Chemistry Education Research and Practice

Accepted Manuscript

This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this *Accepted Manuscript* with the edited and formatted *Advance Article* as soon as it is available.

You can find more information about *Accepted Manuscripts* in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/cerp

Psychometric Analysis of the Thermochemistry Concept Inventory

David Wren[a] and Jack Barbera[b*]
[a]Department of Chemistry, Wake Forest University, Winston-Salem, North Carolina 27109, United States
[b]Department of Chemistry and Biochemistry, University of Northern Colorado, Greeley, Colorado 80639, United States (*jack.barbera@unco.edu)

Abstract

Assessing conceptual understanding of foundational topics before instruction on higher-order concepts can provide chemical educators with information to aid instructional design. This study provides an instrument that can be used to identify students' alternative conceptions regarding thermochemistry concepts. The Thermochemistry Concept Inventory (TCI) has been developed for use in formative and summative assessments. Items on the TCI were developed and refined through qualitative evaluation as well as β and pilot tests. Data for the psychometric analysis of the TCI was gathered during a large data collection (N = 1330) and was analyzed using the Rasch model. Supporting evidence for response process validity, structural validity, and reliability were collected. Testing data shows that the TCI is unidimensional and has acceptable fit to the Rasch model. In support of prior qualitative studies, option probability curves support the response process validity and reliability of the items. With exception of one item, when used in summative testing, all items displayed good item functioning. Test-level analysis indicated that the TCI was well targeted to the ability of students in our testing samples. Performance data from different groups shows that the TCI is a measure of overall student ability, providing evidence of concurrent validity.

Introduction

An individual's concept of thermal energy usually precedes formal instruction on the topic. Knowledge, gained through everyday experiences, is brought into the formal-learning setting and built upon across a variety of courses in many different disciplines. At the tertiary level, chemistry instruction in thermal energy typically begins with the topics of thermochemistry, taught during the first-semester of introductory chemistry courses. For chemistry majors, the fundamentals of thermochemistry are built upon when learning about thermodynamics, in both the second-semester of introductory chemistry and again in physical chemistry. Therefore, an understanding of the topics of thermochemistry is foundational in building knowledge of thermodynamics.

An important goal of science education at the tertiary level is the movement of students towards expert understanding within a discipline area. Progression through any science curriculum requires that new "advanced-level" knowledge and ideas be built upon a strong set of foundational conceptions and skills. Studies within cognitive science, as well as in science education, have documented the impact of prior knowledge on learning (National Research Council, 2000; National Research Council, 2012). Prior knowledge has been categorized into three levels (Chi, 2008; Vonsiadou *et al.*, 2008): 1) no prior knowledge, 2) correct but incomplete knowledge, and 3) incorrect knowledge. A large body of research

across a variety of disciplines has focused on the "incorrect knowledge" level. Discrepancies between the incorrect knowledge of novices and experts have been referred to by a variety of names including: alternative conceptions, misconceptions, naïve conceptions, fragmented ideas, or preconceptions. While the different terms stem from different perspectives, the use of incorrect knowledge by students can hinder the attainment of an expert-level understanding.

The study and evaluation of student conceptual understanding has been an active area of investigation within the physics and chemistry education communities for over two decades (Bailey and Slater, 2005; Barke *et al.*, 2009; Docktor and Mestre, 2011). 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 chemistry education literature. Concept inventory length, format, and intend use can vary widely. Many of the standards used by the measurement and assessment communities are becoming accepted and expected for new assessment instruments published by the Chemistry Education Research community (Arjoon *et al.*, 2013), 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(American Educational Research Association, 1999). 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 will present quantitative evidence, based on the Rasch model, for the intended uses and interpretations of data from the Thermochemistry Concept Inventory (TCI). The use of probabilistic models (i.e., Item Response Theory and Rasch) is not new in the chemistry education research community(Scalise *et al.*, 2006; Claesgens *et al.*, 2009), however, use of the Rasch model for the development and evaluation of concept inventories has increases over the past few years(Herrmann-Abell and DeBoer, 2011; Wei *et al.*, 2012; Barbera, 2013; Hadenfeldt *et al.*, 2013; Pentecost and Barbera, 2013). The TCI is a 10-item multiple-choice assessment that uses identified thermochemical alternative conceptions of college-level general chemistry students as distracter options. The design, development and qualitative evaluation of the TCI items have been detailed previously (Wren and Barbera, 2013). The 10-item version of the TCI is presented along with a detailed psychometric evaluation of data at the item level. Use of qualitative data was used to help understand and explain quantitative results, and provided a complete appraisal of evidence for the validity of interpretations and uses of TCI testing data.

Methodology

Student Participants, TCI Administration and Data Collection

Three phases of quantitative data were collected: β testing of potential TCI items (15 items), pilot testing the TCI (12 items), and large-scale data collection with the TCI (10 items). The

pool of 15 potential TCI items was reduced through evaluation during all three phases of quantitative analysis. The 10-item version of the TCI is presented in the Appendix. Prior to all phases, data collection and analysis protocols, as well as data security protocols were reviewed and approved by the Institutional Review Board of the University of Northern Colorado as well as the institutions involved in the data collection. During each administration of the TCI, informed consent was obtained from participants. Those who did not grant consent were removed from the data prior to analysis.

Beta and Pilot Testing

The β 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 formative assessment administrations. Beta testing took place in both first-semester general chemistry lectures, 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, administered by teaching assistants before instruction on thermodynamics. All student identifiers included a "TA code", such that students could be traced back to a specific TA if needed.

Large Data Collection

The large data collection utilized a large-enrollment general chemistry program located in the Pacific West region of the U.S. (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; these summative assessment scores from the TCI were used for evaluation in the course. Based on the time required for the majority of students to complete the TCI during β and pilot testing, a 30-minute block of time was use for the administration during the large data collection. Most students finished 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 TCI test form; consent was indicated by students on their scantron. Instructors collected both the scantron and paper copy of the TCI. After 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.

Rasch Analysis

All data sets were analyzed separately using the dichotomous Rasch model, raw data was imported into the Winsteps program (Linacre, 2010). There are many benefits to using Rasch model analysis in 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. 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 1, in the dichotomous Rasch model 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).

$$P_{ni}(X_{ni} = 1) = \frac{\exp[B_n - D_i]}{1 + \exp[B_n - D_i]} \qquad \textbf{Equation 1}$$

Thus, for an item of difficulty D_i, a student with a higher ability will be 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 fit statistics(Linacre, 2010). 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 the fit 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.

Prior to evaluating the TCI data for evidence of validity and reliability, the data must be evaluated based on the assumptions of and the fit to the Rasch model. These evaluations are produced within Winsteps and reported as standard outputs during an analysis. The first assumption of the Rasch model is unidimensionality, that is, the data only measures one latent trait. Dimensionality is evaluated in the Rasch model using Principal Component Analysis to evaluate the correlated variance of the standardized residuals of items not explained by the 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 (Linacre, 2010). An additional assumption is that of local independence of items. Local independence of items requires that the probability of getting one item correct is independent of the probability of getting another item correct (Linacre, 2010). To verify local independence, inter-item correlations are evaluated using

the same Principal Component 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. In addition to these two assumptions, the Rasch model assumes that all items are of equal discrimination and that no guessing occurs during participant responses. These assumptions were not evaluated for the TCI as a whole test, in fact, it is expected that TCI items will exhibit differential discrimination and that guessing will be a factor in the response process. While the 3-PL Item Response Theory model allows for inclusion of differential discrimination and guessing, this analysis method requires much larger data sets in order to get reasonable fit estimates for these parameters. The presence and impact of differential discrimination and guessing will be discussed at the item-response level during the evaluation of response process validity evidence.

In addition to evaluating data based on the assumptions of the Rasch model, data can also be evaluated for fit to the 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. Fit is estimated by two statistics, outfit and infit. Both fit statistics are chi-squared statistics divided by the degrees of freedom, producing a mean-squared statistic (MNSQ). MNSQ values are reported with associated Z-statistics to assess statistical significance (Bond and Fox, 2007). 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 (Linacre, 2010). When the average is divided by the degrees of freedom, the result is a mean-square statistic (MNSQ), which is reported by Winsteps (Linacre, 2010). 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 (Bond and Fox, 2007; Linacre, 2010). Every MNSQ value has an associated Z-standardized statistic, ZSTD, to assess statistical significance. MNSQ values with ZSTD values greater than ± 2.00 represent statistically significant ($p > 0.05$) values (Linacre, 2010). However, it should be noted, that for large data sets, ZSTD values will increase due to increased statistical power, and should be evaluated only after observations displaying MNSQ misfit have been identified (Linacre, 2010). 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 infit MNSQ has reduced sensitivity to outliers displayed by the outfit statistic (Linacre, 2010). The infit statistic is calculated the same as outfit, but is weighted by the statistical information (model variance) of observations (Linacre, 2010). 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) (Bond and Fox, 2007). 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 (Linacre, 2010). 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 (Bond and Fox, 2007). Items with poor infit statistics can be evidence against response process validity and should be evaluated using qualitative research methods.

Using the populations and methodologies noted above, this manuscript addresses the following research questions regarding the Thermochemistry Concept Inventory (TCI):

1) Is the TCI data appropriate for analysis using the Rasch model?
2) How do the TCI items function when administered as a summative assessment?
3) What evidence supports the validity and reliability of the TCI data?
4) Can the TCI distinguish between performance groups?
5) What evidence supports the generalizability of the TCI?

Results and Discussion

Rasch model analysis was used to collect evidence for the intended uses and interpretations of TCI testing data. This analysis was 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: all distracters are attractive to a certain proportion of the student sample and item difficulty targets student abilities found in the target population. Evaluation of item functioning during β-testing informed the novice response process validity interviews obtained in the prior study (Wren and Barbera, 2013) and was used to flag specific items or item responses for revision or removal from the TCI. Revisions to items were evaluated in the pilot study, resulting in the 10-item version of the TCI used in the large data collection data set presented in this manuscript. Detailed psychometric analysis of the 10-item version of the TCI will be presented along with addition to information gained from β and pilot testing.

Appropriateness of TCI Data for Rasch Analysis

The 10 TCI items administered in the large data collection were evaluated for unidimensionality and local independence. In evaluating the unidimensionality of the TCI, two items had loadings above the recommended value of 0.40. Items D and G (see Appendix for a list of all TCI items) both had loading values of 0.67 on the second contrast. However, the eigenvalue of this contrast was only 1.4, less than 2.00 cut-off. Therefore, these items are not seen as a threat to the unidimensionality of the TCI. All other items had loadings less than ± 0.4. In evaluating the local independence of the items, no two TCI items had inter-item correlation values greater than 0.19. This low value, compared to the criterion of 0.5, provides evidence for the assumption of local independence. Based on these evaluations, the TCI data meets the Rasch assumptions of unidimensionality and local independence and therefore, this is an appropriate measurement model.

The assumptions of item discrimination equality and guessing were not explicitly evaluated for the TCI as a whole. As the most meaningful data from the TCI is at the item-response level (i.e., option choice and connection to an alternate conception) the impact of these parameters will be discussed in context, at the item-level, during the response process validity portion of this section.

TCI Item Functioning During Summative Assessment

Evidence for Item Fit to the Rasch Model

Raw TCI testing data was used analyzed using the Winsteps program. Item difficulty measure estimates and fit statistics calculated using the Rasch model are shown in Table 1.

Table 1. Item-level psychometric estimates for Rasch model analysis, items ordered from hardest (item D) to easiest (item K).

Item	Difficulty Measure[a]	Infit MNSQ[b]	Outfit MNSQ[b]
D	1.51	1.13	1.23
C	1.37	0.94	0.93
I	0.68	0.95	0.93
F	0.52	0.98	1.00
G	0.28	1.08	1.09
A	0.16	0.95	0.93
E	-0.22	1.00	0.97
B	-0.74	0.98	0.97
H	-0.96	0.93	0.84
K	-2.60	1.10	1.56

a: The more negative the value the easier the item
b: Acceptable range for MNSQ values is 1.00 ± 0.5 (Bond and Fox, 2007; Linacre, 2010)

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.

The outfit MNSQ statistics (shown in Table 1) for all items, except item K (outfit MNSQ = 1.56), are acceptable. No ZTSD statistics are given in Table 1, as the large sample size decreases the utility of this statistic for diagnostic purposes. As shown in Table 1, 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. The student ability measures can be directly compared to item difficulty measures, as they are both on a logit scale. An assessment is most informative when item difficulties are matched with student abilities. Given that a student population will have 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 1.

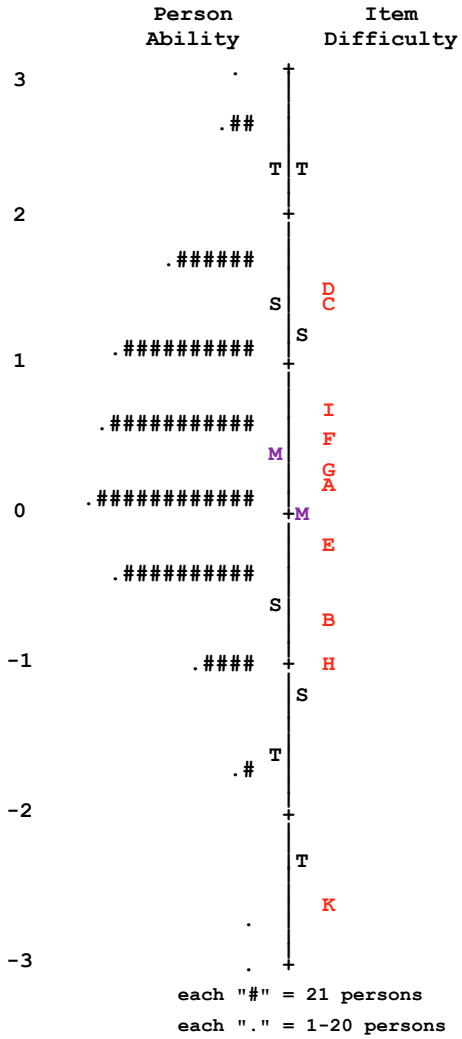


Figure 1. Wright map of item person ability and item difficulty plotted on a logit scale; M indicates mean, S indicates one standard deviation, T indicates two standard deviations.

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

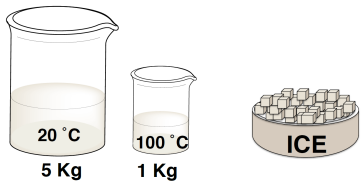
close to one another and the spread of item difficulties covers the range of student abilities, this is an indication of good test targeting. The TCI items display excellent targeting to the population studied, where all items except for one (item K) are well matched with the majority of the student 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 0 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 as low-stakes formative assessment, the average student ability was just below 0 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 (summative) testing. Item targeting also provides evidence for the item reliability. Items that have difficulty measures close to the average student ability will 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 items that at the extremes of the TCI scale (e.g., items D and K). This can help users of the TCI place confidence in item-level interpretations of their testing data.

Evidence for Item Functioning

Distracter analysis of TCI items was used for both the item development process and in the evaluation of the 10-item version of the TCI items used in the large data collection study. Item option probability curves (OPCs) provide a visual representation of the attractiveness of item options, as seen in Figure 2. As both student ability and item difficulty are measured on the same scale, they can be compared(Bond and Fox, 2007). Just as an individual's probability of correctly responding to an item depends on that person's ability in relation to the item difficulty, so to does their probability of choosing an item option. An OPC plots out the probability of choosing an option (plotted on the y-axis) as a function of ability (plotted on the x-axis)(Bond and Fox, 2007; Linacre, 2010). For an individual item, these plots then map out which student abilities are the most likely to choose a given response option to that item. 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 (Herrmann-Abell and DeBoer, 2011). 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. 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 2 is an example of an item with good item distracter functioning.

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
Rasch Difficulty Measure	0.16		

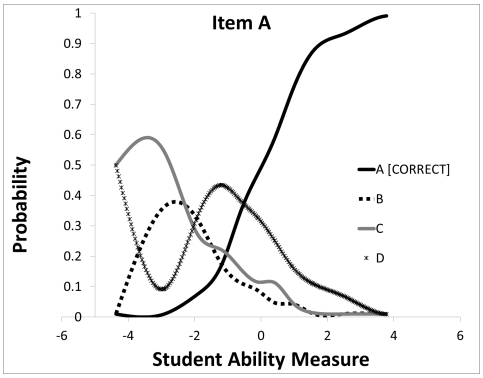


Figure 2. Psychometric information for Item A for both item-level (difficulty measure) and item option-level (option count and frequency; average ability of students choosing option).

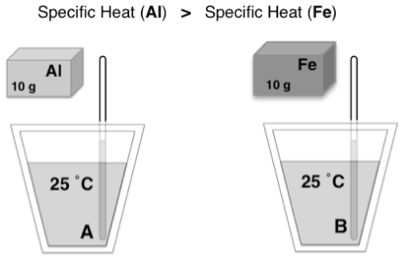
For item A, each response option targets a different student-ability range. This is important for the evaluation of options B and C, which were only selected by 8% and 11% of the student sample, respectively. Options that seem unattractive to students based simply on response frequency might actually 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 alternative conception in a portion of the student population. 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 3), 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) demonstrates discrimination of students based on ability. Option B was the most probable answer for students of the lowest ability, and represents the alternative conception that the rate of thermal energy transfer can be determined using the thermal properties of materials (e.g., specific heat capacity (Wren and Barbera, 2013)). Option C was the most probable for students of average ability, and represents the alternative conception that the temperature of an object is an accurate measure of the total thermal energy of an object (Wren and Barbera, 2013). Based on this analysis, item C has acceptable Rasch item fit statistics and an 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 remain 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
Rasch Difficulty Measure	1.37		

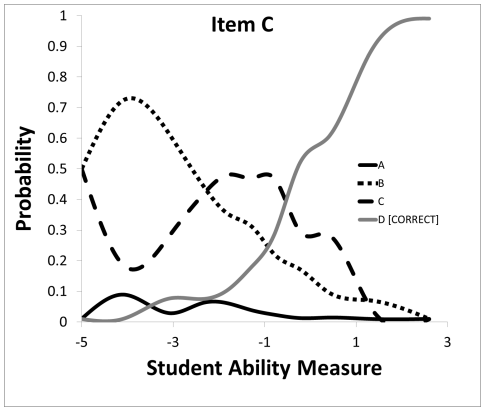


Figure 3. Item C and associated psychometric information demonstrates that option A is unattractive the students (3% option frequency) and does not discriminate among students based on ability (OPC).

Psychometric estimates and item OPCs for all items can be found in the Appendix. In addition to the removal of option A from item C, option A from Item H should also be 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 4, however, it will be retained in the final version of the TCI. Removing this option from item K could increase the threat to validity (construct-irrelevant easiness (American Educational Research Association, 1999; Wren and Barbera, 2013)) 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.

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	

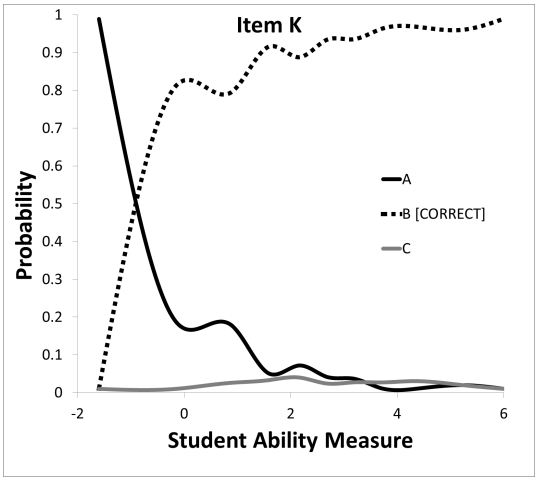


Figure 4. Item K psychometric information demonstrates that option C does not provide useful or reliable information

Validity and Reliability Evidence of TCI Data

Structural Validity

Structural validity is the degree to which the actual test structure matches the designed theoretical structure based on the construct being measured (Loevinger, 1957). 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. As was shown during the evaluation of the assumptions of the Rasch model, the TCI is unidimensional and therefore matches the designed theoretical structure.

Response Process Validity

Option probability curves (OPCs) can be used to support the validity of students' response processes. During the development phase, evidence for response process validity for the TCI items was established through interviews with students from the target population(Wren and Barbera, 2013). These interviews showed that students did not choose response options at random; rather, they used reasoning based on their level of conceptual understanding of the topic being probed by an item. This type of response process data can be further confirmed using the OPCs. It was shown above that the response profile of an OPC could be used to determine differences in student ability as well

as when an option is not uniquely attractive to an ability range. These profiles show that certain response options are attractive to students of a certain ability range and that students are not randomly guessing. In addition, the OPC profiles show that the items have the ability to discriminate between the high and low performers. As noted in the section above, item response options that do not display these response process validity characteristics are will be removed from the final version of the TCI items.

It was noted in the methodology section that the Rasch model does not account for guessing nor differential item discrimination during an analysis and that if estimation of these parameters is required then use of the 3-PL Item Response Model is needed. While we do not have a numerical estimate of the guessing for an item, our OPC profiles show that guessing is not readily apparent. That is, if guessing were a significant factor for an item or response option, then the OPCs would have little correlation between responses and student ability. We are not using this argument to claim that no students guessed when completing the TCI, rather, any guessing that did occur seems to have had little impact on the response process validity of the items. This outcome was also noted in our prior qualitative studies(Wren and Barbera, 2013). In addition, we do not have a numerical estimate of the each item's discrimination in order to gauge their similarity, however, inspection of the OPCs shows clear delineation between the response options chosen by the highest performers compared to the lowest performers. This argument does not imply that all the TCI items have equal discrimination, it simply reflects on use of the data we do have to establish robust item functioning. While not perfect, we feel that the Rasch model does give us a significant amount of evidence with which to support the proper functioning of the TCI at the item-response level.

Reliability

Estimating reliability is situation specific (Meyer, 2010). Just as validity cannot be summarized by one coefficient, evidence for reliability can come from multiple sources and take multiple forms (American Educational Research Association, 1999). For the TCI, the reliability of student item-option responses used to diagnose the use of alternative conceptions requires that students find the options attractive and respond based on their conceptual understanding. Therefore, evidence for reliability of TCI testing data should be derived mainly from item-level analysis of measurement error, rather than measurement error related to the test score (e.g., internal consistency measures). Evidence for the reliability of TCI data will be presented at both the test and item levels.

The internal consistency of the TCI Rasch data was measured for each group from the large scale data collection. The Cronbach's alpha values for these administrations range from 0.26 to 0.46. While these values are below the typically acceptable value of 0.7, it has been argued that internal consistency measures (e.g., Cronbach's alpha or KR-20) may not be an appropriate reliability measure for a concept inventory(Adams and Wieman, 2011). Two recent concept inventory developers found similarly low internal consistency values upon administration of their instruments(Bretz and Linenberger, 2012; McClary and Bretz, 2012). These authors present the case that concept inventories measure students' conceptions, which are typically not coherent. Therefore, a measure designed to elicit a

highly connected structure, such as Cronbach's alpha, is expected to produce low values. To further support the reliability of their instrument data(Bretz and Linenberger, 2012; McClary and Bretz, 2012), the developers performed test-retest measures. However, a major limitation to the test-retest method is that it requires that the items be administered twice, with an appropriate time delay between administrations(Crocker and Algina, 1986). As our large data collection occurred as a classroom quiz, a second administration was not feasible. Therefore, we acknowledge the low internal consistency values of our data at the test level and provide item level data that will further support our findings.

Quantitative evaluation of the randomness of student responses at the item level is produced when using the Rasch measurement model; this evidence can be used as a means to support the reliability of student responses. In the Rasch model, student response data from each item is evaluated for fit to the model. Random student responses would reduce the fit of the data. For example, low performers could randomly choose a correct response for an item above their ability level, thereby reducing the reliability of the responses. Of the 10 TCI items, 9 display good fit to the Rasch model (Table 1), with the exception of item K. Therefore, 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 will be much greater than for the other TCI items.

In addition to Rasch fit statistics, item OPCs profiles support item response reliability. Item options that can discriminate students of different abilities will decrease item option-level error and increase precision (American Educational Research Association, 1999). Item OPCs demonstrate that the majority of TCI item options are attractive to a specific range of student ability (see Appendix); where for some items, each option is the most probable for a specific student ability range.

Additional qualitative evidence to support the reliability of TCI items was gathered during the development and refinement stages (Wren and Barbera, 2013). During these qualitative investigations, most students reported choosing response options based on their conceptual understanding. Therefore, as students were not choosing item options at random, it can be inferred that they would use their conceptual understanding to choose their response options if given the items again.

Thermochemical Conceptual Understanding as a Measure of Student Ability

During data collection, the TCI was administered to four standard sections (A, B, C & D) and one 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 is shown in Table 2; sections A, B, C, and D were combined for statistical comparison to the honors section.

Table 2. Large scale data collection Rasch scores

Section	N	Mean	Standard Deviation
A	330	0.374	1.111
B	315	0.362	0.958
C	338	0.457	1.030
D	310	0.367	1.048
Honors	37	1.452	1.023

There was a significant difference between the TCI average ability for the general sections (M = 0.391, SD = 1.038) and the honors section (M = 1.452, SD = 1.023; t (1330) = 6.13, p < 0.00, two-sided). The magnitude of the difference in the means was small to moderate (eta squared = 0.03). Students in the honors section performed better on all 10 items of the TCI. This provides evidence of the concurrent validity of the TCI, in that the TCI average ability can distinguish students with marked difference in ability.

Generalizability of TCI Across Administration Conditions

Comparison of data from the different testing phases (large data collection, β, and pilot) allows for preliminary analysis of the generalizability of the TCI results. During the different phases of this project, the TCI was given under an array of administration conditions such as different testing environments (e.g., lecture or lab), varied administrators (e.g., course instructor or teaching assistant), and for different stakes (e.g., quiz-based summative assessment or voluntary formative assessment). During β and pilot testing TCI items were administered as a type of formative assessment, being voluntary and having no course evaluation associated with TCI scores. Within the formative assessment administrations, both lecture and laboratory recitation testing environments were assessed.

In comparing results between summative and formative administrations, students, who were given the TCI as a quiz-based summative assessment, and most likely studied thermochemistry prior to administration, did markedly better on certain items. For one item, (item K; presented in Figure 4) that simply addresses the sign convention of the enthalpy of reaction, it was answered correctly by 93% of the students when given as a quiz. However, when the TCI was given as 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 and difference in student ability in the two samples. Most TCI items functioned similarly under the different testing conditions and had acceptable Rasch fit statistics under both conditions.

The TCI was administered by teaching assistants (TAs) in lab recitation sections and by instructors during lecture. As this presents a variety of different testing conditions, increased error associated with the differences in testing environment were a concern. However, under all conditions, items functioned similarly and had acceptable Rasch fit

statistics. In addition, 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.

These varied conditions provide preliminary evidence that the items on the TCI should perform well in both formative and summative assessment and should be invariant to the testing environment. However, these claims should be verified by researchers using the TCI, especially in populations significantly different than those used in this manuscript (Arjoon *et al.*, 2013).

Summary and Conclusions

Evidence for different lines of validity 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, detail was taken to provide explanations that allow readers to evaluate this evidence. 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 (Wilson, 2005).

The use of β and pilot testing of the TCI items provided invaluable information that supplemented evidence collected from qualitative studies. These studies together led to the 10-item version of the TCI tested in the large study. Psychometric evaluation of this 10-item version of the Thermochemistry Concept Inventory using Rasch model analysis provided evidence for structural validity and response process validity. Of the 10 items administered to the large data collection sample, only item K had unsatisfactory psychometric properties. However, this item was designed by 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 (Figure 3) and option A for item H (see Appendix), were unattractive to students of all abilities and will be removed from the TCI. These will be the only changes made to create the final version of the TCI.

Responses to Research Questions

1) Is the TCI data appropriate for analysis using the Rasch model?
In alignment with the assumptions of the Rasch model, data from the TCI was shown to be unidimensional and locally independent. These evaluations were conducted through principal component analysis of data residuals. Quantitative evaluations of guessing and differential discrimination were not conducted for the instrument as a whole; however, examination of the option probability curves was used to evaluate the impact of guessing and the discriminant ability of item options.

2) How do the TCI items function when administered as a summative assessment?
Data from the large scale study showed good item targeting for the population as estimated by the Wright map of the item difficulties and student abilities. All items had acceptable

infit statistics, indicating good item functioning for students within the ability range of the item. One item, item K, had unacceptable outfit statistics with this population, therefore interpretations of this item have more error and should be used with caution. Item K is has been retained on the final version of the TCI due to its acceptable functioning when used in formative assessment administrations.

3) What evidence supports the validity and reliability of the TCI data?
The structural validity of the TCI data was established in the evaluation of the dimensionality analysis noted above. The TCI was designed to be unidimensional; therefore, this finding supports the structure of the results. The response process validity of the items, originally established through qualitative studies(Wren and Barbera, 2013), was further supported by the profiles of the option probability curves, indicating that most option choices correspond to different student ability levels and are therefore not random in nature. While the option probability curves alone do not provide direct evidence of what the students were thinking when they chose a response option, they do support our prior qualitative result that students are selecting options based on their knowledge of the topics presented. The reliability of the TCI data for each section of the course was evaluated using Cronbach's alpha. The low values obtained are expected for an instrument of this type. Additional reliability evidence is provided by the option probability curve profiles and interviews conducted during the development of the items. Each of these sources supports that students are choosing responses based on their conceptual understanding, not at random. Therefore, it is expected that their responses are reliable and reproducible.

4) Can the TCI distinguish between performance groups?
A comparison between honors and non-honors sections of the same general chemistry course revealed that the honors students did perform better as expected. This result was significant with a small effect size. This comparison also provided evidence for the concurrent validity of the TCI data.

5) What evidence supports the generalizability of the TCI?
In addition to the large scale summative assessment administration, the TCI was administered during β and pilot testing under a variety of conditions for formative assessment purposes. While students during the summative assessment performed better on average, the items functioned equally well (better in the case of item K) during formative assessment administrations. In all formative assessment administrations, the performance results and item functioning were invariant to testing conditions (lab vs lecture and graduate TA vs instructor). While these comparisons were made during the developmental stages, it is not expected that the final version of the TCI items will perform differently. However, it is the responsibility of all instrument users to evaluate their data for signs of validity and reliability and how it matches results reported by other users or the developers, this requirement is not specific to 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 alternative conceptions 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(Wren and Barbera, 2013) and in the evaluation of the 10-item version of the test. The use of the Rasch model allowed for high-resolution analysis of item-level testing data in order to evaluate the accuracy (validity) and precision (reliability) of interpretations and uses of testing data at the item-response level.

Intended Uses and Interpretations for TCI testing data

The TCI is designed to be a diagnostic assessment to identify student alternative conceptions 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 could provide evidence for the effectiveness of a new instructional strategy to address alternative conceptions in thermochemistry. Alternatively, the TCI could 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 alternative conception. The challenge with making correlations between a student's incorrect item responses and their use of the alternative conception that the distracter was designed to represent is that there is additional error associated with interpretations at the item-response level then at the item-score or test-score levels (American Educational Research Association, 1999). Therefore, qualitative and quantitative evidence for these interpretations were collected from students in the target population. These studies provided evidence for response process validity. Given that most of the interpretations of TCI testing data is intended to be at the item level.

Potential Users of the TCI

Chemistry instructors and chemical education researchers interested in using the finalized version of the Thermochemistry Concept Inventory (TCI) should contact the corresponding author for an electronic copy. In addition to the finalized version, a detailed answer key has been made for use in formative assessment, to provide students detailed explanations why each distracter is incorrect and what associated alternative conception is associated with each incorrect answer.

References

Adams, W. K. and Wieman, C. E., (2011), Development and Validation of Instruments to Measure Learning of Expert-Like Thinking, *International Journal of Science Education,* **33**, 1289-1312.

American Educational Research Association, A. P. A., National Council on Measurement in Education, (1999), *Standards for educational and psychological testing*, Washington, DC: American Educational Research Association.

Arjoon, J., Xu, X. and Lewis, J., (2013), Understanding the State of the Art for Measurement in Chemistry Education Research: Examining the Psychometric Evidence, *Journal of Chemical Education,* **90**(5), 536-545.

Bailey, J. M. and Slater, T. F. (2005). A contemporary review of K-16 astronomy education research, Highlights in Astronomy, O. Engvold, San Francisco, Astronomy Society of the Pacific, **13**, 1029-1031.

Barbera, J., (2013), A Psychometric Analysis of the Chemical Concept Inventory, *Journal of Chemical Education,* **90**(5), 546-553.

Barke, H. D., Hazari, A. and Yitbarck, S., (2009), *Misconceptions in chemistry; Addressing perceptions in chemical education*, Berlin Heidelburg: Springer-Verlag.

Bond, T. and Fox, C., (2007), *Applying The Rasch Model*, 2nd, New York, NY: Routledge.

Bretz, S. L. and Linenberger, K. J., (2012), Development of the Enzyme-Substrate Interactions Concept Inventory, *Biochemistry and Molecular Biology Education,* **40**(4), 229-233.

Chi, M. (2008). Three Types of Conceptual Change: Belief Revision, Mental Model Transformation, and Categorical Shift, International Handbook of Research on Conceptual Change, S. Vonsiadou, New York, NY, Routledge, 61-82.

Claesgens, J., Scalise, K., Wilson, M. and Stacy, A., (2009), Mapping Student Understanding in Chemistry: The Perspectives of Chemists, *Science Education,* **93**, 56-85.

Crocker, L. and Algina, J., (1986), *Introduction to Classical and Modern Test Theory*, New York, NY: Holt, Rinehart and Winston.

Docktor, J. L. and Mestre, J. P. (2011). A synthesis of discipline-based education research in physics. N. R. Council.

Hadenfeldt, J. C., Bernholt, S., Liu, X., Neumann, K. and Parchmann, I., (2013), Using Ordered Multiple-Choice Items to Assess Students' Understanding of the

Structure and Composition of Matter, *Journal of Chemical Education,* **90**, 1602-1608.

Herrmann-Abell, C. and DeBoer, G., (2011), Using distractor-driven standards-based multiple-choice assessments and Rasch modeling to investigate hierarchies of chemistry misconceptions and detect structural problems with individual items, *Chemistry Education Research and Practice,* **12**, 184-192.

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

Loevinger, J., (1957), Objective Tests As Instruments of Psychological Theory, *Psychological Reports,* **3**(Monograph Supplement 9), 635-694.

McClary, L. M. and Bretz, S. L., (2012), Development and Assessment of a Diagnostic Tool to Identify Organic Chemistry Students' Alternative Conceptions Related to Acid Strength, *International Journal of Science Education,* **34**(15), 2317-2341.

Meyer, J. P., (2010), *Reliability*, New York: Oxford University Press, Inc.

National Research Council (2000). How People Learn: Brain, Mind, Experience, and School. B. Committee on Developments in the Science of Learning, John , Brown, Ann , Cocking, Rodney (Eds.), Commission on Behavioral and Social Sciences and Education. Washington, D.C., National Academy Press.

National Research Council (2012). Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering. C. Committee on the Status, and Future Directions of Discipline-Based Education Research, Singer, Susan , Nielsen, Natalie , Schweingruber, Heidi (Eds.), Board on Science Education, Division of Behavioral and Social Science and Education. Washington, D.C., National Academy Press.

Pentecost, T. C. and Barbera, J., (2013), Measuring Learning Gains in Chemical Education: A Comparision of Two Methods, *Journal of Chemical Education,* **90**(7), 839-845.

Scalise, K., Claesgens, J., Wilson, M. and Stacy, A. (2006). "ChemQuery/Living By Chemistry (LBC)." Retrieved February, 2014, from https://bearcenter.berkeley.edu/project/chemqueryliving-chemistry-lbc.

Vonsiadou, S., Vamvakoussi, X. and Skopeliti, I. (2008). The Framework Theory Approach to the Problem of Conceptual Change, International Handbook of Research on Conceptual Change, S. Vosniadou, New York, NY, Routledge,

Wei, S., Liu, X., Wang, Z. and Wang, X., (2012), Using Rasch Measurement to Develop a Computer Modeling-Based Instrument to Assess Students' Conceptual Understanding of Matter, *Journal of Chemical Education,* **89**(3), 335-345.

Wilson, M., (2005), *Constructing Measures: An Item Response Modeling Approach*, New Jersey: Lawrence Erlbaum Associates, Inc.

Wren, D. and Barbera, J., (2013), Gathering Evidence for Validity during the Design, Development, and Qualitative Evaluation of Thermochemistry Concept Inventory Items, *Journal of Chemical Education,* **90**(12), 1590-1601.

Appendix

Psychometric Analysis of the Thermochemistry Concept Inventory

David Wren[a] and Jack Barbera[b*]
[a]Department of Chemistry, Wake Forest University, Winston-Salem, North Carolina 27109, United States
[b]Department of Chemistry and Biochemistry, University of Northern Colorado, Greeley, Colorado 80639, United States (*jack.barbera@unco.edu)

Potential Users of the TCI

Chemistry instructors and chemical education researchers interested in using the finalized version of the Thermochemistry Concept Inventory (TCI) should contact the corresponding author for an electronic copy. In addition to the finalized version, a detailed answer key has been made for use in formative assessment, to provide students detailed explanations why each distracter is incorrect and what associated alternative conception is associated with each incorrect answer.

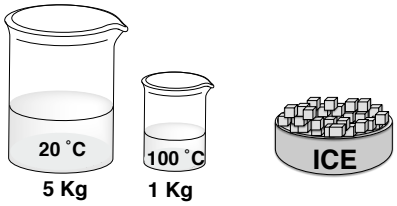
Additional Psychometric Evidence

For each of the 10 items on the Thermochemistry Concept Inventory, a full set of quantitative data is presented.

Data presented includes:
1) The item itself
2) Item response frequencies and Rasch ability levels
3) Rasch difficulty measure values
4) Rasch Item option probability curves

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.



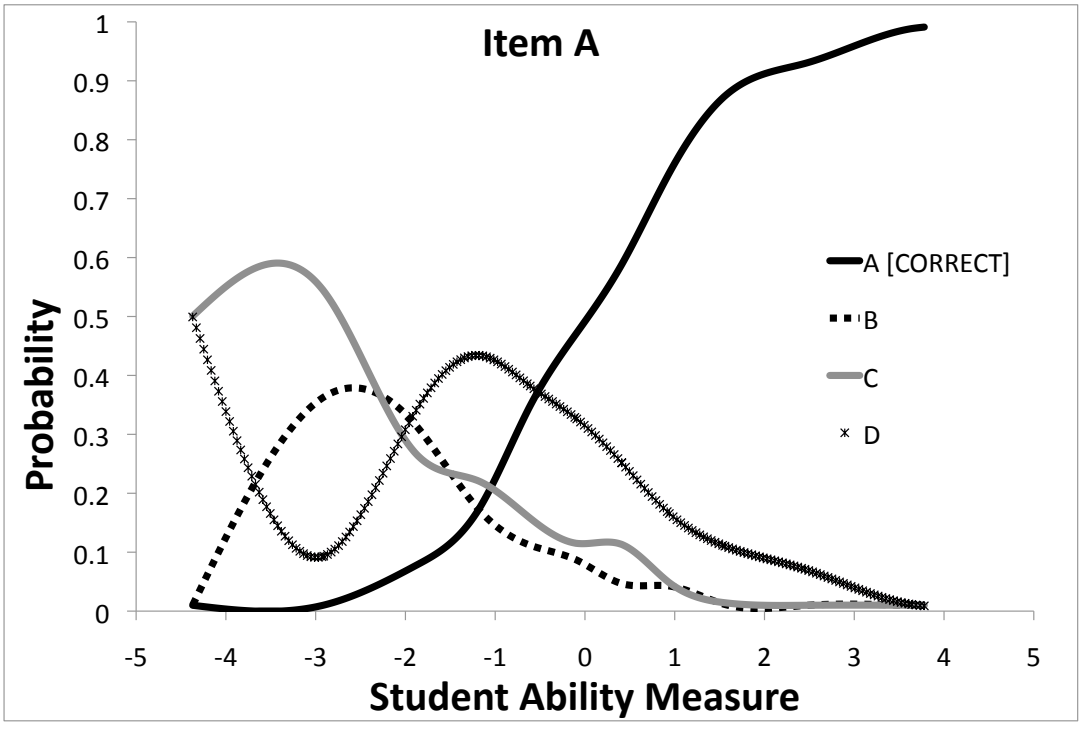
(A) Equal amounts of ice will melt in each beaker

(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

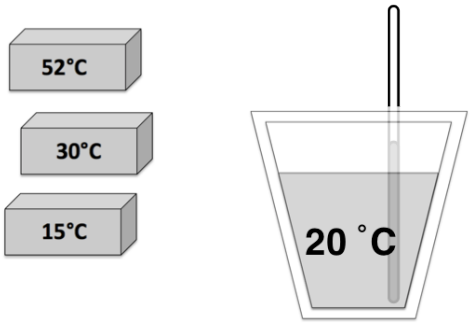
(D) More ice will melt in the beaker with water at 100 °C

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
Rasch Difficulty Measure			0.16



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.

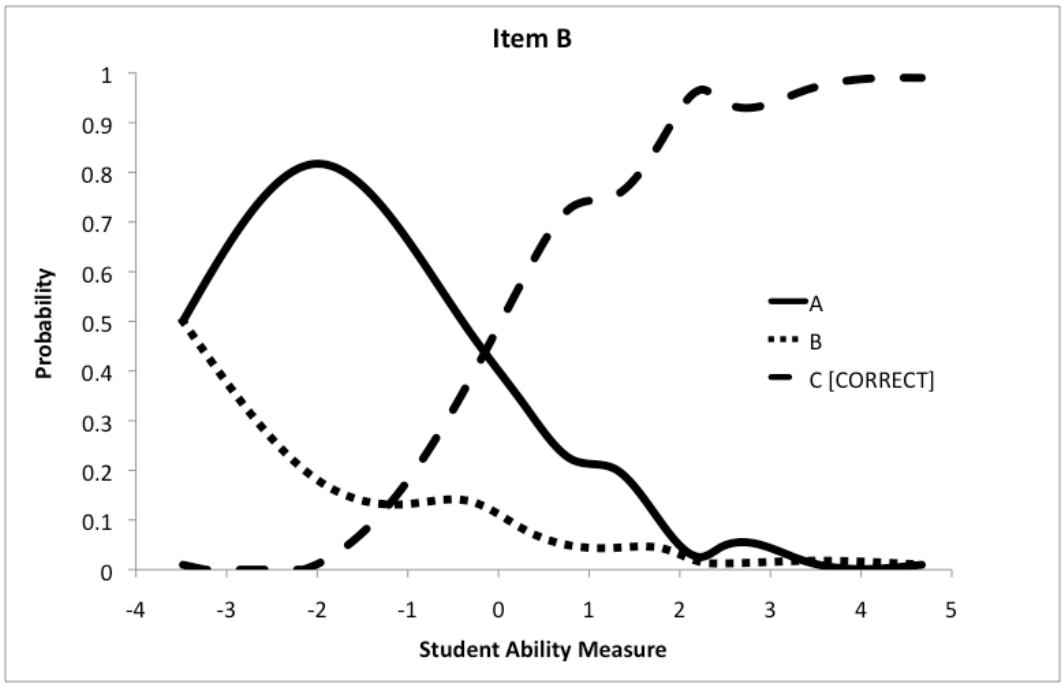


(A) When the block at 30 ˚C is added to the water, thermal energy will flow back and forth between the block and the water until thermal equilibrium is reached

(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

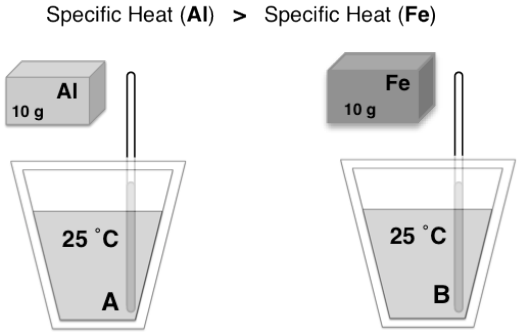
(C) When the block at 15 ˚C is added to the water, the process can be described as an endothermic process with respect to the block

Item B			
Item Option	**Count**	**%**	**Rasch Average Ability**
A	286	22	-0.28
B	75	6	-0.20
C	930	72	0.64
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.



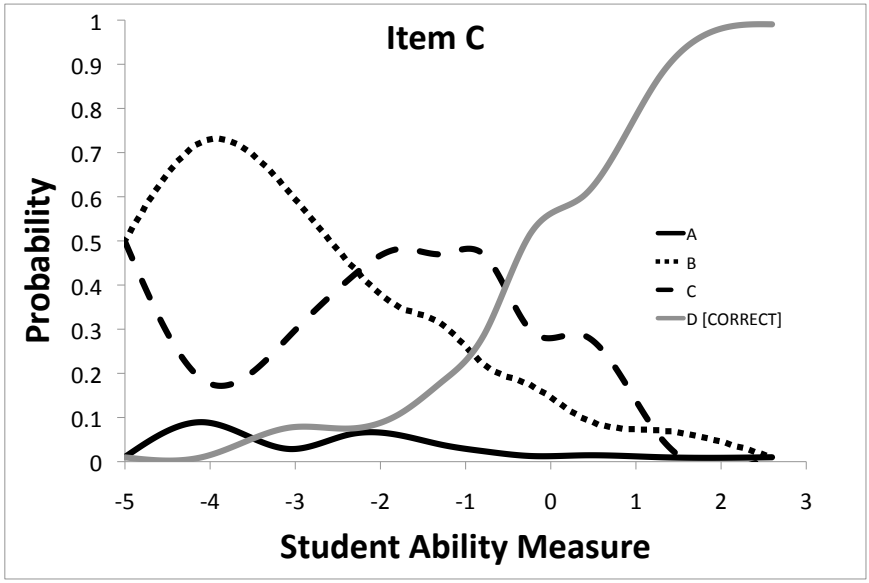
(A) After adding either block to the water, the process can be described as an endothermic process, with respect to the block

(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
Rasch Difficulty Measure	1.37		



Item D

Use the following reaction and the associated standard reaction enthalpy to choose the most accurate answer below.

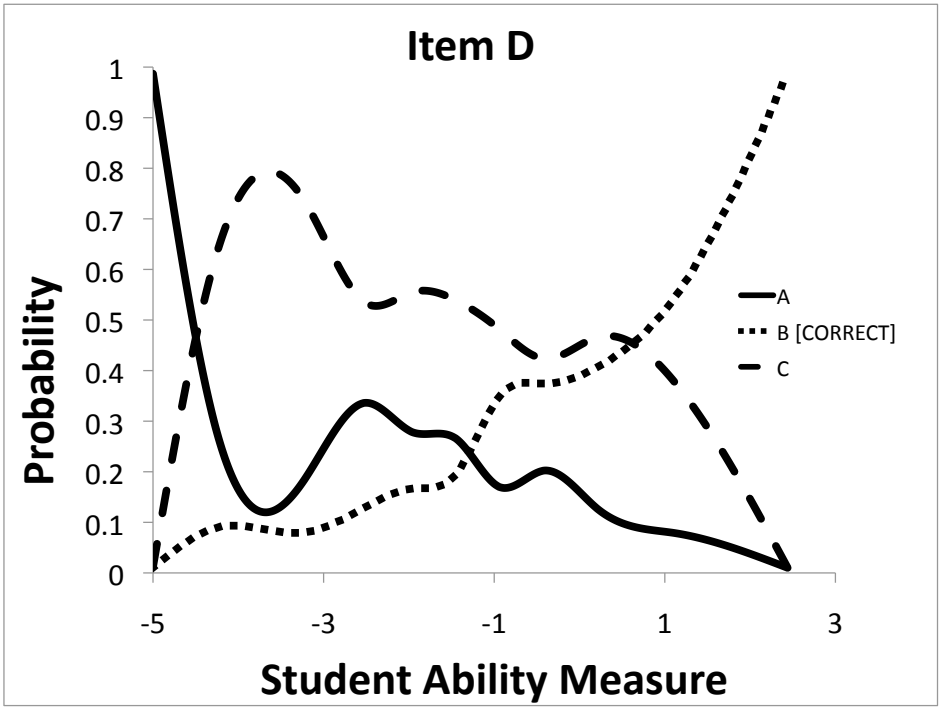
$$A-B + C \rightarrow A + B-C \qquad \Delta H^{\circ}_{rxn} = 400\ kJ/mol$$

(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

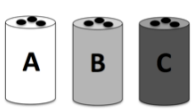
Item D			
Item Option	**Count**	**%**	**Rasch Average Ability**
A	282	22	0.11
B	365	28	0.84
C	644	50	0.26
Rasch Difficulty Measure			1.51

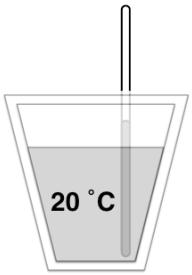


Item E

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





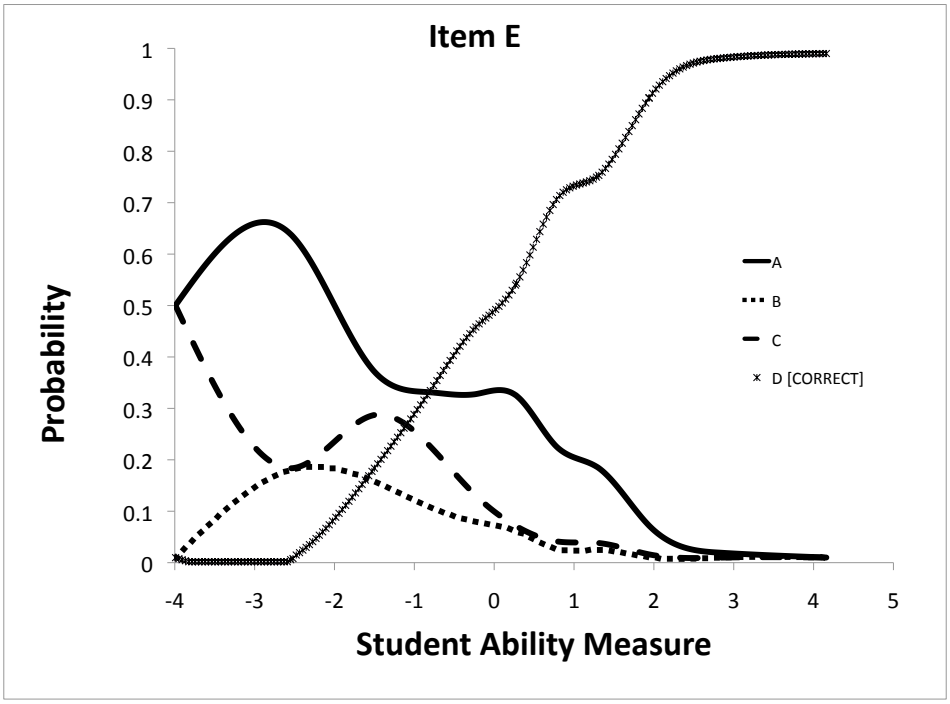
(A) When salt A is added to the water, heat is created

(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
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.

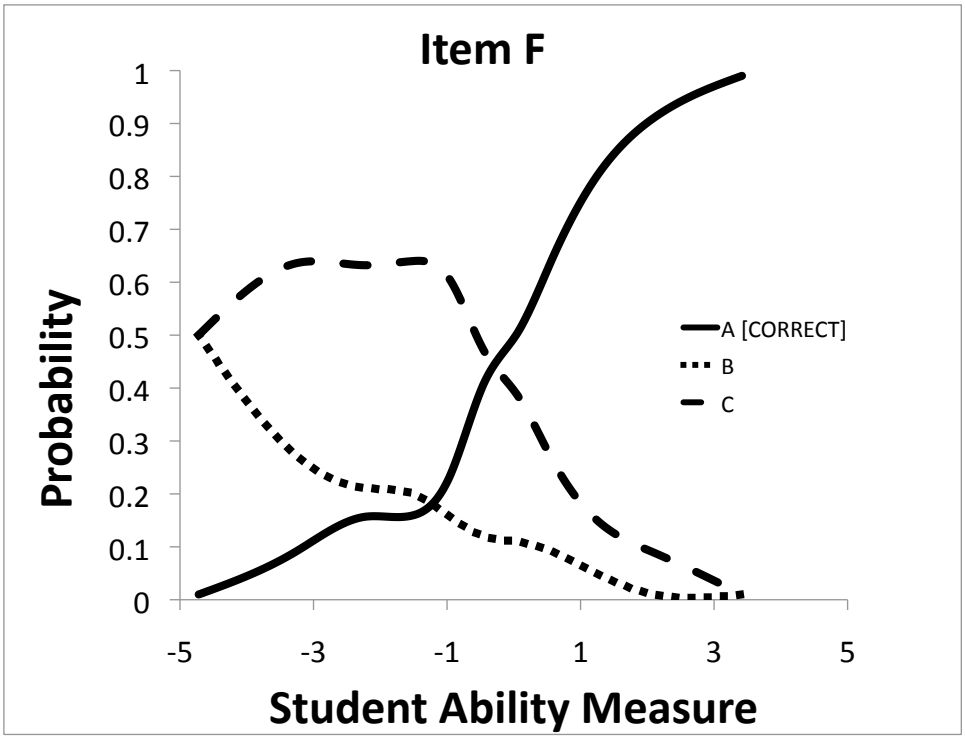
$$C\,(s) + O_2\,(g) \rightarrow CO_2\,(g) \qquad \Delta H^\circ_{rxn} < 0$$

(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
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

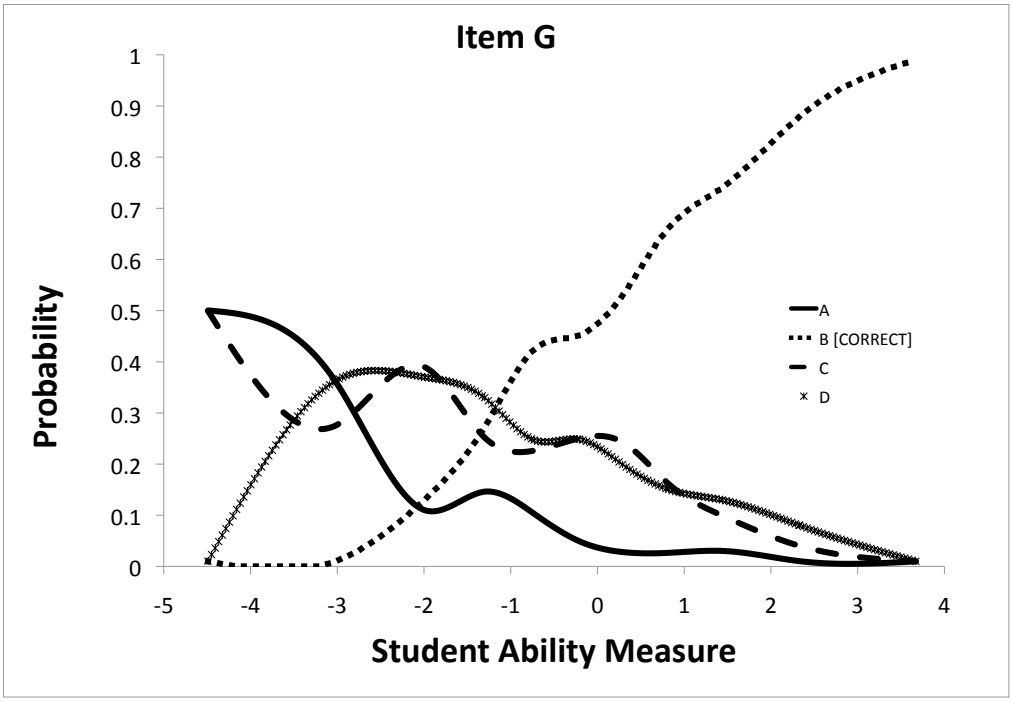
(A) Reaction [2] will reach completion faster than reaction [1]

(B) The bond energy for AB (g) is less than the bond energy for CD (g)

(C) Based on the ΔH°_{rxn} values, neither reaction requires energy to occur

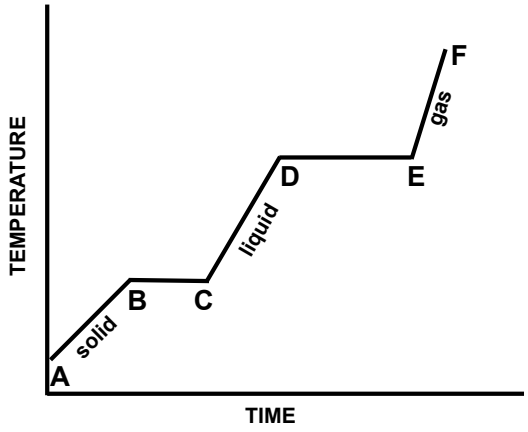
(D) Reaction [1] is more endothermic than reaction [2]

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
Rasch Difficulty Measure			0.28



Item H

Using the heating curve for water provided, select the most accurate answer.



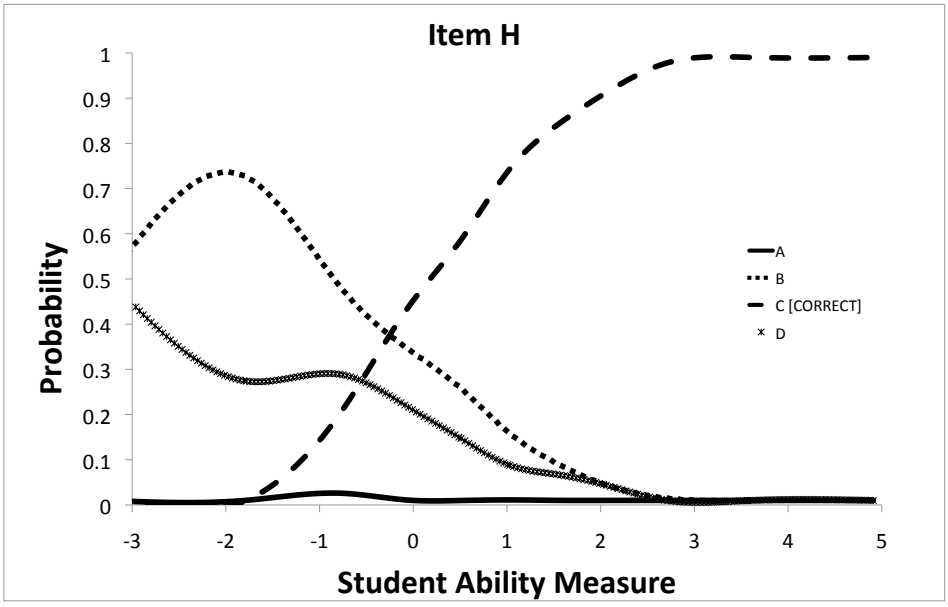
(A) The Y-axis of this graph could also be labeled as *heat*, because temperature and heat are the same

(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

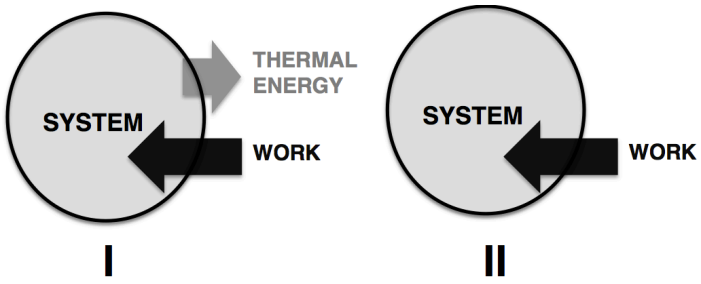
(D) The water at point **C** is a liquid, therefore the temperature cannot be 0 ˚C

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
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.



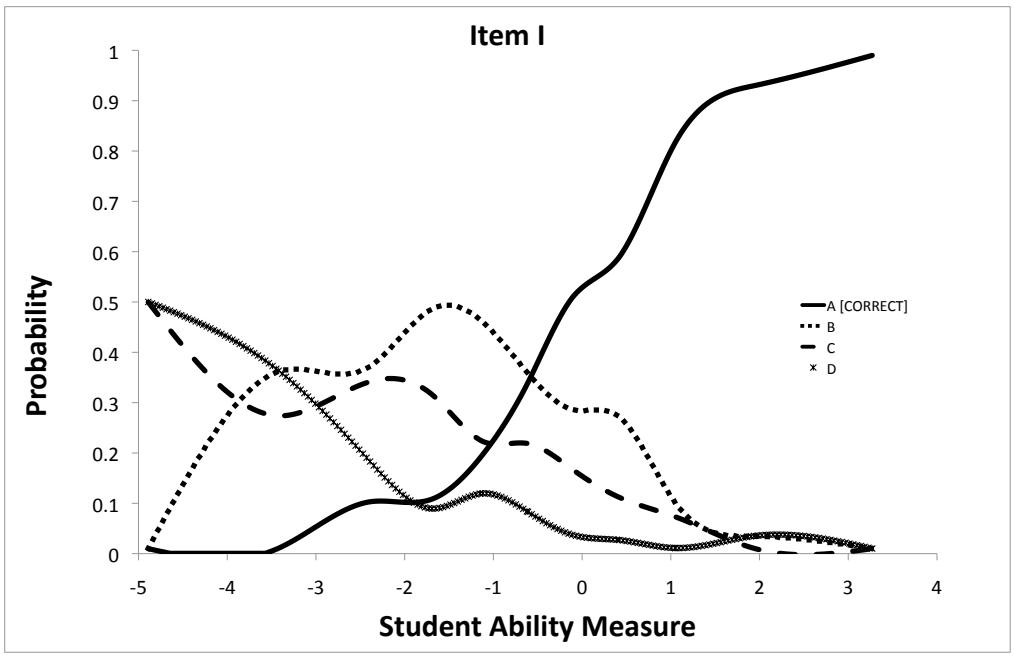
(A) The total energy (internal energy) for system I will increase

(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
Rasch Difficulty Measure	0.68		



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
Rasch Difficulty Measure	-2.60