is greater than the bond enthalpy of B-C. [see 5A, 7A, 8B]*

C: Though the ΔH°_{rxn} is positive 400 kJ/mol (endothermic), the reaction does not necessarily "require" 400 kJ/mol to occur. The ΔH°_{rxn} reflects the energy difference between the products and reactants, but does not give any information about the energy needed for the reaction to occur (see figure below). Similarly, just because a reaction is exothermic (negative ΔH°_{rxn}), one cannot assume energy is not needed for the reaction to occur. [see 8C]



6. A styrofoam coffee cup contains water at 20 ˚C. Three salt shakers are shown below, containing salts A, B & C. Use the reaction enthalpies given to choose the answer that most accurately describes what would happen when equivalent moles of salt are added to the water.

Reaction	$\Delta H^{\circ}_{dissolution}$
A (s) --> A (aq)	-100 kJ/mol
B (s) --> B (aq)	50 kJ/mol
C (s) --> C (aq)	0 kJ/mol



(A) When salt A is added to the water, heat is created

(B) When salt C is added to the water, it will not dissolve

(C) The temperature of the water in the cup will increase when salt B is added

(D) Adding salt A will result in the largest change in temperature

A: Heat is not "created" when salt A is added to the water, because heat is not a substance. Even though we use the term "heat" as a noun in everyday uses, in thermochemistry heat (q) refers specifically to a process where thermal energy is transferred from one body to another. This might seem very subtle, but it would be equivalent to say that work in created, which is also incorrect, because work refers to another process. This is why both work (w) and heat (q) are not state functions, because as processes, they are path dependent. [see 7C]

B: Though unlikely, a $\Delta H^{\circ}_{dissolution}$ of 0 kJ/mol for the dissolving of salt C does not tell you anything about the solubility of salt C. Just like ΔH°_{rxn}, $\Delta H^{\circ}_{dissolution}$ is only related the difference in energies of the products and reactants. [see 4B & 8A]

C: If the temperature of the cup were to increase when salt B is added to the water, that would mean the dissolving of salt B is exothermic. However, the $\Delta H^{\circ}_{dissolution}$ for salt B is positive, meaning it is an endothermic process and the temperature of the water would decrease. [see 1A, 2C, 3C, 4A, 7B, 9B, 10C]

***D**: The largest change in temperature can either be an increase or decrease in temperature. Therefore, only the MAGNITUDE of $\Delta H^{\circ}_{dissolution}$ is important and not the sign. The reaction with the largest $\Delta H^{\circ}_{dissolution}$ is with salt A, which will increase the temperature of the water. [see 8D]*

Thermochemistry Concept Inventory Key

7. The production of carbon dioxide from elemental carbon and oxygen is shown in the reaction below. For this reaction, choose the most accurate statement below.

$$\mathbf{C\,(s) + O_2\,(g) \;\text{-->}\; CO_2\,(g)} \qquad \Delta H^\circ_{rxn} < 0$$

(A) The product is more energetically stable than the reactants

(B) The production of CO_2 (g) is an endothermic process

(C) The change in enthalpy of the reaction depends on the amount of heat contained in the reactants and product

***A**: The negative sign for the ΔH°_{rxn} represents that the energy associated with the reactants is larger than the energy of the product. This is because the ΔH°_{rxn} represents the difference between these two energies. As a general rule, the lower energy state a particular species (e.g. CO_2) is in, the more energetically stable it will be. Since the product CO_2 (g) has a lower associated energy than the reactants, it can be considered more energetically stable. [see 5A, 5B, 7B]*

B: A $\Delta H^\circ_{rxn} < 0$ means the sign of the enthalpy is negative, representing an exothermic process. [see 1A 2C, 3C, 4A, 6C, 9B, 10C]

C: Heat is a process (think "heating") not a substance. The reactants cannot contain heat. However, they can contain thermal energy. Though this might seem trivial, how the thermal energy is lost directly related to heat (q). The value of q can change depending on the pathway, such that we refer to heat as being path dependent. Because energy (including thermal energy) is a state function (path independent), and heat is not, equating these two terms is incorrect, even if we often use the term heat as a noun in everyday "non-chemistry" situations. [see 6A]

Thermochemistry Concept Inventory Key

8. Use the chemical equations below to choose the most accurate response.

Reaction	Equation	ΔH°_{rxn}
[1]	A (g) + B (g) --> AB (g)	-100 kJ/mol
[2]	C (g) + D (g) --> CD (g)	-500 kJ/mol

(A) Reaction [2] will reach completion faster than reaction [1]

(B) The bond energy for AB (g) is less than the bond energy for CD (g)

(C) Based on the ΔH°_{rxn} values, neither reaction requires energy to occur

(D) Reaction [1] is more endothermic than reaction [2]

A: The ΔH°_{rxn} can tell you whether a reaction is exothermic or endothermic or relative energies of products and reactants, but cannot directly be used in questions dealing with how fast, how far, or how soon a reaction will proceed [see 4B & 6B]

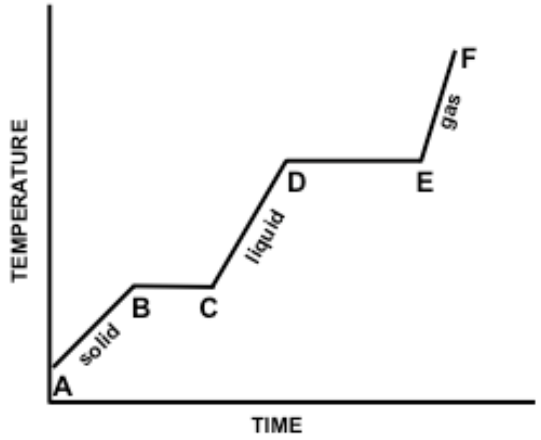
B: *Though no bond energy information is given for the two reactions, both reactions lead to one bond being formed(A-B or C-D). The negative sign for both reactions can be explained by energy being released when either bond is formed, where the magnitude reflects how much energy was released. The amount of energy to break the A-B or CD bond, the bond energy, can be inferred from ΔH°_{rxn}, where CD will have a much larger associated bond energy when compared to AB. [see 5A, 5B, 7A]*

C: Similar to response A, the ΔH°_{rxn} cannot alone provide information on any requirement (or lack of) regarding energy for a reaction to proceed. [see 5C]

D: An important characteristic of ΔH°_{rxn} values is that each contains two pieces of information: what direction is the thermal energy going (the sign) and how much (the magnitude). Thus, a ΔH°_{rxn}=-100 kJ/mol conveys the thermal energy is being transferred from the system to the surrounds (exothermic reaction) and that the amount of energy is 100 kJ for every mole of reactant. Therefore, every ΔH°_{rxn} with a negative sign will be exothermic. Just because the magnitude is closer to zero for reaction [1] than for reaction [2], it does not mean it is more endothermic, it simply means that less thermal energy is transferred per mole of reactant in reaction [1] than in reaction [2]. [see 6D]

Thermochemistry Concept Inventory Key

9. Using the heating curve for water provided, select the most accurate answer.



(A) Moving from point **D** to **E** temperature is constant, therefore no thermal energy is added

(B) The freezing of water, represented by moving from C to B, is an exothermic process

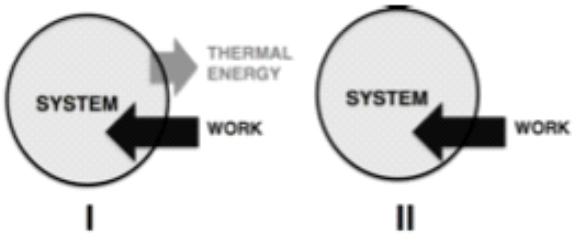
(C) The water at point **C** is a liquid, therefore the temperature cannot be 0 ˚C

*A: Even though the graph represents the temperature remaining constant, it should be remembered that the change in temperature of a substance is not always proportional to the thermal energy added. Many times, the thermal energy is used in other processes that do not directly increase or decrease the temperature. For example, in this diagram, the flat sections (**BC**, **DE**) represent phase changes where thermal energy is used to disrupt interactions within the phase (such as hydrogen bonds in water for **DE**) that would not result in an increase in the solutions temperature. [see 2D & 4C]*

***B**: Liquid water freezes (going from liquid to solid) as thermal energy is lost to the surroundings, which represents an exothermic process.* [see 1A, 2C, 3C, 4A, 6C, 7B, 10C]

*C: Though we commonly know that water freezes at 0 ˚C, it does not mean that all water is frozen (solid) at 0 ˚C at point **C**. The flat section **CB** represents the process in which as thermal energy is removed from the water, liquid water freezes. It is not until point **B** that all of the water is in the solid form.*

10. Two identical systems I and II are shown below. The direction and magnitude of thermal energy transfer and work are represented by arrows. Use this information to choose the most accurate response below.



(A) The total energy (internal energy) for system I will increase

(B) The temperature of system I will decrease

(C) The process shown in system II can be described as endothermic

(D) The sign of the work with respect to system II is negative

A: *The temperature of both systems will increase, as the net change in total energy is positive for both systems. Even though thermal energy is being lost in system 1, the net energy gain is positive. Because the total (internal) energy of the system ($\Delta E_{tot} = q + w$) is the sum of both q and w, a positive ΔE_{tot} will represent an increase in the system's energy. [see 10B]*

B: Similar to the explanation for the correct answer, A, system 1 will have a net energy gain, even if some thermal energy is leaving system 1. A positive total (internal) energy (ΔE_{tot}) will related to an increase in temperature. [see 10A]

C: The term endothermic is related to the process of thermal energy gained by a system, not simply any type of energy. So even if energy is being added to the system through the process of work, one cannot say that the process is endothermic. There are no equivalent terms for the sign of work as there are for heat (exothermic and endothermic). [see 1A, 2C, 3C, 4A, 6C, 7B, 9B]

D: When work is done on the system by the surroundings, as the case for both systems, the sign of work (with respect to the system) is positive.