VALIDATED MULTIPLE-CHOICE TEST ITEMS THAT REQUIRE CRITICAL THINKING SKILLS AND AN ITEM RESPONSE THEORY APPROACH TO TEST ANALYSIS AS TOOLS FOR ASSESSING CRITICAL THINKING IN UNDERGRADUATE BIOLOGY

by

LAUREN JENNIFER IVANS

(Under the Direction of Dr. Julie Kittleson and Dr. Karen Samuelsen)

ABSTRACT

Multiple-choice tests commonly administered in undergraduate biology courses often emphasize factual recall and fail to develop students' critical thinking skills. Since abandoning multiple-choice tests is not realistic for instructors with hundreds of students, this tripartite study was conducted to develop a method for writing and validating multiple-choice items that require critical thinking skills and to demonstrate the advantages of Item Response Theory (IRT) over Classical Test Theory (CTT) for analyzing exams.

Two semesters, Spring 2008 and Spring 2010, of multiple-choice, final exam data from an undergraduate introductory biology course were analyzed using CTT and IRT. Both measurement paradigms generated estimates of item difficulty, student ability, and test reliability. However, the IRT analysis provided more information than the CTT analysis. The IRT analysis showed that the exams did not contain enough difficult questions to precisely measure the ability levels of high achieving students.

The second phase of this study began with the development of 41 multiple-choice items for the undergraduate biology course that proposed to require critical thinking skills. To validate

if the items require critical thinking skills, they were submitted to faculty reviewers who rated whether or not the items required critical thinking skills and they were tested in cognitive think-aloud sessions with undergraduate students. Data from the validation studies provided strong evidence that 32 of the 41 items required critical thinking skills and weak evidence for two items. Data on the remaining 7 items either showed that they did not require critical think skills or were inconclusive. This phase showed the need to investigate the validity of test items and demonstrated a method for doing so.

Twenty-three of the validated critical thinking items were included on the final exam for undergraduate introductory biology course. An IRT analysis was conducted on the exam and again it was found that the test did not contain enough difficult questions to precisely measure the ability levels of high achieving students. The IRT analysis also provided insights into writing multiple-choice items for undergraduate biology that require critical thinking, sources of item difficulty, and areas of student difficulty.

INDEX WORDS: Item Response Theory, Undergraduate biology, Critical thinking, Multiple-choice tests, Validity, Classical Test Theory

VALIDATED MULTIPLE-CHOICE TEST ITEMS THAT REQUIRE CRITICAL THINKING SKILLS AND AN ITEM RESPONSE THEORY APPROACH TO TEST ANALYSIS AS TOOLS FOR ASSESSING CRITICAL THINKING IN UNDERGRADUATE BIOLOGY

by

LAUREN JENNIFER IVANS

B.S., University of California Los Angeles, 2005

M.S., University of California San Diego, 2008

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2012

VALIDATED MULTIPLE-CHOICE TEST ITEMS THAT REQUIRE CRITICAL THINKING SKILLS AND AN ITEM RESPONSE THEORY APPROACH TO TEST ANALYSIS AS TOOLS FOR ASSESSING CRITICAL THINKING IN UNDERGRADUATE BIOLOGY

by

LAUREN JENNIFER IVANS

Major Professors:	Julie Kittleson Karen Samuelsen
Committee:	Kathrin Stanger-Hall David Jackson

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2012

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

A Comparative Analysis of Undergraduate Biology Exams Using Classical Test Theory and Item Response Theory

Abstract:

Two semesters, Spring 2008 and Spring 2010, of Final Exam data from an introductory biology class at a large public university were analyzed through CTT and a 1 parameter IRT model (Rasch Model) to show the potential applications of IRT for university science professors. Fit statistics for all students and items were within acceptable limits, which showed that the Rasch Model was appropriate for the dataset. Another indicator of good model-data fit was that the principle of item parameter invariance was realized in the dataset. While the IRT and CTT analyses both produced estimates of item difficulty, student ability, and test reliability, the IRT analysis was far more informative than the CTT analysis with regard to student ability, test validity, and test refinement, and test validity. From the IRT analysis it was determined that: 1) 2008 students had a higher average ability level than the 2010 students, 2) the exam could only precisely estimate the ability levels of students who were of average and below average ability, 3) ability level estimates for high achievers were associated with the most error, and 4) the overall difficulty level of the exam questions needs to be increased to better measure the ability levels of the high ability students. This comparative analysis therefore demonstrates the benefits and uses of IRT as a framework for designing and evaluating university science exams.

Introduction:

Although many university science exams are rooted in Classical Test Theory, the needs of university science professors have outgrown what Classical Test Theory can provide. Fortunately, measurement theory has an alternative paradigm for the design and analysis of tests that resolves many of the issues with Classical Test Theory.

The two major perspectives in the field of measurement theory are Classical Test Theory (CTT) and Item Response Theory (IRT). These perspectives serve as lenses through which psychometricians design, analyze, interpret, evaluate, and repair tests (Hambleton & Jones, 1993). Classical Test Theory was pioneered by Spearman (1904) and was the dominant framework in testing until the development of IRT models. Thurstone (1925) is often credited with laying the groundwork for IRT (Bock, 1997). IRT evolved slowly until Lord and Novick (1968) catalyzed the field. They extended the nascent concept of IRT to produce a unified theory of testing. At least in the United States, Lord and Novick are credited with developing modern IRT. The Danish mathematician, Georg Rasch (1960), developed his own strand of IRT. Rasch derived a set of models that were used to design instruments to assess reading ability. Rasch's models were also used to develop tests for Denmark's army. Later researchers such as Gerhard Fischer and Benjamin Wright expanded on Rasch's work (Embretson & Reise, 2000). At present, CTT and IRT are both robustly used in measurement theory.

CTT is built on the assumption that an individual's observed test score (X_j) is the sum of his or her true score on the test (T_j) and their error score on the test (E_j):

$$X_j = T_j + E_j \quad (1)$$

An individual's true score is commonly defined as their observed score across parallel tests (tests that assess the same information and/or skills) or as their score on a single test over repeated

testing occasions (M. J. Allen & Yen, 2002; Hambleton & Jones, 1993). While it is conceivable to administer the same test or parallel tests to the same person, a true score is nevertheless an unobservable, theoretical construct. Error scores—another unobservable, theoretical construct—are the random discrepancies between an individual's true score and their observed score. CTT assumes an additive relationship between true scores and error scores (M. J. Allen & Yen, 2002).

To avoid the stumbling block of a single equation with two unknowns, CTT further assumes that: 1) true scores are not correlated with error scores, 2) the average error score across the population of test takers is zero, and 3) error scores on parallel forms of a test are not correlated. The implication of the true score equation is that an individual's overall test performance, rather than performance on specific items, is linked to their true score (Hambleton & Jones, 1993).

Classical Test Theory comes with statistical methods for analyzing both test scores and individual items. The mean and standard deviation are commonly used to analyze test scores while item difficulty indices (p) and discrimination values (r) are used to evaluate items (Hambleton & Jones, 1993). The item difficulty index for item i is calculated as the proportion of test takers who answered the item correctly. A consequence of this combination of mathematical and naming conventions is that easy items (i.e. items that were answered correctly by most of the test takers) have higher difficulty indices than harder items (M. J. Allen & Yen, 2002). An item's discrimination value describes how well the item distinguishes between students with different true scores.

One drawback to statistics associated with CTT is that they are sample dependent and cannot be extended to the general population. Measurements on individual test takers depends on the items included on the test (Hambleton & Jones, 1993). Likewise, statistics on items

depends on the sample of test takers. As an example, an item's difficulty index (*p)* hinges on the rigor of the item's content as well as the capabilities of the students responding to the item (De Champlain, 2010). Furthermore, in CTT an individual's test score is interpreted in the context of a norm group. A norm group is an applicatory sample of people who took the same exam (Embretson & Reise, 2000). With respect to classroom tests, the interpretation of each student's test score depends on the overall performance of the entire class on that test. For example, a test score of 65% would likely be interpreted as evidence of good performance if the class average on the exam was 50%. Conversely, a score of 65% on that same exam would likely be interpreted as evidence of poor performance if the class average on the exam was 95%.

Item Response Theory is a collection of mathematical models and statistical techniques that attempt to model the outcome of an encounter between a person and a test item (Reise, Ainsworth, & Haviland, 2005). The key assumption of IRT is that the probability of a person's correct response to an item can be modeled as a function of item parameter(s) and the extent to which a person possesses a certain latent ability. Essentially, an IRT model relates changes in ability level to changes in the probability of a correct response to an item. Furthermore, a person's responses to test items are used to make predictions about their latent trait(s) (Embretson & Reise, 2000; Molenaar, 1995).

The concept of a latent ability level *(θ)* is unique to IRT. A latent ability is an unobservable trait that is assumed to influence an individual's response to an item that measures that trait (Reise et al., 2005). Latent abilities are hypothetical variables and can take on a variety of forms such as intelligence, multiplication ability (Baker, 2004), and critical thinking skills. Latent ability is measured along a logit scale and higher θ values correspond to more latent

ability. The theoretical continuum of latent ability ranges from - ∞ to ∞; however, in practice the continuum of latent ability tends to range from -3 to 3 (Embretson & Reise, 2000).

The Rasch Model, which is sometimes referred to as a 1 parameter logistic model (1-PLM), is one of the models in the IRT family. The Rasch model models the probability of a correct response as a function of the distance between the person's ability level and the item's difficulty (Wright, 1977). The Rasch Model predicts the probability that a student *j* will answer item *i* correctly through the following equation:

$$p_i(\theta_j) = \frac{e^{(\theta_j - b_i)}}{1 + e^{(\theta_j - b_i)}} \quad (2)$$

In this equation θ_j represents the latent ability level of student *j* and b_i represents the difficulty of item *i*. Item difficulty (b_i) is measured along the same logit scale as latent ability. An item's difficulty level (b_i) is defined as the amount of a latent trait needed to have a 50% chance of getting the item correctly. A person is more likely to fail on an item when the item difficulty exceeds their ability level just as a person is more likely to succeed on an item when the person's ability level is greater than the item's difficulty (Embretson & Reise, 2000).

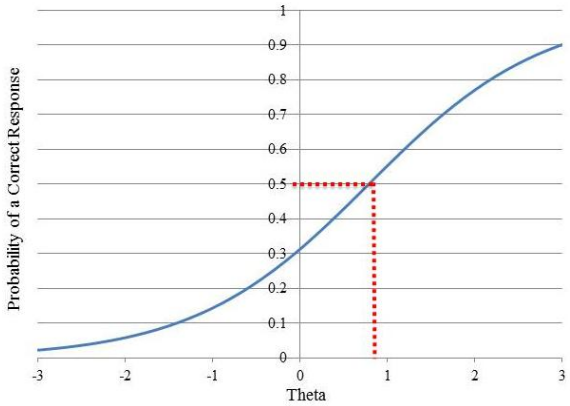


Figure 1.1: Sample Item Response Function with Item Difficulty Location. The x-axis represents the continuum of latent ability that is measured on a logit scale. The y-axis represents the probability of a correct response for that item. The logistic curve depicts the probability that a person with a given ability level (theta) will answer the item correctly. The item's difficulty (b_i) of 0.79 logits is located in red dotted lines.

The output of the Rasch Model—as well as all other IRT models—is an Item Characteristic Curve (ICC), which is sometimes referred to as an Item Response Function (IRF) (Figure 1.1). The ICC is a mathematical function that relates a person's position along a latent ability continuum to their probability of correctly responding to an item that assesses that latent ability (Reise et al., 2005).

More complex IRT models, such as the 2-PLM, 3-PLM, and 4-PLM, build on the Rasch model by progressively incorporating an additional item parameter. The 2-PLM adds an item-specific item discrimination (a_i) parameter to the Rasch model. Whereas the Rasch Model assumes that the ICCs of all items share the same slope at the point of inflection, the (a_i) parameter of the 2-PLM allows each item on the test to have a different slope at the point of inflection. This slope value is also called the item's discrimination. Items with larger slopes are more able to distinguish between people whose ability levels are near the item's difficulty level than items with smaller slopes (Reise et al., 2005). The 3-PLM builds on the 2-PLM by adding in an item-specific pseudo-guessing parameter (c_i). The pseudo-guessing parameter raises the lower asymptote of an item's Item Characteristic Curve to account for the instances whereby low ability students can guess their way to a correct answer on that item. The 4-PLM adds an item-specific "carelessness" parameter (d_i) to the 3-PLM. The "carelessness" parameter lowers the upper asymptote of an item to account for instances in which high ability students err on that item (Linacre, 2004).

IRT is often referred to as a "strong model" because its assumptions are hard to satisfy. In order to use an IRT model, the observed trends in the data must align with the model's predictions. If the data diverge from the model's predictions, the model is not appropriate. Otherwise stated, the key assumption of IRT must be met. Another assumption of IRT is that the

IRT model used fully accounts for the data. The item and person parameters specified in the model are wholly responsible for determining the probability of a correct response. For the aforementioned IRT models, a consequence of this second assumption is the latent ability measured by the test must be unidimensional. The assumption of local independence also implies that each item on a test is considered an independent event (Embretson & Reise, 2000).

The ways in which IRT breaks from CTT come with practical implications. In CTT the item parameters (*p* and *r* values) and person parameter (observed score) are entirely context dependent. The item parameters depend on the sample of test takers and the observed score of a test taker can only be compared to his/her fellow test takers. Without a method of linking or equating the tests, test scores and item parameters from different tests cannot be compared (Hambleton & Jones, 1993). The sample dependent nature of CTT parameters means that results cannot be generalized across testing occasions (Molenaar, 1995). In contrast to CTT, IRT latent ability levels and item parameters are not sample dependent. The ability level estimates for the test takers do not depend on the sample of items they responded to nor are the item parameters linked to the sample of test takers who took those items. As long as the items are calibrated, the same latent trait can be measured with different sets of items thereby eliminating the need for parallel tests. Unlike CTT, the IRT model can generate measurements along an interval scale. When measurements fall along an interval scale, baseline measurements do not have to be identical for changes in test scores to become meaningful. This allows for student progress to be tracked and compared. IRT can also be used to generate shorter tests that have a lower amount measurement error than longer tests. Lastly, scaling item difficulty and latent ability on the same metric allows for person to item comparisons (Embretson & Reise, 2000).

IRT models can be powerful tools for undergraduate science educators who are vested in fostering the critical thinking skills of their students. The ability level estimates obtained on students from an IRT-based test that samples university students' critical thinking skills can provide insight into their capabilities. Furthermore, the non-sample dependent nature of IRT person and item parameters and the interval level of measurement would allow undergraduate science educators to track changes in students' abilities over the course of the semester and over time. Undergraduate science educators could also track trends in students' critical thinking skills across semesters without the need to administer the same items year after year. These are only a few of the potential benefits of applying IRT to university science exams. The full benefits of IRT in undergraduate science education will not be realized until these models are more widely used among science educators.

Despite the many benefits of IRT, it has yet to be widely implemented in university science courses. In this paper we aim to demonstrate that the Rasch Model can be applied to multiple-choice, final exam data from a second semester, introductory undergraduate biology class for science-majors and show that an IRT analysis provides greater insight into the test and the test takers than a CTT analysis.

Description of the Data:

Two semesters, Spring 2008 and Spring 2010, of final exam data from a second-semester introductory biology course for science majors at a large research one institution in the Southern U.S. were obtained. The exam was designed and administered by a faculty member in biological sciences. Course topics included evolution and natural selection, phylogenetics, plant structure and function, animal structure and function, and ecology. Final exam data were obtained from the instructor in the form of excel spreadsheets that contained students' answers (anonymously)

to a set of multiple-choice questions. Each multiple-choice question contained five answer choices (one correct answer and four distractors). A total of 368 students took the Spring 2008 final exam and 469 students took the Spring 2010 final exam. The Spring 2008 exam contained 106 graded multiple-choice items and the Spring 2010 exam contained 118 graded multiple-choice items. In order to satisfy the IRT assumption on unidimensionality, two items were removed from the Spring 2008 exam. The deleted items tested students on their knowledge of active and passive learning rather than biology. All of the items on the Spring 2010 tested students on their knowledge of biology. Thus, the dataset included 118 items from the Spring 2010 exam and 104 items from the Spring 2008 exam. Sixty-two items were identical between the 2008 and 2010 exams. The combined data set yielded a total of 160 items and 837 students.

Methods:

Classical Test Theory Analyses:

Correlation values were obtained in SPSS Version 19.0 (IBM, 2010). All other Classical Test Theory analyses were conducted in Microsoft Excel 2010.

Item Response Theory Analyses:

Winsteps (Version 3.74.0) (Linacre, 2012a) was used for all IRT analyses. Winsteps was chosen because it is user-friendly, readily available, and can cope with missing data. Winsteps is a Rasch only software that relies on Joint Maximum Likelihood Estimation (JMLE). By default, Winsteps centers the item parameters at a mean of 0 logits and a standard deviation of 1 logit. For the IRT analysis, the data from both tests were combined into a single matrix with 160 items and 837 students. The 62 items that were common to both tests served to "link" the two exams. The 42 items from the Spring 2008 exam that were not administered to the Spring 2010 students

as well as the 56 items from the Spring 2010 exams that were not administered to the Spring 2008 students were coded as "9" and treated as missing data.

Results:

Classical Test Theory Results:

The CTT results (Table 1.1) were analyzed to assess each exam as well as each item on the exams. The high Cronbach's alpha values of the 2008 and 2010 exams, suggests that both tests had a high degree of reliability. Likewise, the average point biserial values for the 2008 and 2010 exams suggest that overall, the items successfully distinguished between high and low performing students. However, items with negative discrimination values should be revised. A major deficit with CTT is the lack of adequate methods for comparing students who did not take the same exam. Therefore, the performance of the 2008 students cannot be compared to that of the 2010 students.

Table 1.1 CTT Results for the Spring 2008 and 2010 Final Exams. Two semesters of final exam data were analyzed through Classical Test Theory.

Exam Year:	**2008**	**2010**
Mean:	66.74%	62.31%
Standard Deviation:	13.04	14.34
Range of Item Difficulty Values:	0.962 to 0.158	0.966 to 0.151
Range of Item Point Biserial Values:	-0.008 to 0.466	-0.02 to 0.431
Average Item Point Biserial Value:	0.244	0.257
Cronbach's Alpha:	0.891	0.894

Item Response Theory Analysis Output:

Winsteps output included parameter values and fit statistics for each item and person as well as item estimates for each item. Parameter values are the results of applying the model to the data. The Rasch Model parameter values produced by Winsteps are: a difficulty estimate (b) for each item and an ability estimate (θ) for each student. Winsteps also generates a unique error (SE) value for each b and θ value. The error values can be used to calculate the 95% confidence interval for these parameters. We assume that 95% of the confidence intervals calculated in this manner do indeed encompass the true parameter. Despite the use of the word "error", the error value associated with each parameter is not used to assess whether the model fits the data.

To analyze the extent to which the data on each item and each student conformed to the Rasch Model, Winsteps generated two chi-square based statistics: outfit and infit statistics (fit statistics). Being a stochastic model, the Rasch Model expects there to be a relatively consistent degree of randomness in the data and banks on this inherent randomness when generating an interval scale for the b and θ values. A mean square value for an infit/outfit statistic on a person or item of 1—irrespective of the associated standardized z-score—indicates that the data on the person or item do contain the predicted, uniform level of randomness and do not skew the measurement system. Fit statistics greater than 1.0 are indicative of noise in the data while values less than 1 indicate less randomness than the model predicts (Linacre, 2012b). It has been suggested that the acceptable ranges of mean square values for fit statistics on a High Stakes Multiple-choice test is 0.8 – 1.2. However, informal simulation studies and analyses on hundreds of existing data sets led to the guideline that fit statistics between 0.5 and 1.5 are optimal but an upper limit of 2.0 is begrudgingly allowed (Linacre, 2002; Linacre & Wright, 1994). Fit statistics greater than 2.0 suggest that more than 50% of noise in the data is

unexplained noise. Fit statistics on an item greater than 2.0 suggest that the item does not fall in line with the rest of the items and poses a threat to the assumption of unidimensionality (Linacre, 2009). Fit statistics less than 0.5 do not necessarily degrade the measurement system; rather, they fail to add information (Linacre, 2002; Linacre & Wright, 1994).

The difference between the infit and outfit statistics is that the infit statistic is a weighted statistic whereas the outfit statistic is an unweighted statistic. Weighting the infit statistic by the model variance causes it to be heavily influenced by students' unexpected responses to items that are well matched to their abilities. Conversely, infit statistics are less sensitive to outlier observations. In contrast, as an unweighted statistic, the outfit statistic is heavily influenced by students' unexpected responses to items that are relatively hard or easy for them (Linacre, 2012b).

While fit statistics deal with the influence of each item/person on the overall interval scale of measurement, item indexes shed light on how well each item conforms to the Rasch Model. Item indexes are empirical, *post-hoc* analyses that are not factored into the derivation of the model's parameter estimates. When deriving the parameter estimates, Winsteps assumes that the logistic curve for each item in the test has a slope of 1, a lower asymptote of 0, and an upper asymptote of 1. The *post-hoc*, empirically derived slope, lower asymptote, and upper asymptote serve as rough gauges for how well each item conforms to the Rasch Model and whether or not a more complex model is warranted. The Winsteps item indexes of slope is akin to the discrimination parameter in the 2 parameter model while the indexes of upper and lower asymptote are akin to the pseudo-guessing and mistake-ability parameters in the 3 parameter model and 4 parameter model. Deviations from the assumed values indicate the extent to which

the item deviates from the Rasch Model (Linacre, 2012b) and may suggest the need for a more complex model.

Item Level Data:

Winsteps anchored the mean of the item difficulties at 0 and scaled the difficulties to a standard deviation of 1. The item difficulties ranged from -2.86 to 2.57 (Appendix A). The item level data was used to analyze the extent to which each item fit the Rasch model as well as the extent to which our overall dataset fit the Rasch model.

An analysis of the item fit statistics found that the fit statistics for all 160 items were within acceptable limits. Mean square values for the outfit statistics ranged from 0.73 to 1.29 and mean square values for the infit statistics ranged from 0.89 to 1.20 (Appendix A). These results lend support to the conclusions that, as a whole, the items functioned consistently across the students and that no single item skewed the analysis. These results are evidence that the data are unidimensional enough to be measured by the Rasch model. The range of infit statistics is appropriate for a high stakes multiple-choice test while the range of outfit statistics would be acceptable for a lower stakes multiple-choice test (Linacre & Wright, 1994). Even though this was a high stakes final exam, this is a *post-hoc* analysis of test data that will not be used to determine student scores. Therefore, the more relaxed criteria were implemented and all items were retained in the analysis. Along this line, since all items were used to determine students' grades so the test was analyzed as a whole. Another reason for retaining all items is that high outfit mean square values are less degrading to the analysis than high infit mean square values (Linacre, 2012b).

Having concluded that all items had acceptable fit statistics, the item estimates were combed to determine the extent to which each item conformed to the Rasch Model. Well-fitting

items are those with discrimination values of 1.0, lower asymptotes of 0, and upper asymptotes of 1.0. Two such "good items" are 58 and 62. Item 58 had a difficulty value of -0.14 (SE = 0.08) while item 62 had a difficulty value of -0.07 (SE = 0.08). The items' respective discrimination indexes of 1.03 and 1.09, lower asymptotes of 0, and upper asymptotes of 1.0 are all in accordance with the Rasch model. Outfit and infit mean square values for item 58 were 0.994 and 0.984 and for item 62 were 0.94 and 0.97. The close alignments between the Observed and Expected Score ICCs for these two items provide graphical evidence that the model's predictions for these items fit the data (Figure 1.2 a and b).

Conversely, poorly fitting items are those with discrimination values that deviate from 1.0, lower asymptotes that are greater than 0, and upper asymptotes that are less than 1.0. Two examples of "bad items" that did not conform to the Rasch model's predictions are 95 (b =0.39, SE = 0.11) and 138 (b = 0.84, SE = 0.1) (Figure 1.2 c and d). Mean square values for outfit and infit statistics for item 95 were 1.29 and 1.20, respectively. Item 138 had outfit and infit mean square values of 1.16 and 1.14, respectively. The higher than desired fit statistics showed that these items did not function entirely as predicted when given to well-matched students (high infit) and when given to students who were not matched to the item (high outfit). As with the fit statistics, the item parameters for these two items were not optimal. Item 95 and 138 both show the same pattern of low discrimination values of (0.157 and 0.203, respectively), raised lower asymptotes (0.219 and 0.166), and lowered upper asymptotes (0.807 and 0.805). All of these values are inconsistent with the Rasch model. The low discrimination values show that these items were unable to distinguish between low and high performing students. The pattern of asymptotes suggests that, on the whole, students tended to eliminate one or more distractors and then unsuccessfully guessed at an answer. It is also possible that the majority of students simply

could not decipher items 95 and 138. This may account for their higher than average *b* values. Unfortunately, it was not possible to interview students who took this exam so qualitative data on why these items did not function ideally is unavailable.

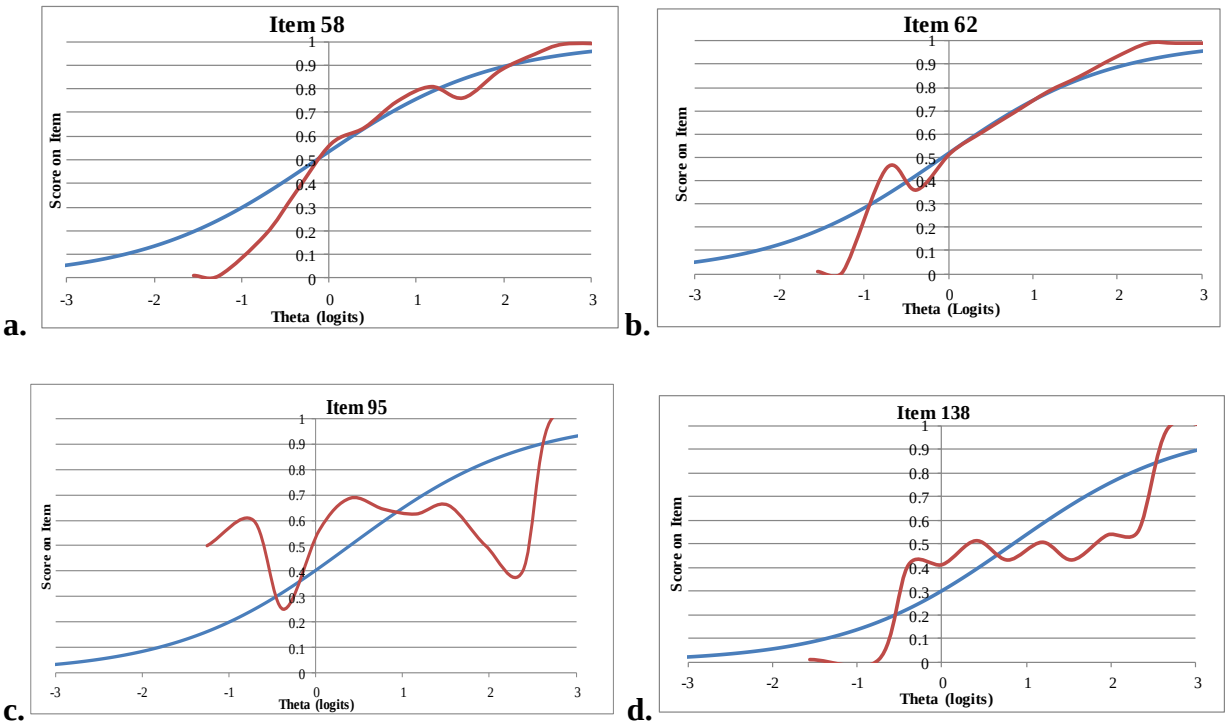


Figure 1.2: Comparisons of Empirical and Modeled Data for "good" and "bad" items. The Expected Score ICC represents the model's prediction of the probability that a student with a given ability level will answer the item correctly (blue curve). The Observed Score ICC is the empirical data of how the students performed on this item (red curve). Student ability level is plotted on the x-axis. **A and B:** Close alignment between the expected and observed score ICCs indicates good model fit. **C and D**: Lack of alignment between the expected and observed score ICCs indicates poor model fit.

Even though not all of our items conformed ideally to the Rasch Model, it is difficult to quantify just how many items are not Rasch-appropriate. Unlike fit statistics, there are no defined guidelines for determining how far an item can deviate from the Rasch Model's predictions before it is deemed inappropriate for the model. Reise and Waller (2003) stated that a guessing parameter (lower asymptote) greater than 0.10 is "substantial". However, they admit that this designation is wholly arbitrary (Reise & Waller, 2003). According to this criterion 32 of

the 160 items had "substantial" lower asymptotes but had acceptable upper asymptotes. If the inverse of this criterion is applied to the upper asymptote values then 5 of the 160 items had a substantially low value for their upper asymptote and an acceptable lower asymptote. Three of the items suffered from substantial lower and upper asymptotes (items: 95, 106, & 138). Taken together, the data show that deviations from the Rasch asymptote were localized to a subset of the items. Therefore, even if the sample size was large enough, neither a 3-PLM nor a 4-PLM would be appropriate.

The data showed that non-uniform discrimination values posed the biggest threat to the ability to apply the Rasch Model to our data. Discrimination values that deviate from 1.0 indicate that the item does not conform exactly to the Rasch Model; however, there are no suggested cutoff points. The empirical discrimination values of our items ranged from 0.01 to 1.49. 33 of the items had discrimination values less than 0.9 while another 33 items had discrimination values greater than 1.1. It should be noted that the range of discrimination values produced by Winsteps tends to be wider than the range of *a* parameters generated by applying a 2 parameter model to the same set of data. The reason for this discrepancy is that, in order to estimate the *a* parameters, many software programs constrain the range of *a* parameters (Linacre, 2012b). Therefore, the spread of discrimination should not be interpreted as direct evidence that a 2-PLM would be a better fit. In the case of our dataset, adding in a discrimination parameter would merely be a weak band-aid for items with gaping wounds. The discrimination values of some of the items, such as item 106 (Discrimination index = 0.010), were entirely unacceptable; therefore, instead of adding in an additional parameter we feel it would be best to either revise these items or eliminate them from the test bank. A better approach would be to eliminate or

revise items that over or under discriminate. Once again, it was concluded that the Rasch model was the most appropriate model for the dataset.

Person Level Data:

Having concluded that the Rasch Model is appropriate for our dataset and that all items should be retained in the dataset, the analysis shifted to the students who took the exam. The ability levels (θ values) of the 837 students ranged from -1.55 to 3.38 and averaged at 0.77 (SE = 0.02)**.** Thus, the overall student ability was slightly greater than the average item difficulty of zero. Winsteps also produced a person reliability estimate of 0.89. The Winsteps person reliability estimate is akin to the classical test theory concept of reliability. High person reliability estimates indicates that it is highly probable that students with higher ability levels do in fact have higher ability levels than students with lower ability levels. In other words, the test successfully stratified the high and low ability students (Linacre, 2012b). Furthermore, a person reliability value of 0.89 and a person separation of 2.91 imply that the test was capable of stratifying students into 3 levels.

As with the items, the extent to which the data on students fit the Rasch model was investigated. The Rasch Model predicts a high probability of a correct response when the student's ability level is greater than the item's difficulty and predicts a low probability of a correct response when the item difficulty value is greater than the student's ability level. For example, student #221 (θ = 1.82, SE = 0.28) followed the model's prediction by answering item 17 (b = 1.34, SE = 0.08) correctly and by answering item 9 (b = 2.57, SE = 0.1) incorrectly. In contrast, student #60 (θ = 1.74, SE = 0.27) troubled the model on items item 41 (b = -2.01, SE = 0.14) and item 81 (b = 2.21, SE = 0.13). Despite having a 98% probability of responding to item 41 correctly, student #60 responded to item 41 incorrectly. Additionally, the probability of

student #60 responding correctly to item 81 was only 38% and yet student #60 answered that item correctly. With the exception of a few blunders and unexpected correct answers, students 34 (θ = 1.17, SE = 0.24) and 570 (θ = 0.66, SE = 0.21) followed the model's predictions (Appendix B). Both students answered the majority of questions below their ability level correctly. Their percentage of correct answers dropped as the difficulty level of the question approached their ability level and reached zero when the difficulty level of the questions far exceeded their ability level.

Once again, fit statistics were used to quantify the extent to which the empirical data match the model's predictions. Outfit statistics for the students ranged from 0.582 to 1.75 and the infit statistics ranged from 0.804 to 1.31. Since the amount of data on each item was greater than the amount of data on each student, the fit statistics on the students were not as tight as the fit statistics on the items. Another reason for looser student fit statistics is that items tend to be more predictable than students (Linacre & Wright, 1994). The high range of outfit statistics could be due to lucky guesses by students who were much less capable than the item and/or blunders by students who were much more capable than the item. As with the items, it was concluded that the fit statistics on the students were acceptable and all students were retained in the analysis.

Whereas CTT assumes that each test score is associated with the same error value, IRT determines the error value associated with each ability level estimate. The error value quantifies how precise an estimated value is and precision decreases as error values increase. Determining the error associated with each ability level estimate is an acknowledgment that the test is not equally able to estimate each student's ability level. Analyzing the error associated with each ability level revealed the range of ability levels over which this test was most able to measure

(Figure 1.3). The data showed that the students whose ability level estimates were $0.8 < \theta < 1.0$ had the least error associated with their ability level estimates while students whose ability level estimates were $1 < \theta$ or $\theta < 0.8$ had the most error associated with their ability level estimate. To relate this to assigning grades from a CTT "curve" perspective, since the mean ability level of the students was 0.77 the test was most able to measure and classify students at the B, C, D grade range (above average, average, below average). The test was less able to distinguish between the A and B students. Given the importance of the coveted A grade, the test should be revised so that it is better able to measure students with high ability levels.

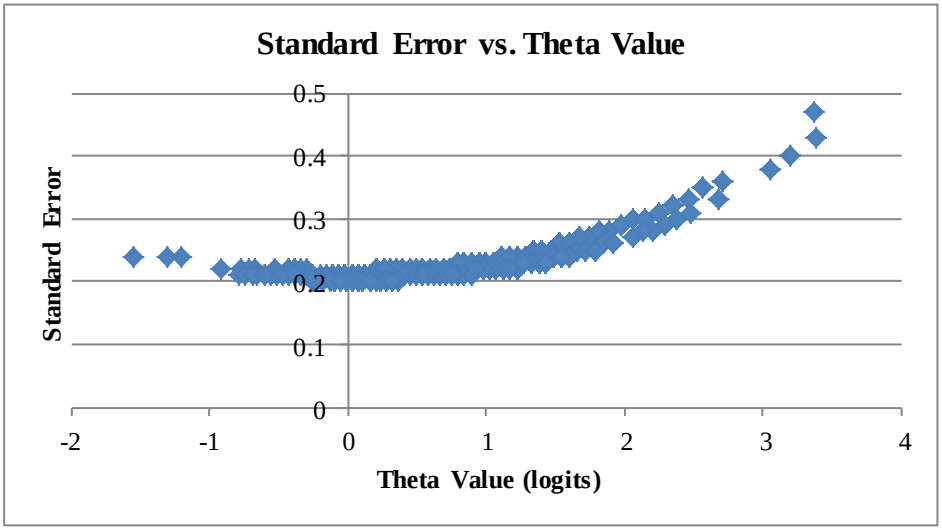


Figure 1.3: Standard Error versus Theta Value. Lower standard errors indicate a higher degree of accuracy of the theta value estimate. Theta value estimates were most accurate for students of average ability and accuracy tended to decrease as ability increased.

Unlike CTT, IRT allowed for a comparison of the average ability level of students across years. The method of Linacare (2012) was used to compare the average ability of the 2008 students to the 2010 students. The entire dataset was first analyzed with the mean ability level (θ) of the 837 students anchored at 0 logits. The item *b* values obtained from this analysis were used to anchor separate analyses of the 2008 and 2010 students. The person θ values produced

by the separate analyses therefore lie on the same reference frame that was defined by the item anchor values (Linacre, 2012b). The average ability for the 2008 students was 0.106 (S.D. 0.713) and the average ability of the 2010 students was -0.082 (S.D. 0.687). An independent samples t-test with equal variances assumed ($F = 2.225$, $p = 0.136$) showed that these two means were significantly different ($t = 3.88$, $df = 835$, $p < 0.001$) and that the Spring 2008 students demonstrated a greater overall ability level than the Spring 2010 students.

Tests of IRT Invariance:

A key principle of IRT is that model parameters are sample independent. The ability level estimates for the test takers do not depend on the sample of items to which they responded nor are the item parameters linked to the sample of test takers who took those items (Embretson & Reise, 2000). This statement should be tempered by noting that absolute invariance of model parameters only occurs when the model and data are an exact match. Since the data never form an exact match to the model, absolute invariance does not occur. Rather, researchers must assess the degree to which the model parameters are invariant (Hambleton, Swaminathan, & Rogers, 1991). The extent to which the principle of invariance holds up is a function of the overall model-data fit (De Ayala, 2010).

The correlation coefficient invariance approach was first used to test the invariance of the 62 items that were common to both exams. The 837 students were randomly divided into calibration groups A and B. Due to the randomization, 50% of the 2008 students were in calibration group A while the other 50% of the 2008 students were in calibration group B. Similarly 50% of the 2010 students were in calibration group A while the other 50% of the 2010 students were in calibration group B. Separate Winsteps analysis of the 62 items that were common to each exam were conducted for each calibration group. By default, Winsteps fixes

the mean of the difficulty values at zero. Therefore, in order to test the invariance of the items, the mean of person ability levels was set to zero for both calibration runs. Fixing the mean of the θ values at zero set a common metric for the item difficulty parameters across the two runs. The correlation between the item difficulty parameters of the 62 common items between the A and B calibration groups was $r = 0.980$ ($p < 0.01$) and a linear relationship was observed between the two sets of item difficulty parameters (Figure 1.4a). This analysis was repeated and the two calibration groups were used to estimate the item difficulty values for all 160 items. The correlation value for all 160 items across the two calibration samples was $r = 0.973$ ($p < 0.01$). Since a correlation value of 0.9 or greater is considered evidence of item invariance (De Ayala, 2010), these data show that the principle of item invariance was realized in the dataset and that the data fit the Rasch model.

The invariance of the theta values was assessed though separate estimations of theta values using the odd and even numbered items (Hambleton et al., 1991). Theta values for all 837 students were first estimated using only the odd numbered items. The theta values for all 837 students were then estimated using only the even numbered items. The item difficulty values for the "odd test" ranged from -2.99 to 2.47 while the item difficulty values for the "even test" ranged from -2.37 to 2.12. Both tests had an average item difficulty level of zero. A correlation of 0.800 ($p<0.01$) was obtained between the set of theta values obtained using the "odd-test" with the set of theta values using the "even-test". This correlation is neither high enough to merit substantial evidence of theta parameter invariance nor is it low enough to regard it as evidence of poor model-data fit. Rather, these data reflect the need for proper test design (Hambleton et al., 1991). The scatterplot of the two sets of estimates (Figure 1.4b) showed that, on the whole, there is a strong linear relationship between the two sets of theta estimates, which did indicate a good

degree of observed invariance. However, the linear relationship broke down at the ends of the theta scale, particularly at the positive end of the theta scale. The inconsistency of the ability level estimates for students at the extremes of the ability level spectrum—especially at the positive end of the spectrum—was due to the high amount of measurement error associated with their ability level estimates (Figure 1.3) as a result of the test's inability to assess these students. Even though there are students whose ability levels exceed $\theta = 3$, the most difficult item on the exam was $b_9 = 2.47$. Thus, the test did not contain questions that were difficult enough to measure these students' ability levels. Therefore, the theta estimates of the high achievers as measured by this exam were moving targets and were not consistent across item samples. Overall these data show that the degree to which invariance is observed hinges on proper test design. Nevertheless, these data are encouraging as they showed the potential for IRT to produce estimates of student ability levels with a high degree of invariance.

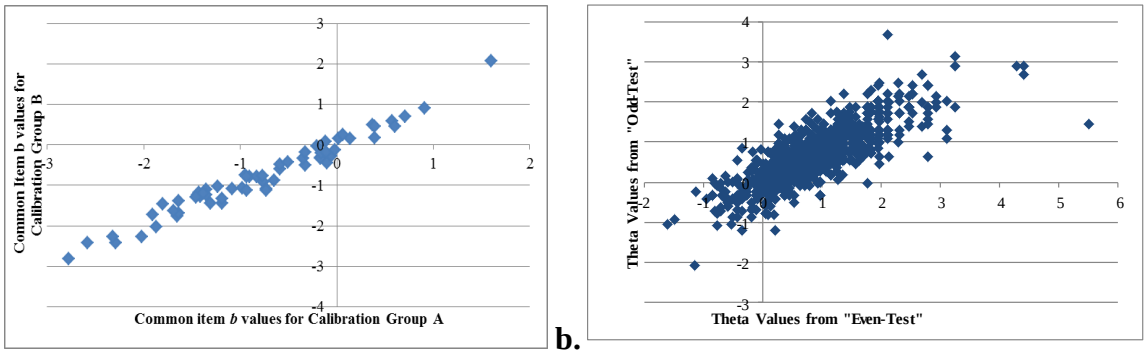


Figure 1.4: Tests of IRT parameter invariance. a) The *b* values for the common items using two random samples are crossplotted. The strong linear trend is evidence of item parameter invariance. **b)** Theta values for all students estimated using the even numbered items are crossplotted against the theta values for all students estimated using the odd-numbered items.

Test Validity Data:

The third phase of the IRT analysis focused on the overall ability of the test to measure student ability levels and to identify ways to improve the ability of the test to measure students.

In this sense, IRT was used to investigate the validity of the test. Even though the "test" analyzed was actually data from two final exams that were combined, they were nevertheless analyzed as a single test. Furthermore, all test validity statements and suggestions for improvements were made as though this was a single test. The rationales for doing so were to demonstrate the method for using IRT to improve tests and to provide a possible starting point for designing future final exams from a Rasch Model perspective.

The person-item bar chart (Figure 1.5) was then analyzed to determine how to revise the test so that it can more precisely measure students with high ability levels. Each item that a student responds to provides information about that student's ability level (θ). The amount of information provided by an item increases when the item's (b) value approaches the student's (θ) value. Conversely, the amount of information provided by an item decreases when the (b) and (θ) values diverge. In order for the test to measure the ability of all students, the distribution of item difficulties should be well-matched to the distribution of students' ability levels. The person-item bar chart for this exam showed a degree of mismatch between the items and students that needs to be corrected. Whereas the test contained a glut of items at the $b_i < 0$ end of the spectrum, there was a dearth of students at the $\theta < 0$ end of the spectrum. This indicates that the number of items with a difficulty value less than $b_i = 0$ can be pared down. The person-item bar chart echoed the results in Figure 1.3 by showing the inadequate number of questions with difficulty levels greater than $b_i > 1.5$. Increasing the number of challenging items on the exam would serve to differentiate the A and B students. Increasing the number of items in the $0.5 < b_i < 1.5$ range would also serve to better differentiate the students of average ability. Overall, the person-item bar chart showed that the difficulty level of the items needs to be increased.

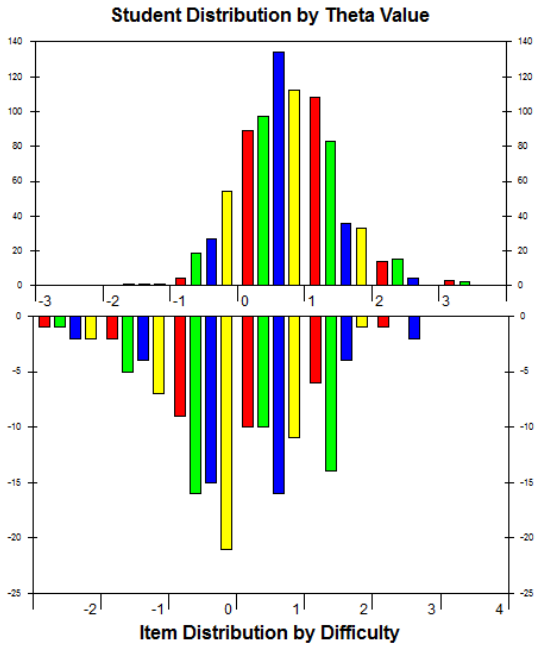


Figure 1.5: Person to Item Bar Chart. The upper panel displays the distribution of students by their ability level. The lower panel displays the distribution if items according to their difficulty scale. Item difficulty and student ability level are both measured along the central x-axis.

The test information function complements the person-item bar chart by depicting the extent to which the exam as a whole was able to gather information about the students' abilities. This can serve as an overall validity check that the test was appropriate for the sample of students to whom it was administered. The amount of information an entire test obtains about a students' (θ) value is simply the sum of the information each item on the test provides about a student of a given (θ) value. The overlay of the test information function with the histogram of student performances illustrated how well the test as a whole captured information on the students who took the exam (Figure 1.6). From this illustration it was determined that the ability range over which the test functioned optimally encompassed the majority of the students; however, the test would have been better suited to a sample of less able students. This result is consistent with earlier results that: 1) the person reliability and person separation estimates that

suggested the test was capable of stratifying students into three levels and 2) ability level estimates (θ) for students of average ability were associated with lower standard errors than students with above average ability level estimates (θ) and 3) the lack of items at the upper end of the item difficulty scale. Whereas the person-item bar chart provided a detailed insight into how well the items were matched to the students and how to better match the item difficulty to the students, the overlay of the histogram with the test information function provided a holistic view of the ability to capture information of students across the spectrum of ability levels. Relying solely on the person-item bar chart would lead to the erroneous conclusion that the test was entirely unable to obtain information on students whose ability level is $\theta > 2.5$.

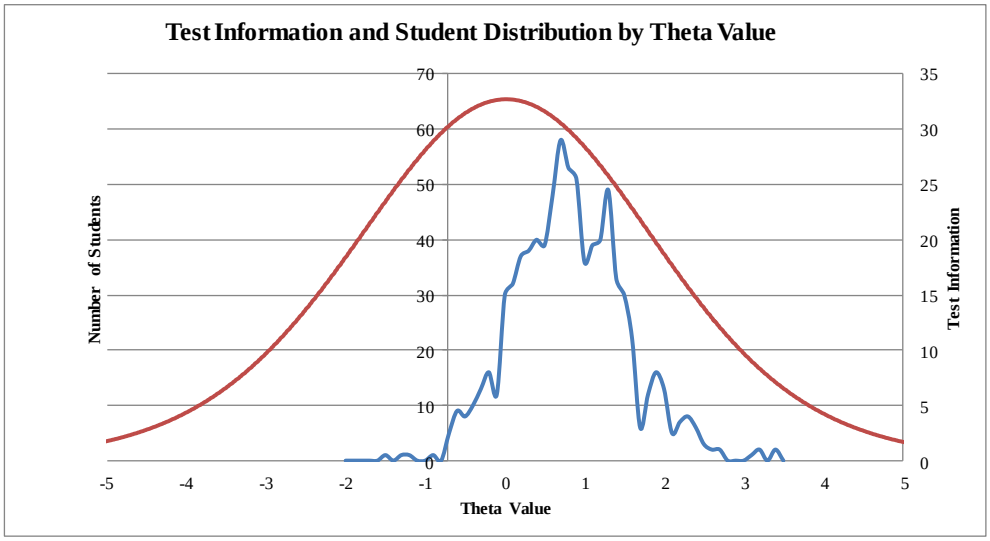


Figure 1.6: **Test information versus student ability level**. The x-axis represents student θ values and the central y-axis represents the total test information. The right y-axis represents the number of students. The red test information function represents the amount of information the test captures at each ability level while the blue histogram depicts the distribution of students by θ value. A comparison of the two functions shows that range of θ over which the test was most able to capture information captures the majority of students.

CTT Parallels to IRT:

Lastly, a comparison of the IRT and CTT data showed that some of the results produced by the two analyses were congruent. A correlation of -0.980 ($p<0.01$) was obtained between the IRT *b* values and the CTT Item Difficulty values. The negative correlation is due to the inverted scales used by the two measurement theories. Difficulty values in CTT range from 0 to 1and increasing difficulty values indicates easier items. In contrast, *b* values in IRT typically range from -3 to 3 and higher *b* values indicate more challenging items. The IRT generated person reliability estimate of 0.89 was on par with the Cronbach's α values for the Spring 2008 and the Spring 2010 exams (0.894 and 0.891, respectively). Both methods can also be used to produce estimates of item discrimination; however, the issue of whether CTT estimates of item discrimination can be directly compared to (a_i) parameters remains controversial. Even for those who agree that item CTT item discrimination values are akin to IRT discrimination values, there is no easy rubric or rule for equating or comparing the two. As for the students, both methods provided each student with a total score and separated the students into high and low performing groups. Since the Rasch Model uses a student's total score as a sufficient statistic for their ability estimate, the two values are highly correlated. The correlations between the Spring 2008 test scores and the IRT θ values and Spring 2010 test scores and IRT θ estimates were both 0.99 ($p < 0.01$) and the correlation between the total scores from the combined exam and the IRT θ estimates was 0.943 ($p < 0.01$). The lower correlation between the θ values for the individual tests with the combined exam is the result of the 2008 and 2010 exams having different total scores. These data show that the IRT results do not contradict the CTT results.

Discussion:

Appropriateness of Model Selection:

Classical Test Theory is often referred to as a weak model because the assumptions behind the model are relatively easy to satisfy. Therefore, we are confident that CTT can be applied to this set of test data. However, the ease of application comes with some trade-offs. When applying CTT one must consider the test as a whole. A person's true score is linked to their performance on the entire test and cannot be applied to their performance on specific items or groups of items. Another disadvantage is the sample dependent nature of CTT data. CTT does not allow for direct comparisons of different exams nor does it allow for comparisons of students who did not receive the same exam. While it does come with a more stringent set of assumptions to satisfy, IRT does not have the drawbacks associated with CTT.

For an IRT model, such as the Rasch Model, to be applied to a dataset the observed data must align with the model's predictions. That is to say, the model must fit the data. Another assumption of IRT is that the person and item parameters considered in the model are solely responsible for determining the probability of a correct response. A consequence of this second assumption is that test must be unidimensional for the latent ability it measures. So a biology exam must assess biology and only biology. Lastly, each item on an exam must be an independent event and cannot be tied to other items on the test (Embretson & Reise, 2000). However, a test is never purely unidimensional. Additional dimensions are always present in the data. Therefore, researchers must assess whether their data are unidimensional enough for the IRT model to be appropriate (Linacre, 2009).

The design of the test is the first source of evidence that Rasch model did indeed fit the data from the final exams. The design of the test serves as the first source of evidence for this

conclusion. Each item on the Spring 2008 and Spring 2010 exam was an independent event and was not linked to any other item on the test. Furthermore, the test items were restricted to the content taught in the biology course. The professor took great care to focus the items on the exams to the material taught in class and the information contained in the assigned readings. Admittedly, the tests contained a diverse array of topics. Nevertheless, the content on the tests was restricted to biology content of a single course. While it may seem counterintuitive that a test that covers a broad range of topics could be unidimensional enough to fit the Rasch Model, prior research shows otherwise. The Biological Science section of the Medical College Admission Test (MCAT) tests contains is a mix of 68-70% biology items and 30 – 32% organic chemistry items (Childs & Oppler, 2000). The biology items cover a broad range of topics including: molecular biology, microbiology, eukaryotic cell biology, genetics, evolution, comparative anatomy, and vertebrate biology (immune, lymphatic, endocrine, muscular, nervous, digestive, and cardiovascular systems). The organic chemistry items cover topics of: covalent bonds, molecular structure and spectra, hydrocarbons, oxygen-containing compounds, amines, and biological molecules (AAMC, 2009). A dimensionality analysis of the Biological Sciences section of the MCAT concluded that, while there was some evidence of multidimensionality in the data, the multidimensionality in the data had a negligible impact on the calibration of the MCAT item bank. Furthermore, the multidimensionality in the data did not impact students' relative score estimates (Childs & Oppler, 2000).

The observations that the fit statistics on the items and students were within acceptable limits and that the item parameters for the common items were invariant across calibration samples provide quantitative support that the data fit the Rasch model. The fit statistics on the items shows that while not all of the items conformed ideally to the Rasch Model, the amount

and extent of the deviations from the model were not severe enough to disrupt the measurement. With the exception of four students who showed outfit statistics greater than 1.5, the infit and outfit statistics on the students all fell within acceptable limits. It should be noted that the outfit statistics on the four students did not exceed 2.0 and therefore it is unlikely any of these students alone disrupted the *post-hoc* analysis. Admittedly, had this analysis been used to assign grades, a more stringent criteria would have been needed. Implementing the more stringent criteria would mean first re-running the analysis without the students who showed high outfit mean square values. The *b* values from the abridged analysis would serve as anchors for another analysis of the entire student set. This method mitigates the influence of the student outliers on the dataset. The same corrective method can also be applied to items with poor fit statistics (Linacre, 2012b). Even though the invariance data on the ability level estimates was inconclusive, the degree to which this mis-matched test showed invariance of ability level estimates is encouraging.

IRT Extensions:

The IRT analysis provided much more insight into the quality of the test and the students who took the exam than the CTT analysis. Use of IRT provided an interesting insight into the students who took the exams. The results showed that the 2008 students demonstrated a higher ability level than the 2010 students. While the class average on the 2008 exam was higher than the class average on the 2010, a CTT analysis does not allow for comparisons between the two groups of students. Therefore, cannot be used to distinguish whether the increase in the class average is due to smarter students or easier items. Thus, another advantage of IRT is the ability to link and scale tests.

The ability of IRT to situate student ability levels along the same metric as item difficulty values provided greater insight into the overall validity of the exam and suggested ways to improve the

exam. Multiple lines of evidence showed that the test is only able to precisely estimate the ability level of students who were average and below average (B-C-D range of grades). This poses a threat to the validity of a test that was used to assign five levels of grades (A, B, C, D, & F). Since grades of D and F are both considered failing, improving the test's ability to distinguish between students in the D and F categories is moot. The crucial improvement to make is to improve the ability of the test to differentiate students it the A/B range. The person-item bar chart was therefore used to identify easy items that can be removed from the exam and to identify the range of difficulty levels where more items are needed.

It may seem counter-intuitive to suggest increasing the difficulty level of the questions on a test in which the class average was slightly above 60%. However, the goal of testing from an IRT perspective is to measure student abilities (Linacre, 2012b). Given this goal, making a hard test even harder is the appropriate course of action as it would reduce the standard errors associated with high scores. By reducing the standard errors associated with high scores the test would be able to accurately and fairly distinguish between A students and B students.

Conclusions:

This endeavor was a *post hoc* analysis of pre-existing test data. Therefore, these results serve as a starting point for future test development. Several issues need to be grappled with when designing and analyzing tests of this sort in the future. Even though it was concluded that the test data did meet the assumption of unidimensionality, it cannot be inferred that biology is a unidimensional construct. Biology encompasses a broad field of areas and specializations. It remains to be determined whether fields such as molecular biology and ecology rely on a single latent ability. More research in this area is needed.

Researchers also need to carefully consider the model they choose to apply to their data. The Rasch Model was applied to this dataset because it was determined that the Rasch Model was the most appropriate IRT model for the test design and sample size. However, the most common reason for items not conforming to the Rasch Model was that their estimated discrimination values differed from 1.0. This result suggests that future researchers and professors may want to consider using 2-PLM; however, a caveat with adopting 2-PLM for assigning ability levels and grades should be mentioned. Unlike the Rasch Model, the 2-PLM does not use a student's total score as a sufficient statistic for their ability level estimate. Therefore, when using a 2-PLM two students can attain the same total score on an exam but receive different ability level estimates and subsequently will receive two different grades. The added benefits of IRT outweigh the time needed to sort through the considerations. One such benefit of using IRT is for determining standards-based cut-off values for grades. Many professors assign letter grades based on a curve. One flaw with the method of the curve is that it is wholly sample dependent. So rather than being based on the standards a professor set for the class, grades are largely determined by overall student performance. IRT can potentially be used to assign grades based on more objective, standards based criteria as well as a realistic assessment of the range of ability over which the test can accurately measure.

In addition to using IRT for assigning grades, professors can also use IRT to track student progress. The interval level nature of the *b* and θ scale allows change scores to become meaningful. As an example, student who began the semester at $\theta = 0.5$ and ended the semester at $\theta = 1.5$ made twice as much progress as a student who began the semester at $\theta = 2.0$ and ended the semester at $\theta = 2.5$. However, we cannot say that a student with $\theta = 2.0$ is twice as smart as a student $\theta = 1.0$.

Along this line, the sample independent nature of IRT parameters allows instructors to link and scale tests. As long as the items are calibrated, the same latent trait (i.e. Biology ability) can be measured with different sets of items (Embretson & Reise, 2000). This opens the door for professors to compare groups of students and to administer make-up exams that are not identical to the in-class exam and to administer different sets of items to students across testing sessions.

The principle of invariance also lends itself to a fairer framework for testing. CTT acknowledges that a student's score depends on the specific items they were given. In contrast, an IRT framework for test design requires professors to build tests with enough items of varying difficulty to measure the ability level of each student. This can lead to shorter exams that more accurately assess the ability of each student.

Another potential application of IRT is to provide diagnostic information on student capabilities. Should a test be laced with validated items that require critical thinking skills, the data from the exam could be used to assess each student's ability to think critically though biology related test items. Granted this type of testing is still in its infancy; however, given the importance of critical thinking skills it is a worthwhile area of research to pursue.

Overall it was shown that the Rasch Model can be applied to our dataset of multiple-choice, final-exam test data from a university biology class and that the Rasch analysis was far more informative than the CTT analysis. Whereas CTT item and person parameters are entirely sample dependent, our Rasch analysis demonstrated the IRT principle of invariance was realized in our dataset. The invariance of our item difficulty levels allowed us to directly compare the 2008 and 2010 students and determine that the 2008 students outperformed the 2010 students. The IRT analysis also provided greater insight into the validity and resolving power of the test than the CTT analysis. The analysis of the test information function-student histogram overlay,

standard error vs. theta value graph, and person-item bar chart showed that the difficulty level of the items on the test needs to be increased so that the test can distinguish the B students from the A students with precision.

The amount of research still needed before IRT can be readily applied in university classrooms is great. However, these conclusions show that added information gained through IRT on test validity, item design, and student capabilities more than justify the efforts.

CHAPTER 2

The Developlment and Validation of Multiple-choice, Critical Thinking Test Items for an Undergraduate Biology Course.

Abstract:

In response to recent data that many undergraduate-level, multiple-choice biology exams contain a glut of factual recall questions, this study aimed to generate a set of multiple-choice test items that require critical thinking skills for a second semester biology course for science majors. It was hypothesized that each item in a pool of 41 items written by the author required critical thinking skills. To test this hypothesis, an expert panel of five faculty reviewers rated the cognitive demand of each item and the items were tested in cognitive think-aloud sessions with undergraduate students majoring in a life science. Results from the faculty reviews and cognitive think-alouds provided strong evidence that 32 of the 41 items required critical thinking skills and weak evidence for two of the items. Six of the 41 items fell short of the goal of requiring critical thinking skills or were flawed and results for 1 of the items were inconclusive. This study showed the importance of validating the cognitive complexity of multiple-choice test items and demonstrated a method for doing so.

Introduction:

With its emphasis on evidence and data, it is reasonable to think that science would be a discipline in which students would learn critical thinking skills. Unfortunately, undergraduate science education also fails to emphasize critical thinking skills (Alberts, 2009; Ennis, 1985; Ennis, Millman, & Tomko, 1985; Lord & Baviskar, 2007; Weld, Stier, & McNew-Birren, 2011;

White et al., 2011). Bruce Alberts, the lead author of the textbook tome *Molecular Biology of the Cell*, placed the majority of the blame for the lack of critical thinking skills in science education on scientists. According to Alberts, undergraduate science education fails to model for students the evaluation of scientific data and explanation, participation in scientific discourse and practices, an appreciation for the nature of and development of scientific knowledge, and the use and interpretation of scientific knowledge. Instead of modeling these essential skills for students, undergraduate science education focuses on filling students with factual knowledge (Alberts, 2009).

Recent studies support Alberts' allegations by showing that undergraduate science assessments tend to contain mostly factual recall questions (Lord & Baviskar, 2007; Momsen, Long, Wyse, & Ebert-May, 2010). Since professors of most undergraduate-level introductory science courses have at least 200 – 300 students and are given little, if any, instructional support, the ease of scoring multiple-choice exams should not be brushed aside (Heyborne, Clarke, & Perrett, 2011; Tomanek & Montplaisir, 2004). However, four years of factual recall questions generate college graduates who are unable to demonstrate an understanding of the information they committed to memory (Lord & Baviskar, 2007).

If science educators are intent on teaching their students critical thinking skills, then they need methods to assess students' critical thinking skills (Bissell & Lemons, 2006; Crowe, Dirks, & Wenderoth, 2008). Not only do assessments provide information for instructors, they also steer student learning. When assessments focus on memorization and recall, students tend to stagnate at these skills. However, when assessments require critical thinking, students make efforts to rise to the challenge (Dancy & Beichner, 2002).

Before science educators can assess their students' critical thinking skills, they must first define "critical thinking skills". Levels 3 – 6 of Bloom's taxonomy have become the consensus definition of critical thinking used by college science educators (D. Allen & Tanner, 2002; Bissell & Lemons, 2006; Crowe et al., 2008; Momsen et al., 2010). Bloom's taxonomy was first published in 1956 and was later revised in 2002. In addition to drawing boundaries around critical thinking skills, Bloom's taxonomy provides educators with shared terminology about learning objectives, a framework for setting learning objectives, a method of assessing the alignment between assessments and learning objectives, and a broad conception of the broad scope of educational goals (Krathwohl, 2002). Bloom's taxonomy assumes that learners engage in distinct thinking behaviors and that these thinking behaviors vary in cognitive complexity (D. Allen & Tanner, 2002). The revised definitions of the Bloom's taxonomy levels of thinking in order of increasing complexity are: remember (retrieving the correct information from the long term memory), understand (deciphering the meaning of information such as written material and images), apply (transferring or using information in a new situation), analyze (dividing material into its component parts and relating the parts to each other as well as the whole), evaluate (using information and standards to judge material, and create (producing an original product or combining elements to yield a novel product). The revised version of Bloom's taxonomy classifies remembering and understanding as lower-order thinking skills and thinking skills as apply, analyze, evaluate, and create as higher-order thinking skills (Bloom, Engelhart, Furst, Hill, & Kratwohl, 1956; Krathwohl, 2002). For a multiple-choice test item to be classified as a critical thinking item it must fall within levels 3 – 5 of Bloom's taxonomy (Krathwohl, 2002). Since multiple-choice items do not give students opportunities to generate a novel answer, they cannot tap into the critical thinking skill of "create" (Crowe et al., 2008).

Writing multiple-choice items that require critical thinking skill is not a simple, straightforward process; therefore, science educators should investigate the validity of the items on their exams both during and after the development phase (NRC, 2001). Validity is classically defined as an overall consideration of the extent to which the interpretations based on test scores are supported by empirical evidence on the test and the theoretical rationale(s) on which the test is based. According to this definition of validity, tests themselves are neither valid nor invalid. Validity calls into question the interpretations of the test scores. The argument based nature of validity means that validity can neither be absolutely proven nor can it be absolutely disproven. Validity is about accumulating evidence to support the use of a test and the interpretations made based on test scores (Messick, 1995). The implication of the concept of validity for science educators is that they cannot use exams to make inferences about their students' cognitive capabilities or about the effectiveness of their ability to teach critical thinking skills unless they can provide sufficient evidence on the cognitive validity of their exams. That is to say, science educators need to validate that the critical thinking items on their exam tap into and assess critical thinking skill(s).

Cognitive think alouds—also referred to as Concurrent Verbal Protocols—are a National Research Council accepted method for establishing the cognitive validity of test items (NRC, 2001). During a cognitive think-aloud test takers are asked to "think aloud" as they solve test items. The essence of this method is that participants are to continuously report the contents of their short-term memory as they work. It has been observed that simply asking people to "think aloud" does increase the time required to complete the exam but does not alter test taker's thought processes (Norris, 1990). The integrity of the data from cognitive think-alouds hinges on the researcher's ability to remain innocuous and to minimize the extent to which they

influence or direct the participant's thought processes. When conducting a cognitive think-aloud, the researcher should only non-intrusively remind the test taker to verbalize everything they are thinking (Norris, 1990; Tan, 2008). Data from cognitive think-alouds can uncover the information recalled and the solution path used by a problem solver during a task (Taylor & Dionne, 2000). While test takers do not share everything they think, cognitive think-aloud data can reveal if students understand the item and whether or not they applied the correct scientific knowledge to the item (Tan, 2008). These data are then used to make inferences about the extent to which the items tap into the intended thinking skills.

Data from faculty reviewers is another commonly used method of validating the thinking skills of multiple-choice test items. This method has been applied to multiple-choice test items from undergraduate science exams (Momsen et al., 2010), graduate school entry tests (Zheng, Lawhorn, Lumley, & Freeman, 2008), and medical school exams (Simpson & Cohen, 1985). Faculty reviewers are asked to review the multiple-choice items and either determine the specific Bloom's taxonomy level or discern the level of thinking skills required by the item.

The combined approach of cognitive think-alouds and expert reviews can be a useful method for science instructors who, due to large class sizes and little grading support, need to rely on multiple-choice exams but do not want to administer exams that emphasize recall. It has been observed that multiple-choice items tend to elicit low level thinking skills whereas constructed response items tend to elicit higher-order thinking skills; however, these tendencies do not reflect inherent properties of the item formats. Rather, they reflect how these item formats are often implemented (Martinez, 1999). Multiple-choice items can elicit higher-order thinking skills such as application, analysis, evaluation, prediction, and problem solving (Crowe et al., 2008; Martinez, 1999). By testing their multiple-choice items in cognitive think-aloud

sessions, science instructors can validate whether their items elicit critical thinking skills and gain insight into how to write multiple-choice test items that require critical thinking skills. Multiple-choice questions that elicit critical thinking skills for college science educators represent a viable option for college science educators who aim to develop their students' critical thinking skills but do not have the capacity to grade hundreds of free response questions.

The aim of this project was to write and validate a set of multiple-choice test items that elicit critical thinking skills in students in a second semester biology course for science majors at a large, public research institution in the Southeastern United States. Validated test items were eligible for the Spring 2012 final exam. Course topics included: evolution and natural selection, osmosis and diffusion, phylogenetics, animal physiology, plant biology, and ecology.

Methods:

Developlent of Test Items:

All test items were written by the author of this paper during the Fall 2011 semester. The author took several measures to ensure that students could not rely on pure recall to solve the items. The author of this paper attended the biology course during the Fall 2011 semester while writing the exam items. Attending the course allowed the author to target the test items to the class and ensured that students could not solve the items by recalling what the professor taught in class. The author also read all associated text and supplementary readings to guarantee that students could not solve the items by recalling information from the course assignments.

In addition to ensuring that the answers to the items could not be found in the course readings and were not stated in the lectures, the author aimed the items at levels 3 thru 5—apply, analyze, and evaluate—of the revised Bloom's Taxonomy (Krathwohl, 2002). The items asked students to either: apply information from the course to a new situation, apply the definition of a

concept learned in class to a new situation, generate inferences based on scientific data, or evaluate sources of evidence. The author derived the data and concepts for many of the items from medical school textbooks and research papers from the PubMed database (National Center for Biotechnology Information & U.S. National Library of Medicine, 2011). These sources were chosen because most—if not all—second semester introductory biology students do not read medical school textbooks and do not search Pubmed for research papers that relate to this course. Additionally, searching Pubmed and reading medical textbooks allowed the author to write questions that relate to the career goals of most science majors. Each multiple-choice item contained five answer choices (one correct answer and four distracters).

In the end, the author wrote an initial pool of 48 test items for the professor of the second semester biology course to approve. The professor and author jointly reviewed all test items to verify that the items were scientifically correct, appropriate for the course, and that the items required critical thinking skills. The professor also provided suggestions for revisions and corrections. Forty-one of the initial pool of 48 items made it past the professor's review (Appendix C). The author also wrote a detailed explanation of the reasoning path to the correct answer for the 41 remaining items (Appendix D).

Validation Studies:

Cognitive think-aloud sessions (Norris, 1990; Taylor & Dionne, 2000) were conducted in the Spring of 2012 as one way to assess whether the remaining 41 items required critical thinking skills. Six students who declared a life science major and had already taken the course for which the test items were written for were recruited for these sessions. Because a subset of the items were to appear on the Spring 2012 exam, students currently taking the biology course could not participate in the cognitive think-aloud sessions. Each student met individually with the author.

Students were instructed to verbalize their thoughts as they worked through the pool of 41 items. Because the students were removed from the content material of the course, they were allowed to refer to the course textbook when working through the items and, when applicable, students were provided with the textbook page references for each item (Campbell, 2010). The textbook provided students with the background material that they had forgotten; however, the textbook did not provide students with answers to any of the items. In an attempt to counterbalance the items, three of the students started with item 1 and the other three students started with item 41. Students were told not to filter their words and to just talk out their thought process. The author remained as non-intrusive as possible during the sessions. The author only spoke when needed to remind students to verbalize their thoughts and to ask students to speak louder. While students were asked to spend 60 minutes working through the items, the sessions ranged from 46 minutes to 90 minutes in length. All cognitive think-aloud sessions were audio recorded and transcribed. To protect their identities, all students were given pseudonyms (Table 2.1).

Table 2.1: Characteristics of student participants. Descriptive information about the students who participated in the cognitive think aloud sessions and their participation is provided.

Pseudonym	Year in School	Major	Semester in which they took the Biology Course:	Post-graduation Plans	Started on Item Number	Items Completed:
Nick	3	Pre-med/Chemistry	Fall 2011	Medical School	1	All except 32
Carol	3	Microbiology	Spring 2011	PA School	41	10,11,15-41 (not 32)
Amanda	3	Biochemistry	Fall 2011	Pharmacy School	1	All
Sandra	3	Biology	Spring 2011	Optometry school	41	All
Victoria	4	Biology/Psychology	Fall 2009	Take a year off and apply for graduate school	41	15-41
Lucy	3	Biology	Fall 2011	Not Sure	1	1 thru 22

Faculty feedback served as another source of validity evidence. An expert panel of five faculty members from the college at which this course is taught at was recruited to review the items. The panel of faculty members represented the departments of Plant Biology, Genetics, Biochemistry and Molecular Biology as well as the schools of Veterinary Medicine and

Medicine. Faculty members were given a copy of the 41 items with textbook page references, an answer key to the items, the written explanation of the rationale and logic behind each question, a review guide, a copy of the course syllabus, and online access to the course textbook. The review guide requested that faculty evaluate all items for their scientific content and to use the revised Bloom's taxonomy to identify any items that were of lower-order thinking (Bloom's levels 1-2) (Krathwohl, 2002). Faculty members were given the option to either submit a written review of the items or to meet with the author of this paper. Qualitative comments about specific items from faculty reviewers were used to revise the items prior to the Spring 2012 exam. Faculty ratings of the items were compiled and items that received ratings of level 1, level 2, and lower-order were investigated to determine if the balance of evidence sided with the ratings of higher-order thinking or lower-order thinking.

Results:

Cognitive Think-Aloud Sessions:

To ascertain whether the student participants used critical thinking skills or not when solving the items, transcripts from the cognitive think-aloud sessions were analyzed based on the criteria of Norris (1990). According to Norris (1990) cognitive think alouds can be used to validate multiple-choice tests if they show that "good thinking" leads to correct answers and "bad thinking" leads to incorrect answers (Norris, 1990, p. 55). In accordance with Norris' (1990) criteria, the analytic questions asked of each item were: 1) Does correct reasoning lead to correct answers? and 2) Does incorrect reasoning lead to incorrect answers. In accordance with these questions, each student's response to an item was first coded as either "correct" or "incorrect" and then subcoded according to how they arrived at their answer. Correct answers were subcoded as resulting from either: correct reasoning, incorrect reasoning, recall, or a guess.

Incorrect answers were sububcoded as resulting from either: correct reasoning, incorrect reasoning, or a guess (Table 2.2). The coding categories were not finer (i.e. no attempt was made to specify the specific thinking skill used by each student) because doing so would likely lead to many erroneous and inconclusive interpretations. Furthermore, solving a test item by recall implies that at some point during the course the student encountered the answer to the question and the test item simply requires them to retrieve that answer. Because it is assumed that the material taught in the course is consistent with the current state of science, a student who relies on incorrect information to solve an item is most likely the result of the student's failure to learn or comprehend the material rather than the item being a true recall item. It is also possible that a student could recall an answer to an item from a source other than the biology course (i.e. previous coursework, laboratory internships, and other prior experiences). However, this is less likely to occur and is harder to control for.

For the purposes of this study, successful items are those in which students who used correct reasoning arrive at the correct answer while students who used incorrect reasoning arrive at an incorrect answer. It is therefore encouraging that none of the students used proper reasoning to arrive at an incorrect answer.

Item 27 (Figure 2.1) is an example of a successful item. Sandra and Carol used correct reasoning to solve item 27. As Sandra said: "I'm going to go with decreased ion flow between cardiac cells because that's how they do their impulses umm just from opening to the next cell…And something in the middle would get in the way." Carol arrived at the same conclusion as Sandra: "I know that ions are important in electrical conductivity umm. And that could lead to heart failure if the umm ions couldn't like spread the impulse fast enough." While Sandra and Carol did not use the all of the technical terms, their rationales were correct. The collagen

deposits that are characteristic of cardiac fibrosis block the transmissions of ions between the gap junctions of cardiac cells thereby interfering with the cardiac cycle. In contrast, Victoria's incorrect reasoning led her to select an incorrect answer: "I do think it that will also misalign the sarcomeres because of all those collagen deposits there." Unlike skeletal muscle cells, the sarcomeres of cardiac muscle cells are not aligned. Furthermore, the question referred to intercellular disruptions not intracellular disruptions. Therefore, it is not possible to "misalign" cardiac sarcomeres and choice c is incorrect. These results show that item 27 requires students to analyze the effects of pathological collagen deposits in the heart and that correct reasoning leads to the correct answer while incorrect reasoning leads to an incorrect answer.

27. Cardiac fibrosis is marked by large collagen deposits between cardiac cells and is commonly seen in patients with chronic heart failure. Packing collagen between cardiac cells can result in:
    a. **Decreased ion flow between cardiac cells.**
    b. Decreased heart size.
    c. Misalignment of cardiac sarcomeres.
    d. Mixing of oxygenated and deoxygenated blood.
    e. All of the above.

Figure 2.1: Preliminary pool item 27. The item is presented as it was given to students and faculty. The correct answer choice is indicated in bold type.

The data on item 20 were less favorable than the data on item 27. When presented with item 20 (Figure 2.2), Victoria used poor logic but answered the item correctly: "Okay it's not just going to be sodium it's not going to be just potassium. It has to be both of the pumps. That's why you drink Gatorade…I'm going to go with both of them. They're usually associated." Victoria displayed a common misconception among students that the concentrations of sodium and potassium ions are somehow linked or coordinated. While the Sodium/Potassium ATPase does transport both ions, neither the concentrations nor the activity of sodium and potassium are linked. The reason that upregulating the Sodium/Potassium Pump in skeletal muscle fibers is the

correct answer is because the ATPase is needed to transport the potassium ions from a region of lower concentration (interstitial space) to a region of higher concentration (skeletal muscle fibers). This item was removed from the pool of items because students can rely on a common misconception to correctly answer the item.

20. Excitation of muscle fibers during exercise results in action potentials. The K^+ that leaves the cell during the repolarization phase of the action potential either diffuses into the capillaries or is reclaimed by the skeletal muscle fibers. Which of the following is an adaptation to exercise that can prevent hyperkalemia (high levels of potassium in the blood) during prolonged periods of exercise:
    a. Increasing the number of K^+ leak channels in skeletal muscle fibers
    b. Increasing the number of Na^+ leak channels in skeletal muscle fibers
    c. **Increasing the number of Na^+/K^+ pumps in skeletal muscle fibers**
    d. Increasing the number of Ca^{2+} ions released per action potential
    e. Increasing the intestinal absorption of K^+.

Figure 2.2: Preliminary pool item 20. The item is presented as it was given to students and faculty. The correct answer choice is indicated in bold type.

Item 11 (Figure 2.3) was eliminated because students incorrectly reasoned their way to a correct answer or relied on recall. Instead of considering the distractors, students tended to look for the term "meiosis" and when they saw that meiosis was not one of the options they selected none of the above. As Nick said: "I would say none of the above because…I would think meiosis." Amanda relied on the same reasoning: "I think the answer should be meiosis which is not one of the choices given." That said, some of the students did go through the distractors to make sure their answer was correct. Lucy went through the distractors but did not show true critical thinking skills: "plasmogamy is the fusion of the cytoplasms…And Karyogamy is when the nuclei are being fused together. Fertilization they have already had them [gametes]. The have already been produced and they're coming together to create the organism so I feel like it's none of the above". Lucy's response showed that she merely needed to recall the definitions of the terms given to solve the item. Item 11 was therefore taken out of the item pool because it

reinforced the misconception that gametes are only produced through meiosis and could be solved by recall.

11. _____ produces gametes (Pages 611, 624, 639, 643, & 802-803)
    a. Cleavage
    b. Fertilization
    c. Karyogamy
    d. Plasmogamy
    e. **None of the above**

Figure 2.3: Preliminary pool item 11. The item is presented as it was given to students and faculty. The correct answer choice is indicated in bold type.

On the surface it appeared as though students could incorrectly reason their way to a correct answer on item 41 (Figure 2.4); however, a closer analysis of the data revealed otherwise. During the menstrual cycle, a surge in the level of luteinizing hormone (LH) triggers ovulation. Victoria incorrectly reasoned that: "failure to ovulate. 'cause it's going to have…there is going to be less estrogen in the body…And it needs the estrogen and insulin to ovulate the egg. The progesterone is the one that is responsible for thickening the layer of the uterus." While low levels of LH will also lead to reduced levels of estrogen, estrogen is not the hormone that triggers ovulation. Victoria was also incorrect about the hormonal control of endometrial development. During the proliferation phase (Days 5 – 14, on average) estrogen stimulates the formation of the endometrial layer. After ovulation, estrogen and progesterone (secreted by the corpus luteum) maintain the endometrial layer. Lastly, implicating insulin as a trigger for ovulation reflects another misconception. In women with polycystic ovary syndrome (PCOS), hyperinsulinemia and insulin resistance can lead to failure to ovulate. However, the failure to ovulate is not due to some inability of insulin to trigger ovulation. Rather, hyperinsulinemia in women with PCOS increases androgen production and high androgen levels interfere with ovulation (Nestler, 2000).

Given all the errors in her reasoning, Victoria's correct response to item 41 can be attributed more to luck than to a flaw in the item.

41. The anterior pituitary of a female with hypogonadism secretes abnormally low levels of LH. Insufficient levels of LH can lead to (Pages 1008-1009):
    a. Developlent of multiple follicles at a time
    b. **Failure to ovulate**
    c. Increased endometrial development
    d. Increased fertility
    e. Abnormally high levels of Inhibin

Figure 2.4: Preliminary pool item 41. The item is presented as it was given to students and faculty. The correct answer choice is indicated in bold type.

Item 17 was the only other item that a student was able to recall the correct answer. When presented with item 17 (Figure 2.5) Sandra relied on information she learned in an upper level physiology course to solve the item: "Lack of ATP in skeletal muscles would result in…net flow of calcium from the sar into the cytoplasm of skeletal muscle. I believe that actually does have something to do with it I remember from physiology not [the biology course]." Sandra is clear that she learned this information in her physiology course, not the biology course for which these questions were intended. Even though Sandra solved the item by recall, students in this biology class will not be able to solve the item by recall as they are not yet eligible for the upper level physiology course that Sandra took.

17. Rigor mortis (stiffness of death) is believed to result from the depletion of ATP in skeletal muscle cells. Lack of ATP in skeletal muscle cells after death would result in (Page 1107):
    a. **A net flow of Ca^{2+} from the sarcoplasmic reticulum into the cytoplasm of the skeletal muscle cell.**
    b. Tropomyosin blockage of myosin binding sites.
    c. The detachment of actin from myosin.
    d. A net flow of Na^{+} from the cytoplasm of the skeletal muscle cell to the extracellular space.
    e. An increase in the rate of glucose metabolism.

Figure 2.5: Preliminary pool item 17. The item is presented as it was given to students and faculty. The correct answer choice is indicated in bold type.

Even though none of the six student participants had ever participated in a cognitive think-aloud, they were surprisingly candid and open. Sandra interrupted her pause to say "I'm not thinking…I would be talking if I was." Sandra's frankness was also evident in her guess to item 1: "This is taking so much longer than I would expect…Okay. I'm going to go with b. Who knows?" In response to a question about plants Nick commented: "I don't like plants. I'm premed for a reason." Nick also had a few thoughts to offer about items that focus on female reproduction: "I don't have to deal with all that kind of stuff…it's biased to women to ask questions like this… it definitely needs to be corresponded to something about male too. Just so it would be balanced." Item 38 about female reproduction evoked a different response from Carol: "I remember that [biology course professor] tried so hard to get us to remember when you have the ability to get pregnant and somehow I managed to forget." Amanda was honest about not understanding the concept of resting potential: "I feel like this is actually testing whether or not you actually understand you know what the meaning of the science terminology. I feel like resting potential is one of those terms that gets thrown around…just a definition. But this is testing if you know what that is in a way…really understanding what it means. 'Cause I can say resting potential is this but not really understand what I'm saying." Victoria was also honest about having forgotten a lot of the course material: "Oh man! [Biology Course Professor] would not be proud of me." Realizing what they had forgotten was a common theme among students. When presented with a phylogenetic tree to interpret Lucy interjected: "What is this thing called?"

An analysis of the overall results of the six cognitive think-aloud sessions revealed several insights into the items (Table 2.2). With the exception of Sandra's performance on item

Table 2.2: Cognitive think-aloud results. Each student's name was entered into the box that matched the code for their response. Only complete answers were coded. No attempt was made at coding items that students skipped, failed to complete, or did not attempt. A column for incorrect answers obtained through incorrect reasoning was not included because no answer was coded as such.

Item	Correct Answer				Incorrect Answer	
	Correct Reasoning	Incorrect Reasoning	Recall	Guessed	Incorrect Reasoning	Guessed
1	Amanda				Nick, Lucy	Sandra
2	Amanda, Nick				Sandra, Lucy	Nick
3					Amanda, Lucy, Nick	Sandra
4	Nick, Lucy				Amanda, Sandra	
5				Sandra	Amanda, Nick, Lucy	
6					Amanda, Nick, Sandra, Lucy	
7	Amanda, Nick, Lucy				Sandra	
8	Nick				Amanda	Sandra
9	Amanda, Nick				Sandra, Lucy	
10	Amanda, Nick, Sandra				Carol	
11		Amanda, Nick	Lucy			Sandra
12	Nick, Sandra				Amanda, Lucy	
13	Amanda, Nick, Sandra, Lucy					
14	Amanda, Nick, Sandra, Lucy					
15	Amanda, Sandra				Nick, Carol, Victoria, Lucy	
16	Nick				Carol, Victoria, Lucy	Amanda, Sandra
17			Sandra		Nick, Amanda, Victoria, Lucy	
18	Nick				Amanda, Carol, Victoria, Sandra, Lucy	
19	Nick, Carol, Victoria, Sandra				Amanda	
20		Carol, Victoria		Sandra	Nick, Amanda, Lucy	
21	Nick, Sandra, Lucy				Amanda, Carol, Victoria	
22	Nick, Sandra				Carol, Victoria, Lucy	Amanda
23	Amanda, Carol, Victoria, Sandra				Nick	
24	Nick, Carol, Victoria				Sandra	Amanda
25					Amanda, Carol, Victoria	Sandra, Nick
26	Nick, Carol, Victoria, Sandra				Amanda	
27	Carol, Sandra				Nick, Amanda,Victoria	
28	Nick, Amanda, Carol, Victoria, Sandra					
29					Nick, Amanda, Carol, Victoria, Sandra	
30	Nick, Amanda, Carol, Victoria, Sandra					
31	Nick, Carol, Victoria, Sandra				Amanda	
32	Victoria, Sandra				Amanda	
33	Nick, Amanda, Carol. Victoria, Sandra					
34	Amanda, Carol, Victoria, Sandra				Nick	
35	Nick, Amanda, Carol, Sandra				Victoria	
36	Nick, Amanda, Sandra				Carol, Victoria	
37	Nick, Amanda, Victoria				Sandra	
38	Carol, Victoria, Sandra				Nick	Amanda
39	Nick, Amanda, Victoria				Carol, Sandra	
40	Nick, Amanda, Carol, Victoria, Sandra					
41	Nick, Carol	Victoria		Amanda	Sandra	

17 and Lucy's performance on item 11, none of the other 39 items could be solved by pure recall. Also, as mentioned earlier, none of the students correctly reasoned their way to an incorrect answer. Taken together, these results indicate that the bulk of the items did force students to go beyond simple recall and that when students employed sound reasoning, they solved the items correctly. Likewise, improper reasoning tended to result in incorrect answers. Items 11 and 20 were eliminated from the pool because students could rely on common misconceptions or recall to arrive at the correct answer. Item 17 was retained because she the student recalled the answer to the item from an upper level physiology course for which this biology course is a prerequisite. The results also show that guessing did not prove to be a good strategy on these items. The vast majority of guesses were incorrect and the prevalence of incorrect guesses suggests that students were largely unable to use test-savvy strategies to discern the correct answer from the distractors.

Faculty Review Data:

Five faculty reviewers were recruited to review the items to assess if any items fell into Level 1 or 2 of the Revised Bloom's Taxonomy (Krathwohl, 2002). Requesting faculty members to rate the items as either lower-order or higher-order thinking is consistent with methodology that has been used to evaluate multiple-choice test items for medical students (Simpson & Cohen, 1985). Faculty reviewers 1, 2, 3, & 4 reviewed all 41 items while faculty reviewer 5 reviewed items 1 through 15. 10 of the 41 items received ratings of "level 1", "level 2", or "lower-order" (Table 2.3). 7 of the 10 items were rated as lower-order by a single faculty reviewer while the remaining three items were rated as lower-order by two faculty reviewers. Three reviewers also commented that item 14 was too easy. Lastly, reviewer #1 suggested the removal of item 21 because the item was true for ligand-gated ion channels (a topic that was

addressed in the biology course) but was not true of metabotropic receptors (a topic not addressed in the biology course). Even though the students who took this course are likely to be unaware of the differences between ligand-gated ion channels and metabotropic receptors, it was important for all items to be scientifically correct.

Table 2.3: Preliminary pool items classified by faculty as easy or lower-order thinking.

Item	Faculty Reviewer				
	1	2	3	4	5
4	level 1	level 2			
6	level 1				
9	level 2				
11		lower order	lower order		
12					lower order
14			too easy	too easy	too easy
20		level 2			
21	level 1				
22	level 1				
25		level 2			
26	level 2	level 2			

Faculty reviewers also provided some comments on the items in general. One faculty reviewer expressed the opinion that while lower-order thinking items are often looked down upon, comprehension items (Bloom's level 2) can be very useful for identifying areas of student weakness. Faculty reviewers also commented on the use of distractors. Reviewers 1 and 5 observed that choice e tended to be the "dumping ground" for bad distractors. Reviewer 1 even suggested reducing the number of distractors for each item to 3—a suggestion that is congruent with the literature that demonstrates that three distractors are sufficient (Haladyna, Downing, & Rodriguez, 2002). These comments from faculty members are helpful insights into writing multiple-choice test items.

Discussion:

Comparing Student and Faculty Responses:

The student transcripts were compared to the faculty reviews to resolve the discrepancies between faculty and student responses to items 4, 6, 9, 12, 22, 25, & 26. Even though none of the items were rated as lower-order by more than two of the five reviewers, it is nevertheless important to try to investigate the apparent contradictions. Items 11, 14, 20, & 21 were not analyzed because they were eliminated from the item pool.

While one faculty member rated item 22 (Figure 2.6) as level 1, few students were able to recall their way to a correct answer. Transcripts from the students who attempted this item show that students revealed some difficulties with the dynamics of actin and myosin dynamics. When solving item 22 Amanda stated: "Well if they are contracted…if it is in a contracted state there is some kind of linkage happening. Because then once it releases, something there had to be energy for it to be released…I'm going to go with the one with ATP." Amanda seemed to have been confused on the dynamics of actin, myosin, and ATP. Amanda's answer implied that, when actin and myosin are crosslinked, one of the two proteins is bound to ATP. Her answer also implies that the protein bound to ATP uses the energy from ATP hydrolysis to break the crosslink. However, this is not entirely correct. The myosin head is bound to ADP and an inorganic phosphate when it forms a crossbridge with actin. The myosin head releases the ADP and inorganic phosphate during the power stroke and the actin-myosin crossbridge is broken when the myosin head binds to another molecule of ATP. Amanda correctly remembered that reversing the actin-myosin crossbridge requires energy in the form of ATP; however, she incorrectly guessed that either actin or myosin needed to be bound to ATP when the two proteins are crosslinked. Lucy faltered on item 22 because she could not jump to a specific point in the

sliding filament theory of actin and myosin binding: "when a muscle contracts the sarcomeres shorten and like it starts with myosin being bound to ATP and hydrolyzing that to ADP and the cycle goes like that… so the answer for that would be that myosin is bound to ATP which is how that all starts." The textbook figure that models the actin-myosin interactions depicts ATP hydrolysis by the myosin head as step 1 of the process. However, item 22 asked students to consider step 3 of the textbook figure which shows the actin-myosin crossbridge. Rather than think through the steps of the sliding filament theory, Lucy fixated on step 1. The faculty rating of this item as recall and the students' difficulty with this item may reflect differences between experts and novices. For students who are struggling to grasp the sliding filament theory, the process is much more than a list of steps to memorize and recall. However, faculty experts can easily recall the proper information to solve this item.

22. Which of the following is true when cardiac sarcomeres are in a contracted state (Pages 903 & 1104):
    a. Actin and myosin are not crosslinked.
    b. Myosin is bound to ATP.
    c. Ca^{2+} is bound to troponin.
    d. Actin is bound to ADP.
    e. This region is in diastole.

Figure 2.6: Preliminary pool item 22. The item is presented as it was given to students and faculty. The correct answer choice is indicated in bold type.

The data on item 6 (Figure 2.7) may also reflect a difference between experts and novices. Even though the item stem did not provide any sort of context for the item, all five faculty reviewers placed this item in the correct context of protobionts and the origins of life which is the topic that this item was written to assess. One of the five reviewers even rated this item as lower-order thinking. The students however, were unable to identify the context of the item. Instead of placing the item in the context of protobionts and the origins of life, the students tended to place this item in the context of a eukaryotic cell. When solving this item Nick

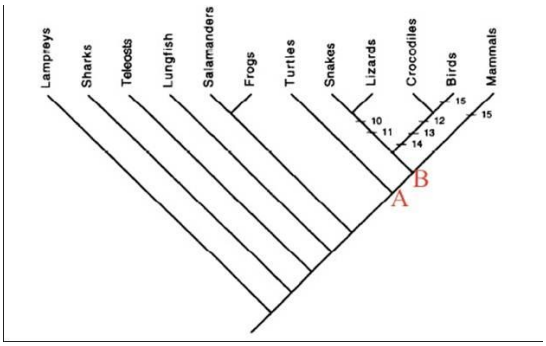
reasoned that: “RNA you know has to come out of the nuclear membrane before it can do stuff.” Amanda made the same contextual error as Nick: “RNA enclosed in a membrane…would have the problem of getting enclosed in the nucleus.” Since students were unable to properly interpret this question, the data on the level of thinking required to solve this item are inconclusive. Providing students with more context in the item stem may help determine if students need to think through this item or if they can just recall the answer.

6. Which of the entities listed below has the greatest chance of being able to carry out both enzyme activity and replication (Pages 509-510):
    a. A protein enclosed in a membrane.
    b. A strand of RNA not enclosed in a membrane.
    c. A strand of DNA enclosed in a membrane.
    d. **A strand of RNA enclosed in a membrane.**
    e. A protein not enclosed in a membrane.

Figure 2.7: Preliminary pool item 6. The item is presented as it was given to students and faculty. The correct answer choice is indicated in bold type.

The students’ responses to item 9 (Figure 2.8) were at odds with the faculty reviewer who rated this item as Bloom’s level 2. In order to solve this item correctly, students needed to read the phylogenetic tree to infer the meaning of an internal node. Since the textbook does not explain the concept of an internal node, students must rely on the tree to derive the meaning of the internal node at point A. Additionally, the format of the phylogenetic tree shown in item 9 differs from the format of the phylogenetic trees shown in the textbook. Therefore, students needed to apply their knowledge from the textbook to interpret the figure. Amanda’s answer showed that she was able to correctly interpret the figure to determine the evolutionary relationships between the organisms and infer what occurred at point A: “frog and salamander diversion happened before…point a. So therefore it is not…one with frogs and salamanders. And then mammals and snakes both happened after turtles…so did lizards and crocodiles.” While her words are slim, Amanda’s answer choice showed that she correctly applied her

knowledge of the time axis in phylogenetic trees to infer that the node at point A represented where the ancestor to turtles diverged from the common ancestor to the mammals, birds, snakes, lizards, and crocodiles. Sandra's answer showed that she was unable to properly apply her knowledge of phylogenetic trees to infer the meaning of the node at point A: "The point at which the ancestor to turtles diverged…from the ancestor to mammals. Okay that makes more sense because that they're the two that diverged from that ancestor." Sandra selected the incorrect answer because she simply looked for a bifurcation at point A between two taxa. Sandra was unable to read the figure to infer that the internal node does not necessarily represent a split between two taxa. An internal node represents a split between evolutionary lineages. Item 9 was therefore retained as a critical thinking item.



Please use the above figure for the questions 8 and 9. Figure modified from: (Morrison, 1996)

9. What does the node at point A in the above figure represent (Page 538)?
    a. The point where the turtle ancestor diverged from the ancestor to snakes, lizards, crocodiles, birds and mammals.
    b. The point where the turtle ancestor diverged from the ancestor to snakes.
    c. The point where the turtle ancestor diverged from the ancestor to mammals.
    d. The point where the turtle ancestor diverged from the ancestor to frogs and salamanders.
    e. The point where the turtle ancestor diverged from the ancestor to frogs.

Figure 2.8: Preliminary pool item 9. The item is presented as they were given to students and faculty. The correct answer choice is indicated in bold type.

Items 25 and 12 (Figure 2.9**)** were two items for which there is weak evidence that these items require critical thinking skills. One faculty reviewer rated item 25 as Bloom's Level 2 and another faculty member rated item 12 as lower-order. More than anything, student responses to these items revealed gaps and misconceptions in their knowledge of the topics on which these items tested them. Amanda's response to item 25 clearly showed her misconception about how muscles grow: "I really think you can't change the size of muscle fibers themselves because the fibers are so dense". Granted muscle fibers are dense; however, muscle hypertrophy results from increases in the amount of actin and myosin filaments within the muscle fiber. Carol's answer revealed multiple gaps in her knowledge of this topic: "I'm not really sure what motor units are but I think it [the answer] probably has more to do with the myofibrils and, and the muscle fiber than it has to do with the motor units… if your muscles are growing they are going to have to have a faster rate of cell division." Carol wrongly inferred that a motor unit has nothing to do with the muscle fibers. A motor unit is comprised of one motor neuron and all of the muscle fibers it stimulates. Carol also wrongly assumed that muscle growth occurs through cell division. Victoria echoed Carol's misconception that muscles grow through increases in cell division: "The rate of cell division in muscle fibers? Well if it is building them, then is it going to be it would be the rate would go up as opposed to just a resting rate. Because now you're actually doing something with them." Item 12 tested students on whether they understand what the process of germination and the structure of a seed. Amanda's response showed her misconceptions with this topic: "Well germination is dealing with reproduction…meiosis is how gamete cells are produced. So I'm going to go with meiosis." Amanda failed to recognize that the seed contains the plant embryo and resorted to equating germination with reproduction and meiosis. Amanda's answer also reflects her aforementioned misconception that gametes are only

produced by meiosis. Nick's response suggests that he was on the right track but the vagueness of his answer is likely due to underlying gaps in his knowledge of seeds: "so the seeds are already there so I would say mitosis just in general cellular division. Yeah 'cause I feel like the seeds are already made." It is unclear what Nick means by his phrase "I feel like the seeds are already made." If Nick intended to say that seeds are produced after fertilization, then he would be correct to eliminate three of the four distractors. However, he was unable to clearly articulate the concepts. Thus, even if items 12 and 25 required critical thinking skills, the students did not remember the proper information to do so. However, item 12 received four ratings of higher-order thinking and item 25 received three rating of higher-order thinking. Therefore, there is weak evidence that that these items require higher-order thinking.

12. A herbicide that kills germinating seeds most likely blocks the process(es) of (Page 624):
    a. **Mitosis**
    b. Fertilization
    c. Meiosis
    d. Gametogenesis
    e. Peptidoglycan formation

25. Strength training can cause all of the following to increase EXCEPT:
    a. The number of recruited motor units.
    b. The number of myofibrils in a muscle fiber.
    c. The rate of cell division in muscle fibers.
    d. The amount of actin and myosin in a muscle fiber.
    e. The size of muscle fibers.

Figure 2.9: Preliminary pool items 12 and 25. The items are presented as they were given to students and faculty. The correct answer choices are indicated in bold type.

The student data on items 4 (Figure 2.10) and 26 (Figure 2.11) sided with the faculty members who rated these items as Bloom's Level 1 and/or Bloom's Level 2. The student responses to item 4 showed that first recalling the effect of cholesterol on membrane fluidity and then making a simple inference based on that effect could solve the item. However, students

tended to over-think this item. Lucy's response is an example of the simple inference students were asked to make: "[cholesterol] reduces the fluidity of the membrane…But one without it [cholesterol] I feel like the fluidity wouldn't be inhibited so…umm. Yeah so I feel like membrane a would transition to a liquid but the b with the 50% cholesterol would remain a gel simply because it has cholesterol in it." Nick made the same inference as Lucy but went on to over-think the question:

"cholesterol would produce like a more stable structure or whatever. So I would think that if you increase the temperature that something without cholesterol it would go into a liquid form more readily than something with less cholesterol… I don't know to what degree you know if this temperature is raised to what degree would it…breakdown or whatever."

In the end Nick stuck to his original inference and answered the item correctly. The student transcripts also concurred with the two faculty reviewers who rated item 26 as level 2. Victoria's response to item 26 showed that it only asked students to comprehend the definition of a heartbeat: "[the SA node] it's going to depolarize one time per beat. 'Cause that's how you get the heartbeat…So if it depolarizes once per beat it's going to be 84." These items are therefore best classified as Bloom's level 2 and do not elicit critical thinking skills as defined by Bloom's taxonomy.

4. Consider two cell membranes. Membrane A does not contain cholesterol (0%) while Membrane B contains 50% cholesterol. At a temperature of 20°C both membranes are in a gel state. Predict what will happen if the temperature is raised from 20°C to 36°C (Page 128):

a. **Membrane A will transition to a liquid while membrane B will remain a gel.**
b. Both membranes will transition to a liquid state.
c. Membrane B will transition to a liquid while membrane A will remain a gel.
d. Both membranes will retain their gel state.
e. The proteins in membranes A and B will denature.

Figure 2.10: Preliminary pool item 4. The item is presented as it was given to students and faculty. The correct answer choice is indicated in bold type.

26. If a person's pulse is 84 beats per minute, how many times per minute does their SA node depolarize (Page 904):

a. 42
b. 21
c. 168
d. 84
e. 336

Figure 2.11: Preliminary pool item 26. The item is presented as it was given to students and faculty. The correct answer choice is indicated in bold type.

There are several limitations associated with the results. It is acknowledged that panels of 5 faculty members and 6 students are small in size. It is possible that if these items had been submitted to larger panels the results would have been otherwise. It is also acknowledged that none of the items were tested in their "native context" of the course for which they were written. The students who participated in the cognitive think-aloud sessions had already moved on from the biology course these items were written. As part of the process of moving on, the students had forgotten some of what they learned in the biology course and acquired information in later coursework that helped them solve the items. The data from the cognitive think-aloud sessions are also limited to the author's interpretations of the student's responses. Even though students were asked and reminded to say everything and anything they were thinking, it is unlikely that the students were fully able to express themselves. It is also possible that the author's interpretations of the students' responses are not necessarily what the students meant to convey. Lastly, the variation among the faculty reviewers went unresolved. It is possible that conducting a focus group with all five faculty members would have settled the discrepancies among the reviewers' opinions. However, validation is not about absolutes. So while the evidence presented has its limitations, it nevertheless supports the claims made in this analysis.

Conclusions:

This study began with the goal of writing a set of multiple-choice items that require critical thinking skills for an introductory biology course for science majors and ended with 32 items that showed strong evidence of requiring critical thinking skills. A total of seven items were eliminated from the set of multiple-choice items. Items 4 and 26, were eliminated because they were deemed to be Bloom's Level 2 and therefore they do not elicit the critical thinking skills of apply, analyze and evaluate. Items 11, 14, 20, & 21 were eliminated from the item pool because they were either solved by recall, too easy, too myopic, or they failed to weed out students with common misconceptions. Item 6 was eliminated because the data on these items were inconclusive. Items 12 & 25 were retained even though there was weak evidence in favor of them. The remaining 32 items were the items that were concluded to have strong evidence that they elicit critical thinking skills in undergraduate science majors. Thus, the final pool of validated items that require critical thinking skills (Bloom's Level 3- 5) contained 34 items.

Despite its limitations, this study demonstrated a method of write multiple-choice test items that require critical thinking skills and for investigating the cognitive validity of those items. The data from the cognitive think-aloud sessions and faculty reviews showed that multiple-choice items that were tailored to elicit critical thinking in a given biology course, actually could elicit critical thinking skills The data from the cognitive think-aloud sessions and faculty reviews showed that, when tailored to the biology course, multiple-choice test items can elicit critical thinking skills. This is encouraging for professors of large undergraduate biology courses who do not have teaching support to grade free response test items.

CHAPTER 3

Stepping It Up: Insights From A Rasch Model Analysis on Multiple-choice, Critical Thinking Items and Sources of Item Difficulty in Undergraduate Biology.

Abstract:

Final exam data from a second semester introductory biology course for science majors were analyzed using the Rasch Model. Twenty-three of the 113 items on the exam were taken from an earlier study that developed and validated a set of multiple-choice items that require critical thinking skills. This study aimed to examine the performance of the items on the test with particular emphasis on the 23 validated critical thinking items. Fit statistics on the items and students, item point to measure correlations, and invariance analyses provided evidence that the Rasch Model was appropriate for the dataset. The high standard errors associated with ability level estimates for high achieving students showed that the test did not contain enough items of high difficulty to precisely measure these students. The results of this study illustrated some of the potential benefits of an Item Response Theory approach to analyzing undergraduate level science exams and yielded insights into how to write multiple-choice test items for undergraduate biology courses that require critical thinking skills as well as factors that influence item difficulty.

Introduction:

Science educators at institutions of higher education are growing increasingly frustrated with the reliance on multiple-choice exams. This frustration is justified in light of recent data that multiple-choice science exams administered to undergraduates emphasized lower level thinking skills (Momsen et al., 2010). Factual recall multiple-choice exams have been blamed

for producing college graduates who cannot demonstrate that they understand the information they memorized during their undergraduate years (Lord & Baviskar, 2007) and for inadvertently teaching students that science is nothing more than an assortment of facts to commit to memory (Wood, 2009). In spite of the frustrations over multiple-choice testing in undergraduate science classrooms, abandoning this method of testing is not entirely feasible.

Large, 200 – 300 student, lecture classes that cover a vast array of topics is the common format for most introductory, undergraduate science courses (Tomanek & Montplaisir, 2004) and multiple-choice exams are most apt for this type of class (Martinez, 1999). Multiple-choice tests are an efficient and inexpensive method of testing. The score reliability of multiple-choice tests tend to be higher than those of essay tests. Whereas an essay exam can only test students on a limited set of concepts, multiple-choice exams can test students on the wide range of topics taught in many introductory, undergraduate science classes. The advantages of multiple-choice items are not limited to pragmatics. Multiple-choice items can elicit critical thinking skills in students (Martinez, 1999). The new Medical College Admissions Test (MCAT) is cited as an example of a multiple-choice exam that requires critical thinking (Zheng et al., 2008). Given the ease and versatility of multiple-choice exams, the goal should not be to abolish multiple-choice exams. Instead the goal should be to improve the quality of multiple-choice exams (Simpson & Cohen, 1985).

Despite the prevalence of multiple-choice tests and the need to improve the quality of multiple-choice tests, the science of item writing is very rudimentary (Haladyna et al., 2002). Literature reviews have assembled suggestions for item writing such as to make all distractors plausible and to vary the position of the correct answer. However, few of these suggestions have been validated through empirical studies (Haladyna & Downing, 1989a, 1989b). Furthermore,

these item writing guides are not specific to the needs of science educators who aim to write critical thinking multiple-choice items.

The "Biology in Bloom Tool" (Crowe et al., 2008) is arguably the most comprehensive and detailed guide for college science educators on assessing critical thinking skills. The Biology in Bloom Tool provides science specific examples of the thinking skills in Bloom's Taxomomy as well as the types of exam questions (labeling, fill in the blank, true or false, multiple-choice, short answer, and essay) that can assess each skill (Crowe et al., 2008). As helpful as the Biology in Bloom Tool is, it does not provide detailed insight into writing multiple-choice items or factors that influence the difficulty levels of the items. More research on the performance of critical thinking items in undergraduate level science exams is needed.

Item Response Theory (IRT) is a powerful research tool for college science educators who aim to investigate the performance of dichotomously scored, multiple-choice exams. Item Response Theory is a family of models that relate an individual's latent ability level to characteristic(s) of a dichotomously scored item which measures said latent ability to predict the probability of a correct response to the item (De Ayala, 2010; Embretson & Reise, 2000). Conversely, IRT models allow researchers to use the patterns of students responses to test items to make inferences about students' latent capabilities (Molenaar, 1995). Thus, by applying an IRT model to student responses to multiple-choice items that measure critical thinking skills, science educators can make inferences regarding their students' critical thinking abilities.

The sample independent (invariant) nature of parameters from IRT models is another key advantage of analyzing multiple-choice test data though an IRT lens. Being invariant, values for item difficulty are not tied to the sample of students who took those items nor are the ability level estimates of the students linked to the specific set of items they were given. An implication of

this principle is that, as long as the items are calibrated, different items can be used to measure the same trait (Embretson & Reise, 2000). Science educators can compare the performances of groups of students who did not necessarily take the same exam.

The Rasch model is the simplest of the IRT family of models as it only takes into account one item characteristic: item difficulty (Embretson & Reise, 2000). A person's latent ability level is measured along the same logit scale as item difficulty and, under the Rasch Model, the distance between the item's difficulty level (b_i) and the person's latent ability level (θ) governs the probability that a person will succeed on that item (Wright, 1977). An item's difficulty level represents the amount of ability needed to have a 50% chance of responding to the item correctly. When a person's ability exceeds the item's difficulty level, he or she is more likely than not to respond correctly to the item. When a person's ability is less than the item's difficulty level, it is more probable that he or she will not solve the item correctly. The logit scale on which item difficulty and person ability are measured is an interval scale; therefore, changes in student performance can be tracked over time (Embretson & Reise, 2000).

The purpose of this study was to use the Rasch Model to investigate the performance of a set of 113 multiple-choice items on a second-semester biology final exam for science majors at a large public research university in the Southeastern United States. Particular emphasis was given to the 23 critical thinking items were written by the author of this paper. The cognitive validity of the 23 items was established through cognitive think-aloud sessions with students and faculty reviews. The Rasch model analysis yielded insights into students' misconceptions, sources of item difficulty, and directions for future research.

Description of the Data:

Data from 113 multiple-choice questions on the Spring 2012 Final Exam of a second semester biology course for science majors at a large public research institution were analyzed. 86 of the 113 multiple-choice questions were provided by the professor. The 86 items provided by the professor contained both critical thinking and recall items. The remaining 27 questions were written by the author of this paper. Twenty-three of the 27 questions written by the author were part of the effort to validate a set of multiple-choice items that require critical thinking skills. The 113 multiple-choice items on this exam covered the topics of: evolution and natural selection, osmosis and diffusion, phylogenetics, plant biology, animal physiology, and ecology. Data from the 358 students who took the exam were obtained as an anonymized Microsoft Excel spreadsheet. All 113 items were administered to all 358 students; therefore, the dataset is complete.

Methods:

Winsteps (Version 3.74.0) (Linacre, 2012a) was used for the IRT and reliability analyses. Winsteps is a Rasch only software that relies on Joint Maximum Likelihood Estimation (JMLE). By default, Winsteps centers the item parameters at a mean of 0 logits and a standard deviation of 1 logit. Since all items were administered to all students, missing data were treated as incorrect responses. SPSS was used for all correlation analyses (IBM, 2010).

Results:

Model Fit Indicators:

Winsteps generates infit and outfit statistics on items and students to evaluate the extent to which the data match the model's predictions. The infit statistic is a weighted statistic that is heavily influenced by unexpected responses by students on items that are targeted to their ability

level. The outfit statistic is an unweighted statistic that is most influenced by students' unexpected responses on items that are not matched to their ability level. Being a probabilistic model, the Rasch model assumes that a certain amount of randomness is present in the data. A student with a 70% probability of responding correctly to an item still has a 30% probability of responding incorrectly to the item. This is where the randomness in the data arises. Mean square values of 1.0 for infit and outfit statistics indicate that the amount of randomness in the data matches the amount of randomness predicted by the model. Mean square values less than 1 indicate that the data fits the model too well while mean square values greater than 1 indicate that the data are too random (Linacre, 2012b). As they apply to students, fit statistics assess the degree to which a student's pattern responses to the test items matches the model's predictions (Bell, 1982). Similarly, fit statistics assess the congruence between the empirical data on an item and the model's predictions (Jackson, Draugalis, Slack, Zachry, & D'Agostino, 2002).

Fit statistics on items can be used to verify whether the assumption of unidimensionality independence holds up (Jackson et al., 2002; Linacre, 2009). The assumption of unidimensionality maintains that the items on the assessment measure a single latent trait and the trait measured by the items is the dimension (Embretson, 2000). However, true unidimensionality is a theoretical construct that is not perfectly realized in an actual dataset. Each item on the test contains multiple dimensions. Creating a unidimensional test means that when the items are grouped together, the common dimension present among all items is stronger than all other dimensions present in the items. The common dimension should correspond to the trait that the test designer intends to measure (Linacre, 2009). Fit statistcs can be used to assess the degree to which an item falls in line with the dimension measured by the rest of the items.

Fit statistics on an item greater than 2.0 suggest that the item measures something different than the rest of the items (Jackson et al., 2002).

Item Level Data:

Fit statistics and point to measure correlations on the items were analyzed to determine if any item or items disrupted the measurement system and to seek evidence on whether the dataset fit the Rasch Model. All items had positive point to measure correlations, which indicates that all items worked toward the Rasch Dimension. Infit mean square values for the items ranged from 0.87 to 1.16 and the outfit mean square values for the items ranged from 0.74 to 1.43 (Appendix E). Since mean square values for infit and outfit statistics between 0.5 and 1.5 are considered optimal for measurement (Linacre, 2012b), it was concluded that all items had acceptable fit statistics and that data satisfied the assumption of unidimensionality. Therefore, all items were retained in the analysis. Furthermore, an item reliability of 0.98 indicates that the sample size was large enough to measure the items.

Having concluded that all items were appropriate for the analysis the item parameters were then investigated. Winsteps anchored the mean item difficulty (*b*) at 0 logits (SE = 0.1) and scaled the item difficulties to a standard deviation of 1 logit. Difficulty values for the 113 items ranged from -2.58 logits to 3.89 logits (Appendix E). The item difficulty levels for the 23 critical thinking items written by the author (items 86 – 109, sans 97) ranged from $b = -1.75$ to $b = 1.67$ with an average of $b = 0.24$. Items at the extreme ends of the difficulty spectrum had the highest standard errors and item standard errors decreased as item difficulties approached the average value of 0 logits (Figure 3.1).

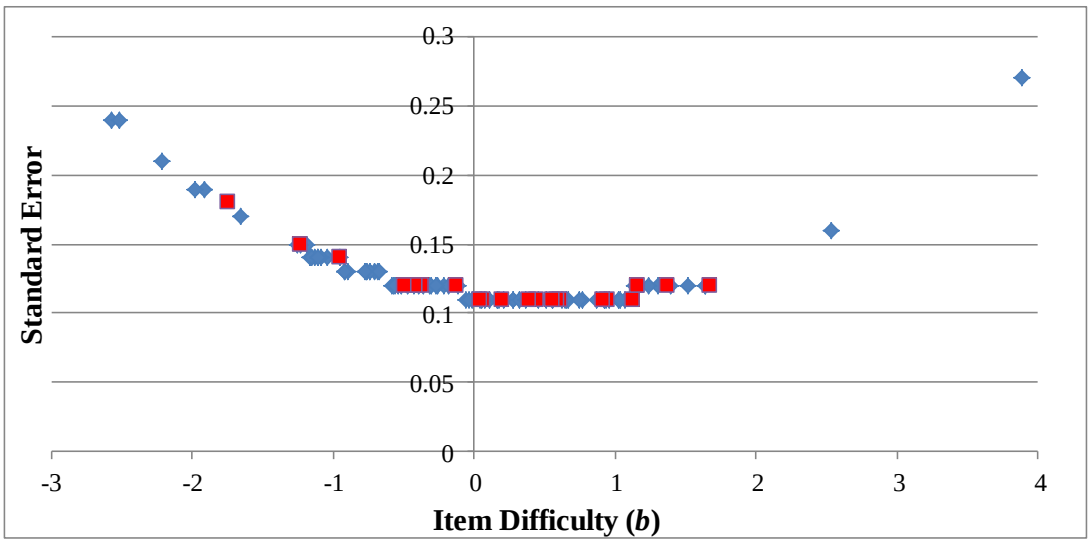


Figure 3.1: Standard Error versus Item Difficulty. Item standard errors increased as item difficulties diverged from the average difficulty of 0 logits. Red boxes represent the 23 validated critical thinking items written by the author while the blue diamonds represent the remaining 90 non-validated items.

In accordance with this assumption, Winsteps set the average slope of the items to 1. The extent to which an item's empirically determined discrimination estimate deviates from 1 reflects the extent to which item deviates from the Rasch model (Linacre, 2012b). Discrimination estimates for the 113 items ranged from 0.319 to 1.671. 29 of the 113 items had discrimination estimates that were greater than 1.1, which indicates that these items over-discriminated. Conversely, 31 of the 113 items had discrimination estimates that were less than 0.9, which indicates that these items were not discriminating enough. The Rasch Model also assumes that all items have a lower asymptote of 0 and an upper asymptote of 1. Again, Winsteps produces empirical estimates of these values to test the extent to which each item conformed to these assumptions (Linacre, 2012b). 24 items had lower asymptotes greater than 0.1 while 7 items had upper asymptotes less than 0.9. These deviations from the Rasch Model were localized to 60 items. Item estimates for 18 of the 113 items showed two or more deviations from the Rasch Model while discrimination estimates alone for 42 items deviated from the Rasch Model. When

considered alongside the infit and outfit mean square values, the item estimates suggest that while the Rasch Model can be applied to the dataset, a 2-PLM may be more appropriate. This analysis persisted with the Rasch Model because the results did not necessarily refute the Rasch Model and because the Rasch Model it is most appropriate for a sample size of 358 students. IRT models require a large sample and as the number of parameters in the IRT model increases, so does the required sample size (Reeve & Fayers, 2005). The Rasch Model can be applied to datasets with as few as 50 -100 people (Linacre, 1994). A study using the marginal maximum likelihood estimation method to apply a 2-PLM to simulated data found a greater amount of bias in the parameter estimates in smaller samples (N = 250) than larger samples (N = 750) (Lim & Drasgow, 1990). Even though it is reasonable to assume that students do guess on multiple-choice exams, applying 3-PLM to a dataset to account for guessing and differing item discrimination values requires a larger sample size than the 2-PLM. A common recommendation is that the 3-PLM should not be applied to datasets with less than 1,000 people (De Ayala, 2009). Furthermore, guessing parameters (c) are often poorly estimated (Baker, 2004).

Student Level Data:

The ability level values (θ) for the 358 students who took the exam ranged from -1.18 to 2.3 logits. The average θ value of 0.55 logits (SE = 0.02) was greater than the average item difficulty value of 0 logits. Infit mean square values for the θ values ranged from 0.83 to 1.28 and therefore fell within acceptable limits. Outfit mean square values for the θ values ranged from 0.66 to 1.88. Even though infit and outfit mean square values between 0.5 and 1.5 are optimal for measurement, mean square values between 1.5 and 2.0 do not degrade the measurement system. Only 4 students had outfit mean square values that were greater than 1.5

and all 4 of these students had acceptable infit mean square values. Therefore, all students were retained in the analysis.

Invariance Analyses:

Model parameters in Item Response Theory are sample-independent. That is to say, ability level estimates for test-takers are not tied to the sample of items on the test. Likewise, item parameters do not depend on the sample of students who took the exam (Embretson & Reise, 2000). Therefore, the extent to which the item and person parameters remain invariant is a measure of the extent to which the test data fit the IRT model (Hambleton et al., 1991).

The odd-even method was used to test the invariance of the theta estimates (Hambleton et al., 1991). Theta estimates for all 358 students were first obtained using only the items on the test with odd numbers. Theta estimates for all 358 students were then obtained using only the items on the test with even numbers. The correlation between the two sets of theta estimates was 0.822 ($p < 0.01$). A scatterplot of the two sets of estimates shows that theta estimates were most invariant for students of below average ability level and were least invariant for high achieving students (Figure 3.2 a**)**. Rather than serving as direct evidence of invariance, these results show that the test was not well matched to the high ability students.

As expected, the item difficulty (*b*) values showed a much greater degree of invariance than the student θ values (Figure 3.2b). Generally speaking, students tend to be less predictable than items. Also, there are 358 pieces of information on each item but there are only 113 pieces of information on each student. The greater amount of data on items increases the precision of the *b* values, which leads to a higher degree of invariance.

To test the invariance of the items, difficulty values for all 113 items were estimated using two different, random samples of 179 students. When estimating the item difficulty values

for the invariance analysis, the mean student ability level for each sample was anchored at 0 logits. Anchoring the mean student ability level at 0 logits set a common metric for comparing the item difficulty estimates. A correlation of 0.966 ($p < 0.01$) was obtained between the two sets of item difficulty estimates (Figure 3.2b). This result shows that the items maintained a high degree of invariance and does serve as direct evidence that the data fit the Rasch model.

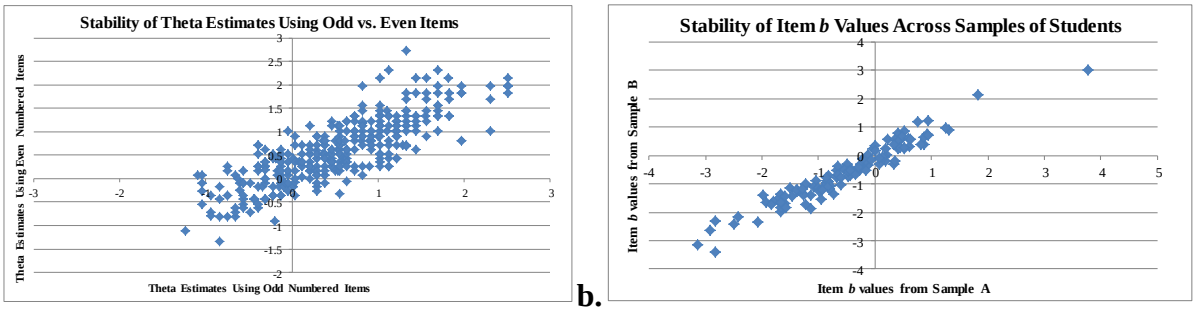


Figure 3.2: Tests of model parameter invariance. Data were analyzed to assess the extent to which the model parameters remained invariant. **A.** Comparison of student ability level estimates using the odd numbered items and the even numbered items. **B.** Comparison of item difficulty values using two random samples of students.

Test Level Data:

Data on the entire test was used to assess the extent to which the test matched the sample of students to whom it was administered. Cronbach's alpha for the exam was 0.90. This value is evidence that, even though the test could not precisely measure the ability levels of high ability students, the test could reliably distinguish between the high and low ability students. The person reliability value of 0.89 suggests that the test resolved the students into three groups based on ability level. The student to item bar chart (Figure 3.3) showed that bulk of the test questions were targeted to average and below average students. With only 5 items with difficulty (*b*) values greater than 1.5, the test was not able to precisely measure the ability levels of students at the upper end of the ability level range. The Student to Item Bar Chart also shows that the test contained an unnecessary number of low difficulty questions. In the ideal case, the average

difficulty of the items (M) would match the average ability level of the students and the item difficulty levels would be spread evenly over the range of M ± 2 standard deviations (Wright, 1977).

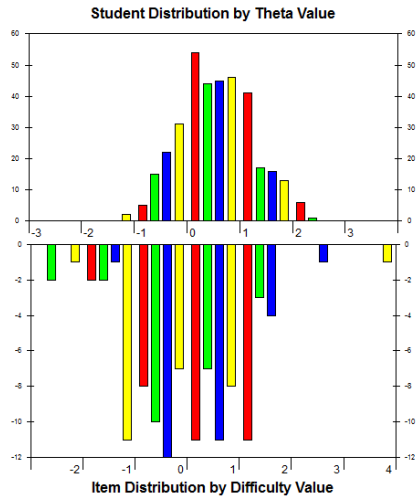


Figure 3.3: Student to Item Bar Chart. The central x-axis represents the logit scale for measuring item difficulty and student ability level and the y-axis represents the number of items/students. The less than symmetrical nature of this chart indicates that the test items were not perfectly matched to the students.

An analysis of the standard errors of measurement for the θ values echoed the data in the Student to Item Bar Chart. The θ estimates were most precise—as defined by lowest standard error of measurement—for average and below average students ($\theta < 1$). Ability level estimates were least precise for students of higher ability level ($\theta \geq 1$) (Figure 3.4). The student to item bar chart shows that most of the questions on the test were between $-1 \leq b \leq 1$. The sharp drop in the number of questions with difficulty $b > 1$ correlates with the sharp increase in standard error of measurement for ability level estimates $\theta > 1$. While the test was well matched to students of below average and average ability level, the test did not contain enough difficult questions to accurately measure the ability levels of high achievers.

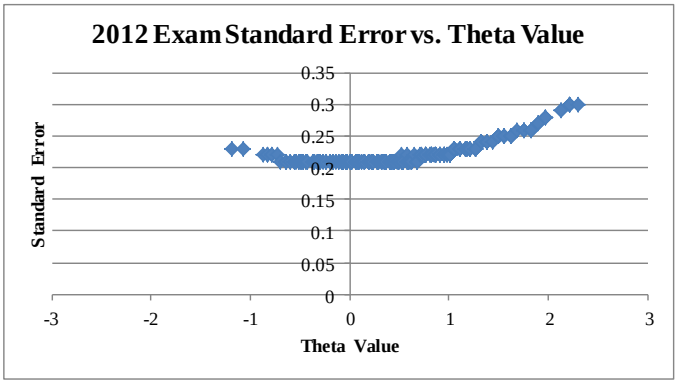


Figure 3.4: Standard Error versus theta value. The average theta value for the 358 students who took the exam was 0.55 logits. Theta values were most precise for students of below average and average ability level. Precision of theta values decreased as ability level increased.

Discussion:

Insights into the Items and Students:

Having concluded that the Rasch Model was appropriate for the dataset and that all students and items should be included in the analysis, data on the individual items was examined to gain further insight into factors that influenced item difficulty as well as insights into the students who took the exam.

Several trends were observed among the items at the ends of the difficulty spectrum. Four of the five easiest questions were factual recall questions that tested students on their knowledge of environmentally friendly practices: 14 (b = -2.58, SE=0.24), 57 (b= -2.52, SE = 0.24), 30 (b = -1.98, SE = 0.19), & 27 (b = -1.91, SE = 0.19). Students easily recalled that switching from fossil fuels to solar, wind, and geothermal energy would reduce the amount of carbon dioxide released into the atmosphere (item 14) and deduced that maintaining keeping fish populations at 70% of the carrying capacity can help prevent the collapse of fisheries (item 57). Students also admitted that the ecological footprint of the United States exceeds its ecological capacity (item 30). One explanation for the ease of these items is that the material was still fresh in students' minds because it was covered in class the week before the exam. Another possible

explanation is that, given the current emphasis on eco-friendly practices in the media and popular culture students were familiar with this information prior to taking this course. Ideally these facts would not be idle knowledge within students. Instead students would apply this knowledge to their daily lives.

At the other end of the difficulty spectrum, the hardest item on the exam zeroed in on a well-documented misconception. Item 85 (b = 3.89, SE = 0.27) tested students on their knowledge of diffusion. When asked to identify which of the following statements about diffusion is true, 287 of the 358 students selected choice b "molecules move in a directional manner from regions of high concentration to regions of low concentration" rather than choice a "the energy needed for diffusion comes from the kinetic energy of the molecules". This result could reflect that the majority of students who took this exam still do not understand that diffusion is driven by the random motion of molecules. Other researchers have documented this same misconception among undergraduate students (Meir, Perry, Stal, Maruca, & Klopfer, 2005; Odom, 1995). It is also possible that students simply overlooked the word "directional" and focused on the movement of molecules down their concentration gradient. Without the opportunity to interview students who took this exam it is difficult to untangle these possibilities.

It is becoming increasingly recognized that visualization skills are key for scientific thinking (Stanger-Hall, Shockley, & Wilson, 2011). Therefore, it is unfortunate that students struggled with items that required an element of three dimensional thinking or visualization as it suggests that students are weak in an area that is needed for scientific thinking. Items 73 (b = 0.92, SE = 0.11) and 87 (b = 0.61, SE = 0.11) tested students on the concept of surface area to volume ratio. Correctly answering these questions required students to consider the three dimensional shapes of organisms and cells. Items 38 (b = 1.64, SE = 0.12) and 89 (b = 1.13, SE

= 0.11) were two hard items that also included a visualization component. Item 38 asked students to trace the path of an action potential along a neuron using voltmeters but did not provide a diagram of the experimental setup nor did the question advise students to draw a diagram. In contrast, item 89 asked students to consider the intracellular and extracellular concentrations of two ions transported by the same antiporter and advised students to draw a diagram of the process. The item characteristic curves for item 38 items suggests that something besides ability level influenced student performances on this item (Figure 3.5a). Albeit to a lesser extent, the item characteristic curve for item 89 also suggests that something besides ability level confounded students' performance on this item (Figure 3.5b). In contrast to items 38 and 89, item 86 (*b*= -0.37, SE=0.12) provided a clear diagram and was an easy item. It is possible that the visualization component of items 38 and 89 influenced their difficulty; however, neither item included distractors specifically written to identify students with poor visualization skills. Therefore, it is not possible to comment on whether visualization skills undergirded students' performances on these items. Furthermore, it is not possible to comment on whether the diagram in item 86 facilitated students' ability to succeed on the item. Nevertheless, these data suggest a need to provide students with opportunities to hone their visualization skills. This suggestion is supported by research that showed that undergraduate students who participated in visualization based workshops improved their performance on both lower-order and higher-order test items (Stanger-Hall et al., 2011).

While Bloom's Taxonomy has been used to predict or to rate the difficulty of an item (D. Allen & Tanner, 2002; Knaus, Murphy, Blecking, & Holme, 2011; Mesic & Muratovic, 2011) the data from this exam showed that Bloom's taxonomy does not necessarily dictate the item's difficulty level.

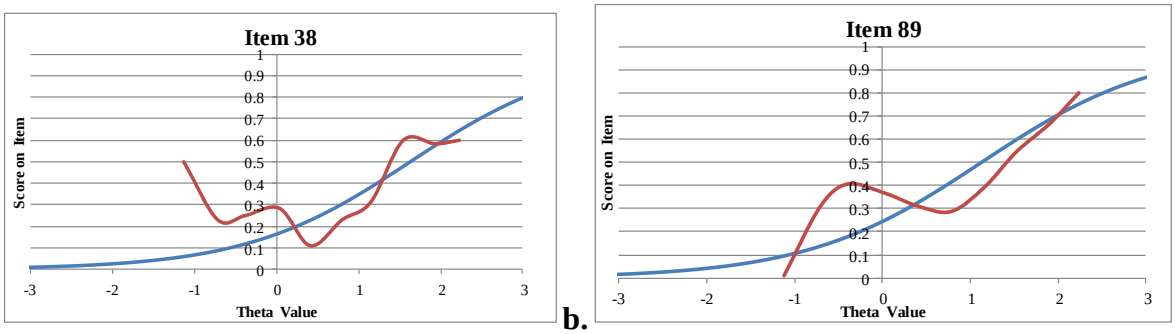


Figure 3.5: Comparisons of empirical and modeled ICCs for items 38 and 89. The blue curve is the Expected Score ICC which represents the model's prediction of the probability that a student with a given ability level will answer the item correctly (blue curve). The red curve is the Observed Score ICC which is the empirical data of how the students performed on this item (red curve). Student ability level is plotted on the x-axis. Lack of alignment between the expected and empirical score ICCs reflect discrepancies between the model's predictions and the empirical data for these two items.

According to the revised Bloom's Taxonomy remembering factual information is the most basic cognitive activity (Krathwohl, 2002). However, item 40 (b = 1.3, SE = 0.12) simply asked students to recall how many million years ago the Cretaceous extinction occurred and yet this item was harder than item 22 (b = 0.87, SE = 0.11) which asked students to apply the definition of aposematic coloration. Items 37 (b = 1.17, SE = 0.12) and 55 (b = 1.24, SE = 0.12) were also factual recall items that fell at the higher end of the difficulty spectrum. The data also showed that there are gradations within the inference level of Bloom's taxonomy. Items 82 (b = -2.22, SE = 0.21), 83 (b = 0.64, SE = 0.11), and 91(b = 0.2, SE = 0.11) all asked students to make inferences based on a phylogenetic tree; however, the complexity of the inferences students needed to make ranged from simple to difficult and the complexity of the inference required by each item was reflected in the item's difficulty value. Earlier in the semester students were tested on how to interpret a phylogenetic tree to identify the closest relative(s) of a taxa. It is therefore encouraging that item 82, which asked students to identify Echinodermata as the closest relatives of Chordata, was an overly easy item. Practice with this type of inference likely facilitated students' ability to solve this item. In contrast, item 83 required students to

identify the subkingdom Bilateria as the closest relatives of the phylum Cnideria. Since Bilateria and Cnideria are on different phylogenetic classification levels, the branching patterns of the phylogenetic tree are more difficult to interpret. Furthermore, students had not been tested on this type of interpretation so they did not have the advantage of practice. Consequently, item 83 was more difficult than item 82. Item 91 asked students to infer the meaning of an internal node on a phylogenetic tree—another novel inference for students. The results suggested that the inference in item 91 was harder than item 82 but not as hard as item 83. As a whole, these data shows that Bloom's taxonomy influences item difficulty but does not solely determine it. These results are consistent with a study of multiple-choice items that appeared on a first year, undergraduate level computer course in computer programming. The researchers found that items that were classified as lower-order thinking according to the revised Bloom's taxonomy were not necessarily easy for students (Shuhidan, Hamilton, & D'Souza, 2009).

Prior studies found that as the cognitive load of an item—the amount of information required to solve an item—increases so does the item's difficulty level (Chalifour & Powers, 1989; Knaus et al., 2011). Of all the items on the exam, items 70 (b=1.04, SE=0.11), 94 (b=1.37, SE=0.12), and 105 (b=0.91, SE=0.11) had the highest cognitive loads, respectively. Students needed to simultaneously consider at least five pieces of information or concepts when solving these items. Most other items on the exam only required students to focus on one concept, one piece of information, or one definition. Item 70 asked students to order a sequence of five evolutionary events. Item 94 asked students to consider the processes of generating skeletal muscle action potentials and the sliding filament theory and predict why a lack of ATP after death would result in rigor mortis. When solving item 105 students needed to consider the functions of seven different hormones and select the correct combination of hormones that would

take in effect in postprandial Joe. With respective discrimination indexes of 1.23 and 1.22, items 70 and 94 sharply discriminated among students. Item 105 was the least cognitively demanding of the three and subsequently item 105 was the easiest and least discriminating of the three items (discrimination index for item 105 = 1.04). Therefore, the cognitive load of items 70, 94, and 105 likely contributed to their high difficulty. These results also suggest that increasing the cognitive load of an item may be one way to distinguish between high and low ability students.

Item level data suggested that the plant life cycles were among the harder topics of this course. Item level data also revealed some underlying gaps in students' knowledge that may have contributed to their difficulties with these topics. Item 8 (b = 2.53, SE=0.16) was a recall question that required students to remember that, in the plant life cycle embryos are produced by the process of mitosis, not plasmogamy, meiosis, dispersal, or fertilization. 231 of the students responded that in the plant life cycle embryos are produced by fertilization. This result could indicate an extreme amount of confusion between a zygote (the unicellular, diploid product of fertilization) and an embryo. However, it could also reflect a rash thought process in which students leapt from fertilization to embryo without considering the steps in between those two stages. Students also struggled with the concept of an embryo on item 69 (b = 0.51, SE = 0.11). Item 69 asked students to recall that the embryo stage in flowering plant reproduction is a multicellular diploid structure and only about half the students were able to do so. 88 of the 358 students responded that an embryo is a single-celled, diploid structure (i.e. a zygote). It is possible that item 8 was much more difficult than item 69 because item 8 allowed did not confront students with their confusion between embryo and zygote while item 69 forced students to consider the definition of an embryo. Along the same line, item 26 (b = 0.55, SE = 0.11) required students to figure out that a plant seed can be likened to an amniotic egg but 108 of the

358 students likened a plant seed to a zygote. Data from items 8, 26, and 69 suggest that these students may have failed to understand the structure of a seed, failed to understand the difference between a zygote and an embryo, failed to understand both concepts, or rushed through the items without giving them proper thought. The distractors for these items are unable to distinguish between the three possibilities. Data from these items may also reflect lingering misconceptions and confusions from the prerequisite course on the concepts of haploid vs. diploid and mitosis vs. meiosis. The gaps in knowledge identified by items 8 and 69 may have influenced the difficulty level of item 92 (b=0.59, SE=0.11) which asked students to apply their knowledge of seeds to figure out that herbicides kill germinating seeds by blocking the process of mitosis. Difficulties with the topic of plant life cycles persisted in items 53 (b = 1.39, SE = 0.11) and 67 (b = 1.02, SE = 0.11). Item 53 asked students to remember that pollen is the gametophyte stage of a flowering plant while item 67 asked students to remember that plants generate spores through the process of meiosis. As a whole, the data from items 8, 26, 53, 69, and 92 supported previous research that the generalized plant life cycle is a challenging topic for introductory biology students to grasp (Stanger-Hall et al., 2011).

Performance of the Author's 23 Critical Thinking Items:

A goal of this project was to examine the performance of the author's multiple-choice items, critical thinking items on an undergraduate-level biology final exam (critical thinking items written by the professor were not considered in this analysis). The author of this paper wrote an initial pool of 41 items that were hypothesized to require critical thinking skills. Validation studies described in chapter 2 of this manuscript showed that only 34 of the 41 items required the critical thinking skills of apply, analyze, or evaluate and that 2 of the 41 items were best classified as Bloom's level 2. The remaining 5 items were eliminated from the pool. The

professor selected 23 of the 34 critical thinking items to include on the exam (Appendix F). The content of the 23 critical thinking items spanned the course syllabus as the items tested students on the topics of: evolution and natural selection, phylogenetics, diffusion, plant biology, animal physiology (renal, nervous, muscular, and endocrine systems), and ecology. Results on the 23 items provided insight into what influences item difficulty and hit on several issues regarding assessing critical thinking in the context of a biology final exam.

A theme among the easier critical thinking items was that they either provided students with all the information they needed to solve the item in a user friendly format (items: 86 (b = -0.37, SE = 0.12), 96 (b = 0.04, SE = 0.11), and 109 (b = -1.23, SE = 0.15)). Item 86 included a clear diagram of a proximal tubule kidney cell and its immediate surroundings and asked students to figure out which change of conditions would increase the rate of GLUT2-mediated glucose export from the cell. Item 109 provided students with the equation for Fick's Law as it pertains to insulin exchange and asked them to infer which variable in the equation changes when the number of capillaries recruited to skeletal muscles is increased. The stem of item 96 provided students with a scientific hypothesis and asked students to select the piece of evidence that best supported the hypothesis. Thus, item 96 was a self-contained item that tested scientific reasoning. Items 88 (b = 0.44, SE = 0.11), 101 (b = 0.95, SE = 0.11), and 104 (b = 1.16, SE = 0.12) also provided students with all the information they needed to solve the item; however, the information provided was not direct and they turned out to be harder items. Items 88 presented students with a graph of lung volume vs. time and asked them to identify the time point(s) when the sarcomeres of the thoracic diaphragm would be contracting. Item 101 presented the information in the form of a bar chart of rates of glucose disposal vs. time and asked students to identify which of the five answer choices was a correct interpretation of the graph. Item 104

provided students with blood and urine osmolality values and asked students to identify the organism for which these values are inconsistent with. The pattern of distractors for item 101 supports the notion that the need to read a graph increased the difficulty level of item 101. Item 101 contained two distractors designed to lure students who failed to correctly interpret the graph and 190 of the 358 students selected one of these two distractors. A wrong answer to item 101 could result from an inability to read a graph, insufficient critical thinking skills, or both. Therefore, while solving item 101 requires critical thinking skills, the item cannot be used as an indicator of critical thinking skills. As with item 101, an incorrect answer to item 88 and/or item 104 could be due the student's inability to understand the information given, insufficient critical thinking skills or both.

The distinction between items 86, 96, and 109 which provided students will all the information needed to solve the item and items such as 94 (b = 1.37, SE = 0.12) 100 (b = 1.67, SE = 0.12), and 102 (b = 1.16, SE = 0.12) which required students to correctly recall the appropriate information needed to solve the item points toward another source of item difficulty for instructors to consider. The difference in difficulty between these two types of items suggests that having to first discern what information is needed to solve the item and then correctly recalling that information is inherently more difficult than simply working through a self-contained item that provides all of the information needed to solve the item. It is also possible that having to retrieve information from the long-term memory and retain it in the short-term memory is more taxing than working with a set of written facts. This notion is grounded in theories of cognitive load. According to theories of cognitive load, the working memory is limited in capacity and it must divide its resources between information storage and information processing. As the information load of a task increases, information processing capacity

declines. Likewise, the short-term memory becomes less able to store information as the information processing aspect of a task increases (Anderson, Reder, & Lebiere, 1996; Barrouillet, Bernardin, Portrat, Vergauwe, & Camos, 2007; Just & Carpenter, 1992). Cognitive load theories also cite the ability to activate the necessary information from the long-term memory and retrieve it into the short-term memory as a fundamental limitation on an individual's ability to complete a task (Anderson et al., 1996; Just & Carpenter, 1992). Test items that provided students with all information circumvented the problem of retrieving information from the long-term memory and lessened the amount of information the working memory needed to store thereby freeing up more capacity for information processing. Given this distinction between item types, items 86, 96, and 109 may be just basic tests of mental processing rather than true tests of critical thinking ability.

The data on the items highlights another issue for college science instructors who want to assess their students' critical thinking skills to consider. In order to be appropriate for the final exam, the majority of the critical thinking items written by the author served the dual purposes of assessing the content material of the course and of requiring critical thinking skills. Solving these items required students to first correctly recall the necessary and appropriate content information and then critically think about the information. Incorporating these two goals into the 23 items caused many of the items to lose focus. The items were not elegant enough to distinguish between students who simply did not recall or know the correct content material and students who recalled the correct content material but were not enough of a critical thinker to solve the item. Item 102 (b = 1.16, SE = 0.12) is an example of this loss of focus (Figure 3.6).

102. The kidneys of healthy individuals secrete erythropoietin. However, individuals with severe kidney disease are unable to produce erythropoietin. Which of the following do you predict is associated with severe kidney disease?

a. High amounts of Red Blood Cells (n = 9)
b. Increased chance of blood clots due to abnormally viscous blood (n = 146)
c. Poor Oxygen delivery to tissues (n = 131)
d. A right-shift of the Oxygen binding to hemoglobin curve (n = 32)
e. Low levels of waste products in the blood (n = 40)

Figure 3.6: Exam Item 102 with pattern of student responses. Exam item 102 is presented as an illustration of an item that requires recall and inference. Answer choice c, indicated in bold, is the correct response. The numbers in parentheses indicate the number of students who selected that answer.

In order to solve item 102 students needed to focus in on erythropoietin, recall that erythropoietin is a hormone that stimulates red blood cell proliferation, and infer that a lack of erythropoietin will lead to anemia and poor oxygen delivery to tissues. The main function of the kidneys is to filter waste products from the blood. Did students gravitate to answer choice b because they simply expected kidney disease to result in something abnormal with the blood? Alternatively, did students fail to focus on erythropoietin and/or fail to correctly recall the function of erythropoietin? Or were students unable to correctly infer the consequences of insufficient levels of erythropoietin? The item is wholly unable to identify where students erred and what they struggled with. Overall the items written by the author of this paper challenged the students to go beyond simply recalling facts. However, the items cannot be used to diagnose students' critical thinking abilities.

Conclusions:

In this paper the Rasch Model was applied to final exam data from a second semester introductory biology course for science majors. The exam contained 113 multiple-choice items that covered a broad range of topics. 23 of these items were written by the author of this paper with the intention of assessing students' critical thinking skills. It was concluded that the Rasch

Model was most appropriate for this dataset. Results from the Rasch analysis provided insights into item difficulty levels, student difficulties, the exam, and assessing critical thinking in undergraduate biology courses.

With respect to item difficulty levels, items that hit on familiar topics tended to be the easiest items whereas items that hit on common misconceptions tended to be the hardest. The data from the Rasch analysis also suggest that increasing the amount of information students need to consider to solve an item increases the item's difficulty and discrimination. Even though Bloom's taxonomy is often regarded as a hierarchical scale, the data presented here showed that there are gradations within the inference level of the taxonomy and that an item's position in the taxonomy is not the sole determiner of item difficulty. Finally, the data in this paper indicate the need for future research into the effects of requiring visualization skills on item difficulty and students' visualization skills.

The students who took this exam displayed some common mistakes among undergraduate students regarding the topics of diffusion and mitosis vs. meiosis. Students' patterns of responses also showed that their misconceptions and/or rash thought process hindered their ability to master the topic of plant life cycles. These data provide insight for teaching future crops of students as they will likely share the same misconceptions as the Spring 2012 students.

The Rasch analysis showed that the exam was well matched to the majority of students but was unable to precisely assess the ability levels of the students at the top of the class. These results show that the exam was in need of harder questions. Since the set of 23 critical thinking items written by the author were not challenging enough, it can be concluded that the level of difficulty of the critical thinking items should be increased as well. The goal is to write an exam that contains a set of items whose difficulty values encompasses the range of student ability

levels. Furthermore, the items should be evenly distributed over the range of item difficulty values (Wright, 1977). The results from this analysis suggest that instructors can increase the difficulty level of critical thinking items by increasing the cognitive load of items and incorporating elements of visualization into items.

The overall message from the set of 23 critical thinking items written by the author is that it is extremely important to be clear about the purpose of each item. If the purpose of the item is to assess students' basic thinking skills rather than their ability to recall factual information, then the item should contain all the information needed to solve the item. However, if the goal of the item is to assess if students are fluent enough with the material to recall the correct content information and critically think about it, then the approach used in this paper would be appropriate. While the conflated approach of requiring students to recall information and think critically used to write critical thinking items for this exam was not optimal, requiring students to be fluent enough with the content material of the course that they can correctly retrieve the appropriate information to solve a test item is a worthy instructional goal and may be best addressed through free response items. The results of this analysis suggest that multiple-choice items should be more specific. Item 87 (Figure 3.7) is an example of an item that should be revised so that it targets students' specific weaknesses.

87. Which of the following would result from reducing the surface area to volume ratio (SA/V) of a cell (same shape):
    a. **The time it takes O_2 to diffuse from the cell surface to the mitochondria would increase. (n = 174)**
    b. The amount of glucose needed to fuel the cell would decrease. (n = 133)
    c. The amount of genomic DNA would increase. (n = 6)
    d. The surface area would increase at a faster rate than the volume. (n = 37)
    e. The cell membrane would become porous. (n = 8)

Figure 3.7: Exam Item 87 with pattern of student responses. Exam item 87 is presented as an example of an unfocused item in need of refinement. The correct response is in bold. The numbers in parentheses indicate the number of students who selected that answer.

To solve item 87, students need to recall that if the surface area to volume ratio of a cell has decreased, the cell has grown in size. This information was developed and concluded in class and is present in the textbook so it is most appropriately classified as factual recall; however, students who cannot recall this information can evaluate the fraction to figure it out. Since the volume is greater, the distance oxygen needs to diffuse to reach the mitochondria increases and

therefore so does the amount of time it takes for this to occur. Unfortunately the distractors do not provide insight into students' sources of error. It is possible that distractor choice b was the most popular because students read the word "reducing" in the question stem and gravitated to the word "decrease" in distractor b. This item was refined so that it provides useful information about students' sources of confusion and so that it contains two distractors that refer to a decrease (Figure 3.8). Additionally, the number of distractors was reduced to three to prevent inclusion of an oddball distractor.

87. Which of the following would result from reducing the surface area to volume ratio (SA/V) of a cell (same shape):
    a. **The time it takes O_2 to diffuse from the cell surface to the mitochondria would increase.**
    b. The cell's size would decrease.
    c. The amount of glucose needed to fuel the cell would decrease.
    d. The surface area would increase at a faster rate than the volume.

Figure 3.8: Refined version of exam item 87. The refined version of this item is presented as an item that could identify student misconceptions. The correct answer is indicated in bold type.

Students who can recall that reducing the surface area to volume ratio of a cell means that the cell's size has increased are unlikely to select distractor b while students who can successfully work with the SA/V fraction are unlikely to select distractor d. Students who either recall or analyze their way to the correct information are then left with choices a and c. Since a

larger cell requires more glucose, distractor c is incorrect. Students can then verify that answer choice a is correct by inferring the aforementioned explanation that if the cell has grown in size, it will take longer for oxygen to diffuse a greater distance to the mitochondria because it has a greater distance to diffuse. However, these hypothesized responses to the refined version of item 87 need to be validated by piloting the refined item with students.

Item 87 illustrates that the process of developing and validating multiple-choice test items that require critical thinking skills used in this paper is iterative. Test items were first written and pilot tested. Items that survived the pilot test stage were revised as needed and included on a final exam. The Rasch Model was then applied to the dataset. After verifying that the Rasch Model was appropriate for the dataset, item performance were then analyzed to identify poorly performing items and to yield insights into both characteristics of item difficulty and students' misconceptions. While this experiment did not provide the opportunity to solicit qualitative feedback on the items from students who took this exam, such data would be very useful. As time consuming as the methods used in this experiment were, they are well worth the time because multiple-choice items that require critical thinking skills and identify students' misconceptions are a very powerful tool for instructors who aim to foster their students' critical thinking skills. The author of this paper encourages college science instructors to build on the qualitative and quantitative methods used in this research study. A potential first step would be to investigate the potential for items of high cognitive-load items and for visualization items to reliably distinguish between A and B students.

REFERENCES

AAMC. (2009). Content Outline for the Biological Science Section of the MCAT.

Alberts, B. (2009). Redefining Science Education. *Science, 323*(5913), 437. doi: 10.1126/science.1170933

Allen, D., & Tanner, K. (2002). Approaches to Cell Biology Teaching: Questions about Questions. *Cell Biology Education, 1*(3), 63-67. doi: 10.1187/cbe.02-07-0021

Allen, M. J., & Yen, W. M. (2002). *Introduction to Measurement Theory*. Long Grove: Waveland Press, Inc.

Anderson, J. R., Reder, L. M., & Lebiere, C. (1996). Working Memory: Activation Limitations on Retrieval. [doi: 10.1006/cogp.1996.0007]. *Cognitive Psychology, 30*(3), 221-256.

Baker, F. B. (2004). *Item response theory : parameter estimation techniques* (2nd ed., rev. and expanded. ed.). Marcel Dekker: New York.

Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., & Camos, V. r. (2007). Time and cognitive load in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(3), 570-585. doi: 10.1037/0278-7393.33.3.570

Bell, R. L. (1982). Person Fit and Person Reliability. *Education Research and Perspectives, 9*(1), 105-113.

Bissell, A. N., & Lemons, P. P. (2006). A New Method for Assessing Critical Thinking in the Classroom. *BioScience, 56*(1), 66-72. doi: 10.1641/0006-3568(2006)056[0066:anmfac]2.0.co;2

Bloom, Engelhart, B. M., Furst, E., Hill, W., & Kratwohl, D. (1956). *Taxonomy of educational objectives: The classification of educational goals.* New York: McKay.

Bock, R. D. (1997). A Brief History of Item Response Theory. *Educational Measurement: Issues and Practice, 16*(4), 21-33.

Campbell, N. A. (2010). *Biology* (Vol. Ninth Edition). San Francisco: Benjamin Cummings.

Chalifour, C. L., & Powers, D. E. (1989). The Relationship of Content Characteristics of GRE Analytical Reasoning Items to Their Difficulties and Discriminations. *Journal of Educational Measurement, 26*(2), 120-132.

Childs, R. A., & Oppler, S. H. (2000). Implications of Test Dimensionality for Unidimensional Irt Scoring: An Investigation of a High-Stakes Testing Program. *Educational and Psychological Measurement, 60*(6), 939-955. doi: 10.1177/00131640021971005

Crowe, A., Dirks, C., & Wenderoth, M. P. (2008). Biology in Bloom: Implementing Bloom's Taxonomy to Enhance Student Learning in Biology. *CBE Life Sciences Education, 7*, 368-381.

Dancy, M. H., & Beichner, R. J. (2002). But Are They Learning? Getting Started in Classroom Evaluation. *Cell Biology Education, 1*(3), 87-94. doi: 10.1187/cbe.02-04-0010

De Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York: The Guilford Press.

De Ayala, R. J. (2010). Item Response Theory. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 155-172). Routledge: New York.

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education, 44*(1), 109-117. doi: 10.1111/j.1365-2923.2009.03425.x

Embretson, S. E. (2000). *Item response theory for psychologists*. L. Erlbaum Associates: Mahwah, N.J.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*: L. Erlbaum Associates.

Ennis, R. H. (1985). A Logical Basis for Measuring Critical Thinking Skills. *Educational Leadership, 43*(2), 44-48.

Ennis, R. H., Millman, J., & Tomko, T. N. (1985). Cornell Critical Thinking Tests. *CCTT*.

Haladyna, T. M., & Downing, S. M. (1989a). A Taxonomy of Multiple-Choice Item-Writing Rules. [Article]. *Applied Measurement in Education, 2*(1), 37.

Haladyna, T. M., & Downing, S. M. (1989b). Validity of a Taxonomy of Multiple-Choice Item-Writing Rules. [Article]. *Applied Measurement in Education, 2*(1), 51.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education, 15*(3), 309-333. doi: 10.1207/s15324818ame1503_5

Hambleton, R. K., & Jones, R. W. (1993). An NCME Instructional Module on. *Educational Measurement: Issues and Practice, 12*(3), 38-47. doi: 10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications: Newbury Park, Calif.

Heyborne, W., Clarke, J., & Perrett, J. (2011). A Comparison of Two Forms of Assessment in an Introductory Biology Laboratory Course. *Journal of College Science Teaching, 40*(5), 28.

Holmang, A., Mimura, K., Bjorntorp, P., & Lonnroth, P. (1997). Interstitial muscle insulin and glucose levels in normal and insulin-resistant Zucker rats. *Diabetes, 46*(11), 1799 - 1804.

IBM. (2010). IBM SPSS Statistics 19.0 (Version 19.0).

Jackson, T. R., Draugalis, J. R., Slack, M. K., Zachry, W. M., & D'Agostino, J. (2002). Validation of Authentic Performance Assessment: A Process Suited for Rasch Modeling. *American Journal of Pharmaceutical Education, 66*(Fall), 233-243.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*(1), 122-149. doi: 10.1037/0033-295x.99.1.122

Knaus, K., Murphy, K., Blecking, A., & Holme, T. (2011). A Valid and Reliable Instrument for Cognitive Complexity Rating Assignment of Chemistry Exam Items. *Journal of Chemical Education, 88*(5), 554-560. doi: 10.1021/ed900070y

Krathwohl, D. R. (2002). A Revision of Bloom's Taxonomy: An Overview. *Theory into Practice, 41*(4), 213-218.

Lim, R. G., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology, 75*(2), 164-174. doi: 10.1037/0021-9010.75.2.164

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328.

Linacre, J. M. (2002). Optimizing Rating Scale Category Effectiveness. *Journal of Applied Measurement, 3*(1), 85-106.

Linacre, J. M. (2004). Discrimination, Guessing, and Carelessness Asymptotes: Estimating IRT Parameters with Rasch. *Rasch Measurement Transactions, 18*(1), 959-960.

Linacre, J. M. (2009). Local Independence and Residual Covariance: A Study of Olympic Figure Skating Ratings. *Journal of Applied Measurement, 10*(2), 1-14.

Linacre, J. M. (2012a). Winsteps (Version 3.74.0) [Computer Software] (Version 3.74.0). Beaverton, Oregon: Winsteps.com. Retrieved on April 18th, 2012. Available at http://www.winsteps.com/. Retrieved from http://www.winsteps.com/

Linacre, J. M. (2012b). *Winsteps Rasch Measurement computer program User's Guide.* Beaverton, Oregon: Winsteps.com.

Linacre, J. M., & Wright, B. D. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Lord, T., & Baviskar, S. (2007). Moving Students from Information Recitation to Information Understanding: Exploiting Bloom's Taxonomy in Creating Science Questions. *Journal of College Science Teaching, 36*(5), 40-44.

Martinez, M. E. (1999). Cognition and the question of test item format. [doi: 10.1207/s15326985ep3404_2]. *Educational Psychologist, 34*(4), 207-218. doi: 10.1207/s15326985ep3404_2

Meir, E., Perry, J., Stal, D., Maruca, S., & Klopfer, E. (2005). How Effective Are Simulated Molecular-level Experiments for Teaching Diffusion and Osmosis? *Cell Biology Education, 4*(3), 235-248. doi: 10.1187/cbe.04-09-0049

Mesic, V., & Muratovic, H. (2011). Identifying Predictors of Physics Item Difficulty: A Linear Regression Approach. *Physical Review Special Topics - Physics Education Research, 7*(1), 010110-010111-010110-010115.

Messick, S. (1995). Validity of Psychological Assessment. *American Psychologist, 50*(9), 741-749.

Molenaar, I. W. (1995). Some Background for Item Response Theory and the Rasch Model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models Foundations, Recent Developments, and Applications* (pp. 3-14). New York: Springer-Verlag.

Momsen, J. L., Long, T. M., Wyse, S. A., & Ebert-May, D. (2010). Just the Facts? Introductory Undergraduate Biology Courses Focus on Low-Level Cognitive Skills. *CBE Life Sci Educ, 9*(4), 435-440. doi: 10.1187/cbe.10-01-0001

Morrison, D. A. (1996). Phylogenetic tree-building. [doi: 10.1016/0020-7519(96)00044-6]. *International Journal for Parasitology, 26*(6), 589-617.

National Center for Biotechnology Information, & U.S. National Library of Medicine. (2011). PubMed (Internet), from http://www.ncbi.nlm.nih.gov/pubmed/

Nestler, J. (2000). Obesity, insulin, sex steroids and ovulation. *International Journal of Obesity, 24*(Suppl 2), S71-S73.

Norris, S. P. (1990). Effect of Eliciting Verbal Reports of Thinking on Critical Thinking Test Performance. *Journal of Educational Measurement, 27*(1), 41-58.

NRC, N. R. C. (2001). *Knowing what students know : the science and design of educational assessment*. National Academy Press: Washington, DC.

Odom, A. L. (1995). Secondary & College Biology Students' Misconceptions about Diffusion & Osmosis. *The American Biology Teacher, 57*(7), 409-415.

Reeve, B. B., & Fayers, P. (2005). Applying item response theory to modeling for evaluating questionnaire items and scale properties. In P. M. Fayers & R. D. Hays (Eds.), *Assessing*

quality of life in clinical trials: Methods and practice (Vol. 2, pp. 55-73). New York, NY: Oxford University Press.

Reise, S. P., Ainsworth, A. T., & Haviland, M. G. (2005). Item Response Theory: Fundamentals, Applications, and Promise in Psychological Research. *Current Directions in Psychological Science, 14*(2), 95-101.

Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods, 8*(2), 164-184. doi: 10.1037/1082-989x.8.2.164

Roth, S. M., Martel, G. F., Ivey, F. M., Lemmer, J. T., Metter, E. J., Hurley, B. F., & Rogers, M. A. (2000). High-volume, heavy-resistance strength training and muscle damage in young and older women. *J Appl Physiol, 88*, 1112-1118.

Shuhidan, S., Hamilton, M., & D'Souza, D. (2009). A taxonomic study of novice programming summative assessment. In M. Hamilton & T. Clear (Eds.), *Eleventh Australasian Computing Education Conference (ACE 2009)* (Vol. 95 of CRPIT, pp. 147-156). Wellington, New Zealand: ACS.

Simpson, D. E., & Cohen, E. B. (1985). Problem Solving Questions for Multiple Choice Tests: A Method for Analyzing the Cognitive Demands of Items.

Stanger-Hall, K. F., Shockley, F. W., & Wilson, R. E. (2011). Teaching Students How to Study: A Workshop on Information Processing and Self-Testing Helps Students Learn. *CBE-Life Sciences Education, 10*(2), 187-198. doi: 10.1187/cbe.10-11-0142

Tan, R. J. B. (2008). *A Mixed-Methods Approach to Test Evaluation Using Explanatory Item Response Modeling and Think-Alouds.* Doctor of Philosophy, University of California, Berkeley.

Taylor, K. L., & Dionne, J.-P. (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology, 92*(3), 413-425. doi: 10.1037/0022-0663.92.3.413

Tomanek, D., & Montplaisir, L. (2004). Students' Studying and Approaches to Introductory Biology. *Cell Biology Education, 3*(4), 253-262. doi: 10.1187/cbe.04-06-0041

Weld, J., Stier, M., & McNew-Birren, J. (2011). The Development of a Novel Measure of Scientific Reasoning Growth Among College Freshmen: The Constructive Inquiry Science Reasoning Skills Test. *Journal of College Science Teaching, 40*(4), 101.

White, B., Stains, M., Escriu-Sune, M., Medaglia, E., Rostamnjad, L., Chinn, C., & Sevian, H. (2011). A Novel Instrument for Assessing Students' Critical Thinking Abilities. *Journal of College Science Teaching, 40*(5), 102.

Wood, W. B. (2009). Innovations in teaching undergraduate biology and why we need them. *Annu Rev Cell Dev Biol, 25*, 93-112. doi: 10.1146/annurev.cellbio.24.110707.175306

Wright, B. D. (1977). Solving Measurement Problems with the Rasch Model. *Journal of Educational Measurement, 14*(2), 97-116.

Zheng, A. Y., Lawhorn, J. K., Lumley, T., & Freeman, S. (2008). Application of Bloom's Taxonomy Debunks the "MCAT Myth". *Science, 319*(5862), 414-415. doi: 10.1126/science.1147852

Appendix A: Item parameters, fit statistics, and estimates for the 2008 and 2010 exams.
The *b* values are the item parameters, the infit and outfit mean square values are the fit statistics, and the discrimination and asymptote values are the item estimates.

Item	*b* value	Error	Infit Mean Square	Outfit Mean Square	Discrimination	Lower Asymptote	Upper Asymptote
1	0.61	0.07	0.970	0.957	1.170	0.000	1.000
2	-0.07	0.08	0.971	0.975	1.073	0.000	1.000
3	-0.93	0.09	1.027	1.169	0.951	0.000	0.990
4	-0.24	0.08	0.978	1.008	1.041	0.000	1.000
5	-0.6	0.09	0.972	0.970	1.037	0.000	1.000
6	0.18	0.08	1.012	1.045	0.930	0.000	0.975
7	-0.06	0.08	1.008	1.005	0.981	0.000	0.996
8	-1.57	0.12	0.954	0.832	1.047	0.000	1.000
9	2.57	0.1	1.084	1.215	0.897	0.022	1.000
10	0.24	0.07	1.041	1.050	0.844	0.096	0.973
11	-0.56	0.09	1.034	1.063	0.941	0.686	0.989
12	-0.45	0.08	0.960	0.901	1.086	0.002	1.000
13	-0.53	0.08	1.043	1.140	0.913	0.000	0.980
14	0.36	0.07	0.931	0.906	1.321	0.000	1.000
15	-0.03	0.08	0.897	0.856	1.301	0.000	1.000
16	0.53	0.07	1.045	1.052	0.766	0.086	0.963
17	1.34	0.08	0.989	0.993	1.033	0.000	1.000
18	-0.25	0.08	0.965	0.930	1.087	0.000	1.000
19	-0.35	0.08	1.002	1.002	0.995	0.024	1.000
20	-0.88	0.09	0.967	0.935	1.040	0.000	1.000
21	0.65	0.07	0.981	0.973	1.107	0.000	1.000
22	0.85	0.07	1.007	1.014	0.959	0.000	0.981
23	0.51	0.07	1.059	1.076	0.691	0.146	0.960
24	1.19	0.07	1.020	1.048	0.892	0.042	1.000
25	1.19	0.07	1.000	1.013	0.991	0.000	0.979
26	0.01	0.08	0.929	0.884	1.232	0.000	1.000
27	-0.52	0.08	0.961	0.916	1.073	0.000	1.000
28	1.05	0.07	1.046	1.060	0.774	0.049	0.945
29	-1.36	0.11	0.969	0.884	1.035	0.000	1.000
30	0.3	0.07	0.957	0.961	1.170	0.000	1.000
31	0.61	0.07	0.983	0.970	1.105	0.000	1.000
32	0.43	0.07	0.951	0.928	1.249	0.018	1.000
33	1.46	0.08	1.030	1.040	0.906	0.023	0.997
34	0.5	0.07	0.928	0.924	1.356	0.000	1.000
35	0.75	0.07	1.047	1.045	0.749	0.074	0.964
36	-0.31	0.08	1.024	1.001	0.962	0.139	0.994
37	-0.52	0.08	0.996	1.022	0.998	0.000	1.000
38	-0.12	0.08	0.933	0.907	1.173	0.000	1.000
39	0.01	0.08	0.963	0.959	1.114	0.062	1.000
40	0.52	0.07	0.908	0.882	1.486	0.000	1.000

Item Number	*b* value	Error	Infit Mean Square	Outfit Mean Square	Discrimination	Lower Asymptote	Upper Asymptote
41	-2.01	0.14	0.991	0.941	1.011	0.000	1.000
42	-0.73	0.09	1.069	1.203	0.879	0.376	0.973
43	0.58	0.07	1.040	1.043	0.790	0.038	0.952
44	-0.84	0.09	0.995	0.923	1.021	0.109	1.000
45	-1.16	0.1	1.065	1.271	0.907	0.000	0.981
46	-0.88	0.09	1.005	1.019	0.992	0.653	1.000
47	-1.71	0.12	0.988	1.030	1.002	0.000	1.000
48	0.63	0.07	1.002	1.010	0.980	0.007	1.000
49	-0.57	0.09	1.044	1.132	0.917	0.508	0.983
50	-0.51	0.08	1.024	1.069	0.953	0.444	0.992
51	-1.51	0.11	0.977	0.938	1.021	0.000	1.000
52	-1.03	0.1	0.992	0.963	1.012	0.000	1.000
53	0.7	0.07	0.977	0.973	1.126	0.000	1.000
54	-0.07	0.08	1.078	1.145	0.762	0.124	0.946
55	-0.6	0.09	0.965	0.936	1.056	0.000	1.000
56	0.91	0.07	0.933	0.921	1.355	0.000	1.000
57	1.28	0.07	0.993	0.996	1.022	0.001	1.000
58	-0.14	0.08	0.984	0.994	1.028	0.000	1.000
59	1.66	0.08	1.068	1.124	0.818	0.042	0.899
60	-0.47	0.08	0.964	0.922	1.072	0.000	1.000
61	0.92	0.07	1.088	1.096	0.544	0.096	0.886
62	-0.07	0.08	0.971	0.939	1.090	0.000	1.000
63	-0.22	0.12	1.116	1.181	0.760	0.614	0.952
64	0.48	0.11	1.043	1.039	0.823	0.036	0.964
65	0.01	0.12	1.107	1.151	0.717	0.226	0.940
66	-1.28	0.16	0.997	0.941	1.010	1.000	1.000
67	1.34	0.11	1.031	1.003	0.908	0.000	0.868
68	-0.07	0.12	0.984	0.972	1.039	0.000	1.000
69	-0.12	0.12	1.054	1.111	0.859	0.224	0.973
70	0.39	0.11	1.001	1.026	0.988	0.136	1.000
71	-0.07	0.12	1.007	1.039	0.970	0.000	0.993
72	-0.06	0.12	0.957	0.943	1.105	0.000	1.000
73	0.04	0.12	1.043	1.072	0.878	0.115	0.977
74	0.77	0.11	1.078	1.102	0.574	0.178	0.967
75	-0.19	0.12	0.952	0.931	1.097	0.000	1.000
76	1.52	0.11	0.955	0.976	1.127	0.000	1.000
77	-0.44	0.13	1.076	1.120	0.871	0.200	0.973
78	0.38	0.11	1.070	1.074	0.735	0.163	0.958
79	1.32	0.11	1.015	1.007	0.950	0.016	1.000
80	-2.38	0.25	0.943	0.745	1.042	0.000	1.000

Item Number	*b* value	Error	Infit Mean Square	Outfit Mean Square	Discrimination	Lower Asymptote	Upper Asymptote
81	2.21	0.13	0.961	0.922	1.067	0.000	1.000
82	-0.12	0.12	0.996	0.938	1.038	0.071	1.000
83	-1.31	0.17	0.968	0.847	1.042	1.000	1.000
84	-0.9	0.15	0.921	0.802	1.099	0.000	1.000
85	-0.49	0.13	1.028	1.043	0.958	0.000	0.992
86	-0.61	0.13	0.906	0.824	1.139	0.141	1.000
87	-1.72	0.19	0.953	0.897	1.035	0.000	1.000
88	-0.39	0.13	0.960	1.001	1.050	0.000	1.000
89	-0.31	0.12	0.998	0.998	1.005	0.361	1.000
90	0.43	0.11	0.892	0.872	1.444	0.000	1.000
91	-1.15	0.16	0.926	0.857	1.073	0.000	1.000
92	1.99	0.12	0.979	1.004	1.026	0.000	1.000
93	0.7	0.11	0.945	0.927	1.292	0.000	1.000
94	1.48	0.11	1.008	1.034	0.952	0.018	1.000
95	0.39	0.11	1.201	1.294	0.157	0.219	0.807
96	-0.59	0.13	0.910	0.825	1.135	0.000	1.000
97	-0.41	0.13	1.008	0.996	0.989	0.016	1.000
98	-1.23	0.16	0.916	0.727	1.094	0.000	1.000
99	1.08	0.11	1.069	1.093	0.640	0.088	0.945
100	-0.74	0.14	0.969	0.865	1.058	0.280	1.000
101	-2.59	0.28	1.010	0.948	0.998	0.000	1.000
102	-0.18	0.12	0.924	0.892	1.161	0.000	1.000
103	1.45	0.11	1.057	1.064	0.802	0.047	0.986
104	-0.88	0.14	1.048	1.095	0.942	1.000	0.990
105	1.36	0.1	1.014	1.035	0.946	0.019	1.000
106	0.95	0.1	1.187	1.251	0.010	0.213	0.748
107	0.92	0.1	0.975	0.972	1.128	0.003	1.000
108	0.77	0.1	1.031	1.043	0.813	0.075	1.000
109	1.3	0.1	0.982	0.993	1.051	0.003	1.000
110	-0.5	0.11	0.929	0.884	1.136	0.000	1.000
111	1.49	0.1	1.146	1.205	0.600	0.083	0.654
112	-0.7	0.12	0.981	1.005	1.019	0.000	1.000
113	2.56	0.13	1.068	1.109	0.936	0.012	0.845
114	-1.05	0.13	1.023	1.036	0.980	1.000	0.997
115	-0.13	0.1	0.999	0.974	1.017	0.051	1.000
116	1.74	0.11	0.961	0.945	1.078	0.000	1.000
117	0.57	0.1	0.974	1.014	1.110	0.000	1.000
118	-0.49	0.11	0.913	0.834	1.181	0.091	1.000
119	0.6	0.1	0.975	0.956	1.172	0.000	1.000
120	-0.52	0.11	1.076	1.183	0.836	0.142	0.963

Item Number	*b* value	Error	Infit Mean Square	Outfit Mean Square	Discrimination	Lower Asymptote	Upper Asymptote
121	0.73	0.1	0.957	0.938	1.270	0.000	1.000
122	-2.08	0.18	0.992	0.956	1.006	0.000	1.000
123	-0.87	0.12	0.899	0.802	1.142	0.000	1.000
124	-2.47	0.22	0.977	0.802	1.023	0.000	1.000
125	-2.86	0.26	0.965	0.874	1.026	0.000	1.000
126	-1.86	0.17	0.937	0.770	1.056	0.000	1.000
127	0.46	0.1	1.006	1.015	0.954	0.020	1.000
128	1.42	0.1	0.989	1.022	1.014	0.000	1.000
129	-0.45	0.11	0.930	0.884	1.145	0.000	1.000
130	-0.23	0.1	0.929	0.901	1.178	0.000	1.000
131	-0.69	0.11	1.054	1.075	0.916	0.000	0.981
132	-0.61	0.11	0.951	0.912	1.084	0.000	1.000
133	-0.12	0.1	0.983	0.939	1.073	0.124	1.000
134	-0.96	0.12	0.933	0.855	1.089	0.000	1.000
135	1.35	0.1	0.999	1.014	0.990	0.014	1.000
136	-1.22	0.13	0.931	0.790	1.084	0.000	1.000
137	-1.1	0.13	0.970	0.946	1.033	0.000	1.000
138	0.84	0.1	1.145	1.159	0.203	0.166	0.805
139	-0.45	0.11	1.007	0.990	0.990	0.000	0.999
140	-0.43	0.11	1.010	0.983	0.993	0.321	1.000
141	0.01	0.1	1.018	1.023	0.936	0.229	1.000
142	-1.65	0.16	0.984	0.938	1.015	0.000	1.000
143	-0.27	0.11	1.020	1.054	0.935	0.000	0.985
144	-0.13	0.1	0.933	0.893	1.205	0.000	1.000
145	1.28	0.1	1.038	1.048	0.868	0.038	1.000
146	0.77	0.1	0.980	0.969	1.126	0.012	1.000
147	0.71	0.1	1.071	1.108	0.542	0.102	0.907
148	0.23	0.1	1.001	0.983	1.012	0.023	1.000
149	1.06	0.1	1.093	1.117	0.573	0.096	0.901
150	0.47	0.1	0.995	0.988	1.033	0.000	1.000
151	1.72	0.11	0.964	0.939	1.077	0.000	1.000
152	0.69	0.1	0.993	0.988	1.045	0.041	1.000
153	-1.95	0.17	1.017	1.051	0.986	0.000	0.999
154	0.11	0.1	1.016	0.998	0.953	0.215	1.000
155	-1.39	0.14	1.065	1.128	0.936	1.000	0.988
156	1.1	0.1	0.975	0.967	1.112	0.000	1.000
157	1.38	0.1	1.016	1.060	0.929	0.020	1.000
158	0.05	0.1	1.082	1.126	0.682	0.121	0.933
159	-0.9	0.12	1.010	1.030	0.982	0.000	0.997
160	0.99	0.1	1.078	1.073	0.643	0.084	0.940

Appendix B: Arrangement of student responses by item difficulty. Items are ordered by increasing difficulty. Student responses are coded as 1 (correct), 0 (incorrect), or . (not administered). Students tended to respond correctly to items with difficulty levels that were below their ability level and tended to respond incorrectly to items with difficulty levels that were greater than their ability level.

Item Number	Item *b* value	Student Number (Theta Value):		
		34 (1.17)	264 (0.79)	570 (0.66)
125	-2.86	.	.	1
101	-2.59	1	1	.
124	-2.47	.	.	1
80	-2.38	1	1	.
122	-2.08	.	.	1
41	-2.01	1	1	1
153	-1.95	.	.	1
126	-1.86	.	.	1
87	-1.72	1	1	.
47	-1.71	1	1	1
142	-1.65	.	.	1
8	-1.57	1	1	1
51	-1.51	1	1	1
155	-1.39	.	.	1
29	-1.36	0	1	1
83	-1.31	1	1	.
66	-1.28	0	1	.
98	-1.23	1	0	.
136	-1.22	.	.	1
45	-1.16	1	1	1
91	-1.15	1	1	.
137	-1.1	.	.	1
114	-1.05	.	.	1
52	-1.03	1	1	1
134	-0.96	.	.	0
3	-0.93	1	1	1
84	-0.9	1	1	.
159	-0.9	.	.	1
20	-0.88	1	1	1
46	-0.88	1	1	1
104	-0.88	1	1	.
123	-0.87	.	.	1
44	-0.84	1	1	0
100	-0.74	1	1	.
42	-0.73	1	0	1
112	-0.7	.	.	1
131	-0.69	.	.	1
86	-0.61	1	1	.
132	-0.61	.	.	1
5	-0.6	0	0	1

Item Number	Item *b* value	Student Number (Theta Value):		
		34 (1.17)	264 (0.79)	570 (0.66)
55	-0.6	1	1	1
96	-0.59	1	1	.
49	-0.57	0	1	1
11	-0.56	1	0	1
13	-0.53	1	1	1
27	-0.52	1	1	1
37	-0.52	1	1	0
120	-0.52	.	.	1
50	-0.51	1	1	1
110	-0.5	.	.	1
85	-0.49	1	0	.
118	-0.49	.	.	1
60	-0.47	1	1	1
12	-0.45	1	1	1
129	-0.45	.	.	1
139	-0.45	.	.	1
77	-0.44	1	1	.
140	-0.43	.	.	1
97	-0.41	1	1	.
88	-0.39	0	1	.
19	-0.35	1	1	1
36	-0.31	0	1	1
89	-0.31	0	1	.
143	-0.27	.	.	1
18	-0.25	1	1	0
4	-0.24	1	1	0
130	-0.23	.	.	0
63	-0.22	1	1	.
75	-0.19	1	0	.
102	-0.18	1	0	.
58	-0.14	1	1	1
115	-0.13	.	.	1
144	-0.13	.	.	1
38	-0.12	1	0	1
69	-0.12	1	1	.
82	-0.12	1	1	.
133	-0.12	.	.	1
2	-0.07	1	1	0
54	-0.07	1	0	1
62	-0.07	1	0	1

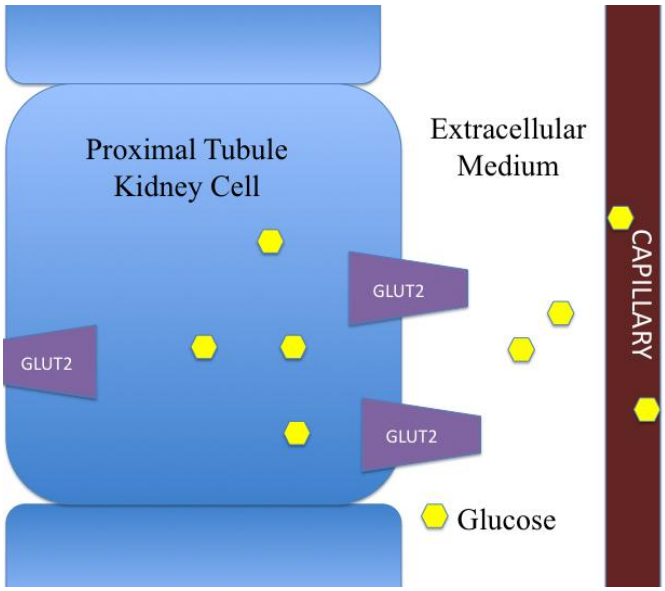
Item Number	Item *b* value	Student Number (Theta Value):		
		34 (1.17)	264 (0.79)	570 (0.66)
68	-0.07	1	1	.
71	-0.07	1	1	.
7	-0.06	1	0	1
72	-0.06	1	1	.
15	-0.03	1	0	0
26	0.01	1	1	0
39	0.01	0	1	1
65	0.01	1	0	.
141	0.01	.	.	1
73	0.04	0	1	.
158	0.05	.	.	1
154	0.11	.	.	1
6	0.18	1	0	1
148	0.23	.	.	0
10	0.24	0	1	1
30	0.3	1	0	0
14	0.36	1	1	0
78	0.38	1	1	.
70	0.39	1	1	.
95	0.39	0	1	.
32	0.43	0	0	0
90	0.43	1	0	.
127	0.46	.	.	0
150	0.47	.	.	0
64	0.48	1	1	.
34	0.5	1	1	1
23	0.51	1	1	0
40	0.52	0	1	0
16	0.53	1	0	1
117	0.57	.	.	0
43	0.58	1	0	1
119	0.6	.	.	1
1	0.61	1	0	0
31	0.61	1	0	0
48	0.63	1	1	1
21	0.65	1	1	0
152	0.69	.	.	1
53	0.7	1	0	1
93	0.7	0	1	.
147	0.71	.	.	0

Item Number	Item *b* value	Student Number (Theta Value):		
		34 (1.17)	264 (0.79)	570 (0.66)
121	0.73	.	.	0
35	0.75	0	1	0
74	0.77	0	1	.
108	0.77	.	.	0
146	0.77	.	.	0
138	0.84	.	.	0
22	0.85	1	1	1
56	0.91	0	0	1
61	0.92	0	0	1
107	0.92	.	.	0
106	0.95	.	.	0
160	0.99	.	.	1
28	1.05	1	0	0
149	1.06	.	.	1
99	1.08	1	0	.
156	1.1	.	.	1
24	1.19	0	1	1
25	1.19	0	1	0
57	1.28	1	0	0
145	1.28	.	.	0
109	1.3	.	.	1
79	1.32	0	0	.
17	1.34	1	0	1
67	1.34	0	0	.
135	1.35	.	.	0
105	1.36	.	.	0
157	1.38	.	.	0
128	1.42	.	.	0
103	1.45	0	0	.
33	1.46	0	1	0
94	1.48	0	1	.
111	1.49	.	.	0
76	1.52	0	1	.
59	1.66	1	0	0
151	1.72	.	.	0
116	1.74	.	.	0
92	1.99	1	0	.
81	2.21	0	0	.
113	2.56	.	.	0
9	2.57	0	0	0

Appendix C: The Preliminary Pool of Multiple-Choice Items. Correct answers are indicated in bold type. All page numbers refer to the course textbook (Campbell, 2010).

1. The intracellular calcium concentration [Ca^{2+}] of cardiac muscle cells is lower than the extracellular [Ca^{2+}]. This is important for heart function. To maintain this lower intracellular [Ca^{2+}], cardiac muscle cells rely on a $3Na^{+}/1Ca^{2+}$ antiporter. Given that the import of Na^{+} drives the export of Ca^{2+} ions, which of the following is true (HINT: Draw a diagram) (Page 137):
    f. **The antiporter transports Na^{+} down its concentration gradient**.
    g. Downregulating the sodium/potassium pump will not affect the $3Na^{+}/1Ca^{2+}$ antiporter.
    h. The antiporter utilizes ATP for energy.
    i. The antiporter maintains equal intracellular [Na^{+}] and [Ca^{2+}].
    j. The antiporter stops working when the intracellular [Na^{+}] drops below the intracellular [Ca^{2+}].

2. In the diagram below, GLUT2 is a facilitated diffusion transporter that transports glucose out of proximal tubule cells in the kidney into the extracellular medium. The glucose then moves from the extracellular medium into the capillary. Which of the following would increase the rate of GLUT2 activity (Pages 134-135):
    a. Increasing the fructose concentration gradient across the proximal tubule cell membrane.
    b. **Increasing the concentration of glucose in the proximal tubule kidney cells.**
    c. Increasing the rate at which kidney cells metabolize glucose.
    d. Shuttling more ATP to GLUT2.
    e. Decreasing the rate of glucose entry into the capillary.



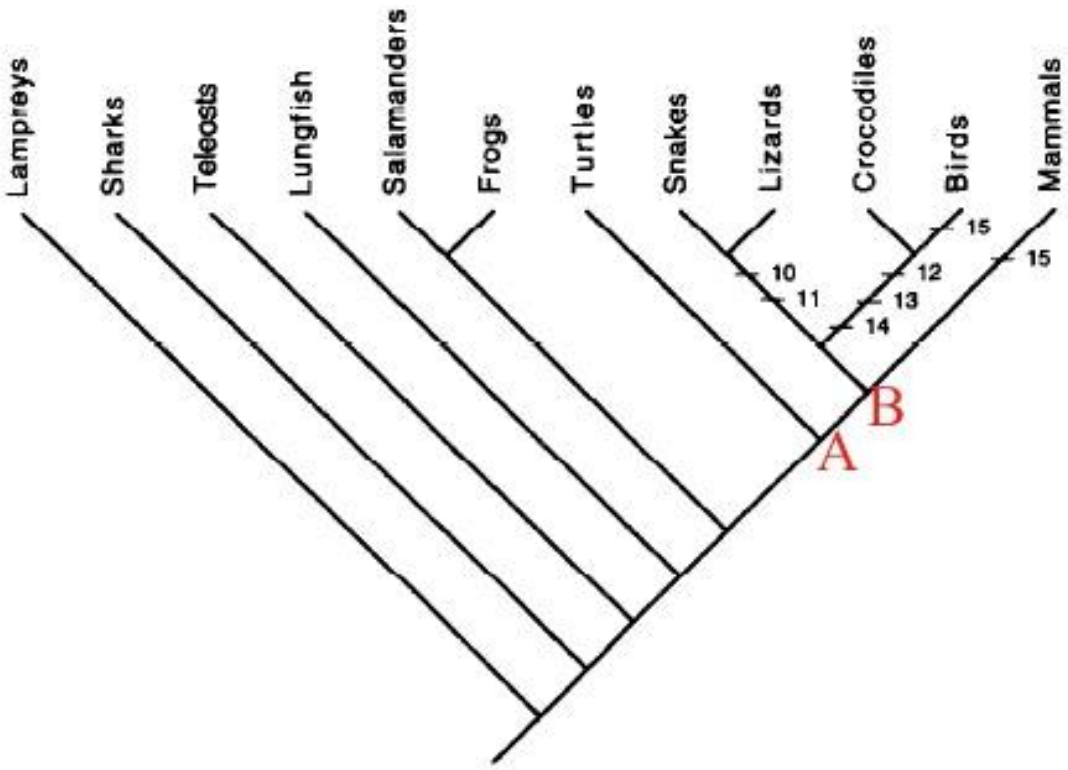
3. All of the following are advantages of having membrane enclosed structures EXCEPT (Pages 98 &100):
    a. Larger surface area for membrane bound biosynthetic enzymes.
    b. **Increased solubility of hydrophobic molecules in the cytosol.**
    c. Cell compartments with pH that is lower than the cytoplasmic pH.
    d. Separation of distinct biochemical pathways.
    e. Protection of the cell's DNA from destructive cytoplasmic enzymes.

4. Consider two cell membranes. Membrane A does not contain cholesterol (0%) while Membrane B contains 50% cholesterol. At a temperature of 20°C both membranes are in a gel state. Predict what will happen if the temperature is raised from 20°C to 36°C (Page 128):
    a. **Membrane A will transition to a liquid while membrane B will remain a gel.**
    b. Both membranes will transition to a liquid state.
    c. Membrane B will transition to a liquid while membrane A will remain a gel.
    d. Both membranes will retain their gel state.
    e. The proteins in membranes A and B will denature.

5. Which of the following would result from reducing the surface area to volume ratio (SA/V) of a cell (same shape) (Page 99):
    a. **The time it takes O_2 to diffuse from the cell surface to the mitochondria would increase**.
    b. The amount of glucose needed to fuel the cell would decrease.
    c. The amount of genomic DNA would increase.
    d. The surface area would increase at a faster rate than the volume.
    e. The cell membrane would become porous.

6. Which of the entities listed below has the greatest chance of being able to carry out both enzyme activity and replication (Pages 509-510):
    a. A protein enclosed in a membrane.
    b. A strand of RNA not enclosed in a membrane.
    c. A strand of DNA enclosed in a membrane.
    d. **A strand of RNA enclosed in a membrane.**
    e. A protein not enclosed in a membrane.

7. Two different species of animals have a homologous trait: long curved claws on their forelimbs. Based on this information you can infer that (Pages 540-541):
    a. The two species of animals are the closest living relatives.
    b. **The most recent common ancestor of the two species had long, curved claws.**
    c. The two species of animals use their long, curved claws for similar functions.
    d. The gene that confers long, curved claws is identical between the two species.
    e. The two species of animals also have long curved claws on their hindlimbs.



Please use the above figure for the questions 8 and 9. Figure modified from: (Morrison, 1996)

8. Birds and Mammals are the only two taxa in the cladogram shown above that maintain a stable internal body temperature (homeothermy). Based on this information you can infer that (Pages 538-540):
    a. Homeothermy is a homologous trait.
    b. **Homeothermy arose twice independently during evolutionary history.**
    c. Organism "B" was a homeotherm.
    d. Birds and Mammals are a monophyletic group.
    e. Crocodiles lost the ability to maintain a constant body temperature.

9. What does the node at point A in the above figure represent (Page 538)?
    a. **The point where the turtle ancestor diverged from the ancestor to snakes, lizards, crocodiles, birds and mammals**.
    b. The point where the turtle ancestor diverged from the ancestor to snakes.
    c. The point where the turtle ancestor diverged from the ancestor to mammals.
    d. The point where the turtle ancestor diverged from the ancestor to frogs and salamanders.
    e. The point where the turtle ancestor diverged from the ancestor to frogs.

10. In nature, the bacteria *Agrobacterium tumefaciens* transforms the plant *Arabidopsis thaliana* by injecting its DNA into the plant's ovules. The bacterial DNA is integrated into the chromosomal DNA of the megaspore. This does not affect the ability of the plant to continue its life cycle. The bacterial genes are undergoing (Page 756):
    a. **Horizontal gene transfer followed by vertical gene transfer.**
    b. Horizontal gene transfer only.
    c. Vertical gene transfer only.
    d. Vertical gene transfer followed by horizontal gene transfer.
    e. Endosymbiosis.

11. _____ produces gametes (Pages 611, 624, 639, 643, & 802-803)
    a. Cleavage
    b. Fertilization
    c. Karyogamy
    d. Plasmogamy
    e. **None of the above**
12. A herbicide that kills germinating seeds most likely blocks the process(es) of (Page 624):
    a. **Mitosis**
    b. Fertilization
    c. Meiosis
    d. Gametogenesis
    e. Peptidoglycan formation

13. You encounter a new organism that has the following characteristics: it is photoautotropic, it shows alternation of generations, it lacks vascular tissue, and it soaks up water through its surface. What kind of organism could this be (Pages 607, 611, & 802)?
    a. A grass
    b. A cyanobacterium
    c. **A moss**
    d. A fern
    e. A cyanobacterium or a moss

14. A seed is planted 10 inches below the soil surface. At the moment of germination it carries out all of the following processes EXCEPT (Pages 821-822, & 824):
    a. **Photosynthesis**
    b. Cell Division
    c. Shoot Elongation
    d. Gravitropism
    e. Cell Expansion

15. Multiple Sclerosis (MS) is characterized by inflammation and loss of myelination of the neurons of the Central Nervous System. Which of the following do you predict would occur in patients with MS (Pages 1054 &1066):

    a. Disruption of impulses along the somatic motor neurons.
    b. Disruption of impulses along neurons of the medulla.
    c. Spinal cord neurons are more susceptible to damage than sensory neurons.
    d. Patients with MS do not respond to anti-inflammatory drugs.
    e. **Both B and C**.

16. α-dendrotoxin (DTX) is a neurotoxin found in puffer fish. When applied to a neuron preparation in the lab, DTX increases the frequency of action potentials. Which of the following would explain this (Page 1052)?
    a. DTX makes the threshold potential less negative.
    b. DTX lengthens the time it takes for the voltage gated K^+ channels to close.
    c. DTX lengthens the time it takes for the voltage gated Na^+ channels to open.
    d. **DTX decreases the magnitude of the undershoot.**
    e. DTX makes the resting membrane potential more negative.

17. Rigor mortis (stiffness of death) is believed to result from the depletion of ATP in skeletal muscle cells. Lack of ATP in skeletal muscle cells after death would result in (Page 1107):
    a. **A net flow of Ca^{2+} from the sarcoplasmic reticulum into the cytoplasm of the skeletal muscle cell.**
    b. Tropomyosin blockage of myosin binding sites.
    c. The detachment of actin from myosin.
    d. A net flow of Na^+ from the cytoplasm of the skeletal muscle cell to the extracellular space.
    e. An increase in the rate of glucose metabolism.

18. The resting potential of a neuron is measured in the lab by a voltmeter. Which of the following would cause the voltmeter to register a membrane potential more negative than -70mV (Page 1048 & 1050)?
    a. Reversing the positions of the reference and measurement electrodes.
    b. Adding a chemical that opens Na^+ channels.
    c. **Adding a chemical that opens K^+ channels.**
    d. Adding a chemical that blocks Cl^- channels.
    e. Adding glucose to the extracellular fluid.

19. Another way to state that the resting potential of a neuron is -70mV is (Page 1048):
    a. **The inside of the neuron is 70mV more negative than the extracellular environment.**
    b. The amount of negative charges in the neuron is greater than the amount of positive charges in the neuron by a factor of 70mV.
    c. Ions have ceased to cross the membrane and the charge difference across the membrane holds constant at 70mV.
    d. There are 70mV more positive charges inside the neuron than in the extracellular environment.
    e. At rest the neuron is only permeable to negative charges and there is a net flow of negative charges into the neuron.

20. Excitation of muscle fibers during exercise results in action potentials. The K^+ that leaves the cell during the repolarization phase of the action potential either diffuses into the capillaries or is reclaimed by the skeletal muscle fibers. Which of the following is an adaptation to exercise that can prevent hyperkalemia (high levels of potassium in the blood) during prolonged periods of exercise:
    a. Increasing the number of K^+ leak channels in skeletal muscle fibers.
    b. Increasing the number of Na^+ leak channels in skeletal muscle fibers.
    c. **Increasing the number of Na^+/K^+ pumps in skeletal muscle fibers.**
    d. Increasing the number of Ca^{2+} ions released per action potential.
    e. Increasing the intestinal absorption of K^+.

21. All of the following processes are required to transmit a signal across a chemical synapse EXCEPT (Pages 1055-1056):
    a. **Endocytosis**
    b. Diffusion
    c. Exocytosis
    d. Facilitated Diffusion
    e. All of the above are required

22. Which of the following is true when cardiac sarcomeres are in a contracted state (Pages 903 & 1104):
    a. Actin and myosin are not crosslinked.
    b. Myosin is bound to ATP.
    c. **Ca^{2+} is bound to troponin.**
    d. Actin is bound to ADP.
    e. This region is in diastole.

23. Which of the following pieces of evidence BEST supports the hypothesis that the Monoamine Oxidase (MAO) cleaves norepinephrine (NE) to attenuate NE signaling in the autonomic nervous system:
    a. Smooth muscle cells produce MAO.
    b. Norepinephrine is classified as a monoamine.
    c. Pharmacologic inhibitors of MAO do not increase the frequency of skeletal muscle contractions.
    d. Age related changes in MAO activity correlate with cognitive defects.
    e. **Mice deficient in MAO show increased levels of norepinephrine in the brainstem.**

24. The image below is an electron micrograph of damaged skeletal muscle and the red box surrounds the site of damage. What part of the muscle has been damaged (Pages 1104-1105)?

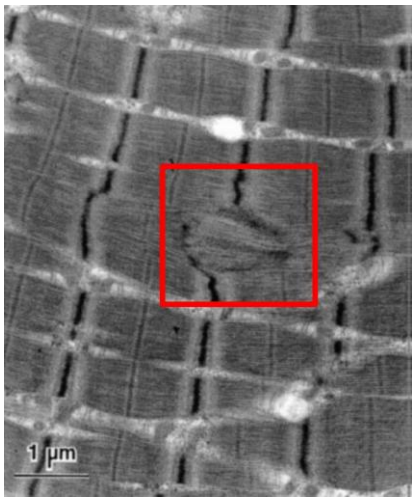


Figure modified from: (Roth et al., 2000)

    a. **The line between sarcomeres.**
    b. The line through the center of the sarcomere.
    c. The muscle fiber.
    d. The T-tubule.
    e. The sarcoplasmic reticulum.

25. Strength training can cause all of the following to increase EXCEPT:
    a. The number of recruited motor units.
    b. The number of myofibrils in a muscle fiber.
    c. **The rate of cell division in muscle fibers.**
    d. The amount of actin and myosin in a muscle fiber.
    e. The size of muscle fibers.

26. If a person's pulse is 84 beats per minute, how many times per minute does their SA node depolarize (Page 904):
    a. 42
    b. 21
    c. 168
    d. **84**
    e. 336

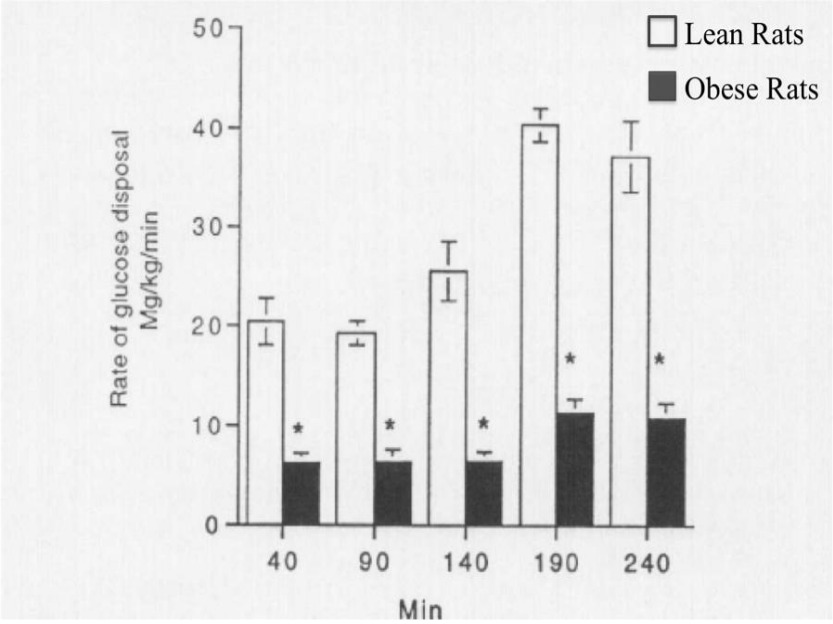
27. Cardiac fibrosis is marked by large collagen deposits between cardiac cells and is commonly seen in patients with chronic heart failure. Packing collagen between cardiac cells can result in:
    a. **Decreased ion flow between cardiac cells.**
    b. Decreased heart size.
    c. Misalignment of cardiac sarcomeres.
    d. Mixing of oxygenated and deoxygenated blood.
    e. All of the above.

28. In a double circulatory system, the pressure that sends the blood to the systemic circuit is generated during _____ by the _____ (Pages 903-904):
    a. Diastole, Left Ventricle
    b. Systole, Right Atrium
    c. Diastole, Right Ventricle
    d. **Systole, Left Ventricle**
    e. Systole, Left and Right Atria

29. Which of the following would increase the rate of glucose diffusion from the capillaries to the surrounding cells (Pages 905-906):
    a. Decreasing the amount of glucose in the blood.
    b. **Decreasing the flow velocity of blood through the capillaries.**
    c. Increasing the flow velocity of blood though the veins.
    d. Decreasing the cross sectional area of the capillaries.
    e. Increasing the number of glucose active transporters in the capillary membrane.

30. Fick's Law has been applied to gas exchange and insulin exchange. As it applies to insulin exchange between the plasma and the muscle interstitium: $Q = PS\ (C_p - C_I)$. In this equation Q = rate of insulin exchange, P = the permeability of the surface to insulin, S = the surface area for exchange, C_P = plasma insulin concentration, C_I = interstitium insulin concentration. Increasing the number of capillaries recruited to muscle tissue would enhance insulin delivery to muscle by:
    a. Increasing P
    b. **Increasing S**
    c. Increasing C_P
    d. Decreasing C_I
    e. Decreasing PS

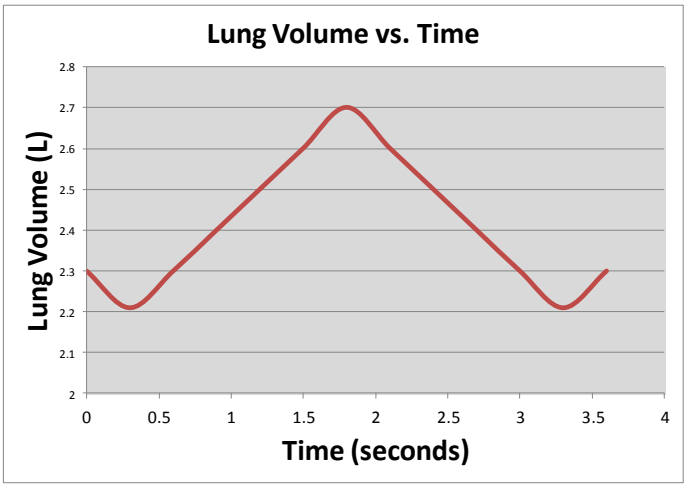
31. The graph below shows the rate of glucose uptake by tissues in lean (unfilled bars) and obese (filled bars) rats after being given equal amounts of glucose. The y-axis shows how fast glucose is taken up by the cells. Based on the graph, which of the following is true:



Time in Minutes post Glucose dose
Figure modified from: (Holmang, Mimura, Bjorntorp, & Lonnroth, 1997)

a. The lean rats have higher blood sugar values at the 140 minute mark than at the 90 minute mark.
b. The obese rats have the same blood sugar values at the 40 minute and 90 minute marks.
c. The obese rats have higher blood sugar values than the lean rats at all time points.
d. The lean rats were more active at the 40 minute mark than at the 190 minute mark.
e. The obese rats had higher glucose uptakes than the lean rats.

32. The graph below describes the typical breathing pattern for an adult male at rest. At which time point in the graph would you expect the sarcomeres of the thoracic diaphragm to be contracting (Pages 918-919):



a. 0 seconds
b. **1 second**
c. 2.5 seconds
d. 0.2 seconds and 3.2 seconds
e. At rest sarcomeres will not actively contract

33. Based on the graph below, which hormone do you predict is responsible for the trend in blood sugar levels that begins at the 10 minute time point (Pages 893, 982, & 986)?

a. Insulin
b. Aldosterone
c. Antidiuretic hormone
d. **Glucagon**
e. Parathyroid Hormone

34. The kidneys of healthy individuals secrete Erythropoietin. However, individuals with severe kidney disease are unable to produce Erythropoietin. Which of the following do you predict is/are associated with severe kidney disease (Page 913)?
    a. Low amounts of Red Blood Cells.
    b. Increased chance of blood clots due to abnormally viscous blood.
    c. Poor Oxygen delivery to tissues.
    d. A right-shift of the Oxygen binding to hemoglobin curve.
    e. **Both A and C.**

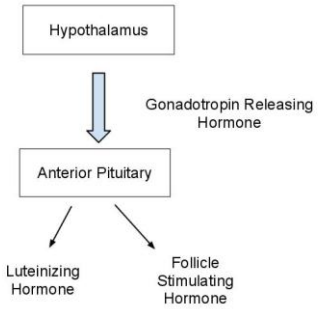
35. After being filtered into the filtrate, Drug XYZ is both poorly reabsorbed by the kidneys. Based on this information, which of the following is most likely true about Drug XYZ (Pages 960-961):
    a. The kidneys are inefficient at clearing Drug XYZ from the blood.
    b. **Individuals taking Drug XYZ excrete large amounts of the drug in their urine.**
    c. The molecular size of Drug XYZ is larger than a Red Blood Cell.
    d. A decrease in blood pressure will increase the rate at which the kidneys filter Drug XYZ from the blood.
    e. Secretion of renin will lower the rate at which the kidneys filter Drug XYZ from the blood.

36. A new animal previously unknown to humans was just discovered! The identity of this animal is kept secret, but you hear about some lab results. You learn that the animal has blood values of 638 milli-osmoles/L and produces urine that is 17564 milli-osmoles/L. Of the choices listed below, this animal is most likely a previously unknown type of (Pages 966-967):
    a. Snake
    b. Marine fish
    c. **Cat**
    d. Frog
    e. Freshwater fish

37. Joe woke up in the morning and ate a full breakfast of salted bacon, an orange, and toast. Which combination of hormones do you predict is taking effect in Joe after this breakfast (Pages 970, 986, & 989-990):
    a. Renin, angiotensisn, insulin
    b. Glucagon, ADH, Calcitonin
    c. Insulin, Calcitonin, Renin
    d. Glucagon, PTH, angiotensin
    e. **PTH, Insulin, ADH**

38. It has been observed that secretion of progesterone by the corpus luteum causes a woman's body temperature to rise by 0.5°F. If the woman does not get pregnant, the body temperature drops by 0.5°F around the time of menstruation. Based on this information, at what point during a woman's 28 day cycle would you predict this increase in temperature occurs (Pages 1008-1009):
    a. Day 7
    b. Day 10
    c. **Day 16**
    d. Day 22
    e. Day 28

39. Consider the following sequence of events. This sequence of events begins at (Pages 1009-1010):



a. Birth in females only
b. Birth in males and females
c. Puberty in males only
d. Puberty in females only
e. Puberty in males and females

40. The tiny Corkus organism lives on the skin of large, hairy Spudnus and causes the Spudnus to develop itchy boils. Which of the following would make this interaction an example of mutualism (Page 1199):
    a. The Spudnus provides the Corkus with warmth.
    b. The Corkus feeds off the Spudnus' secretions.
    c. The hair of the Spudnus protects the Corkus from UV Rays.
    d. All of the above
    e. None of the above

41. The anterior pituitary of a female with hypogonadism secretes abnormally low levels of LH. Insufficient levels of LH can lead to (Pages 1008-1009):
    a. Developlent of multiple follicles at a time
    b. **Failure to ovulate**
    c. Increased endometrial development
    d. Increased fertility
    e. Abnormally high levels of Inhibin

Appendix D: Reasoning and Logic Behind the Items in the Preliminary Pool.

1. This question requires students to sort through the various mechanisms of transport. The stem informs students that the import of Na^+ drives the export of Ca^{2+}; hence, students need to recognize that this is an example of cotransport. Students then need to apply their knowledge of cotransport to sort through the distractors. In order for the import of Na^+ to power the export of Ca^{2+}, Na^+ must be moving down its concentration gradient. Therefore answer a is correct and answer c is incorrect. Choices e and d are for students with the misconception that the cotransport is affected by the difference between the sodium and calcium concentrations. Choice b is a trap for students who do not understand that the sodium/potassium pump is responsible for maintaining a higher extracellular sodium concentration, which is necessary for the three sodium/one calcium antiporter.

2. Question 2 asks students to consider the movement of glucose from the kidney cells to the blood and to apply their knowledge of facilitated diffusion. Students need to remember that facilitated diffusions does not require energy (choice d), facilitated diffusion channels are specific for their substrate (choice a), and that facilitated diffusion is driven by concentration gradients. Increasing the concentration of glucose in the proximal tubule cell will increase the glucose concentration gradient across the membrane and will therefore increase GLUT2 activity (choice b—correct answer). Choices e and c will decrease the glucose concentration gradient and are therefore incorrect. Choice a is for students with the misconception that facilitated diffusion channels are not specific for a substrate.

3. This question tests students' understanding of membranes. Membranes allow for separation of aqueous compartments. Since they are separated, the various aqueous compartments can have distinct environments. Therefore, choices c and d are true. Membranes also serve as scaffolds for biochemical reactions and provide protection for the contents they surround—hence, choices a and e are true. However, membranes cannot make a hydrophobic molecule dissolve in an aqueous medium and choice b is therefore incorrect (and the correct answer to this question).

4. Students need to apply their knowledge of membrane fluidity to answer this question. Cholesterol stabilizes the membrane. That is to say that at higher temperatures cholesterol keeps the membrane from becoming too fluid while at lower temperatures cholesterol prevents the membrane from becoming too solid. Without cholesterol the 0% membrane will transition to a fluid state upon being heated. Conversely, since it is packed with cholesterol, the 50% membrane will remain a gel. Therefore choice a is the correct answer. While proteins do denature at high temperatures, students need to realize that 36° C is not high enough to cause the proteins in the membrane to denature. Since 36° C is the normal temperature for a human, the membrane proteins do not denature at this temperature and choice e is incorrect.

5. Students need to realize that a decrease in the surface area to volume (SA/V) ratio of a growing cell means that the increase in volume is greater than the increase in surface area. A larger cell needs more energy to live—hence b is incorrect. And since the cell is larger, the O_2 has to travel a greater distance to reach the mitochondria so choice a is correct. Choice c is a review from the perquisite course—the size of the genome does not increase as the cell grows. Choice d checks to see if students have a basic understanding of the SA/V ratio. Choice e is just

a distractor that seems plausible but is not true (students should use what they have learned about cell elongation and the cell cycle to know that cells grow without developing pores in their membranes).

6. Answering this question requires students to apply their knowledge of RNA, DNA, and membranes. Molecules that are surrounded by a membrane are more protected than naked molecules. Therefore, choices b and e are incorrect. Proteins can carry out metabolism but are not self replicating so choice a is incorrect. DNA is replicated however, it is incapable of enzymatic activity so choice c is incorrect. RNAs can have catalytic activity so choice d is correct.

7. This item requires students to apply the definition of a homologous trait. The long curved claws seen in the two species are the result of their common ancestry. Therefore, the trait must have been present in the most recent common ancestor to the two species and choice b is correct. While the two species share a common ancestry, they are not necessarily the closest relatives of each other—hence, choice a is incorrect. As the two species diverged from each other the gene that encodes this homologous trait changed over time so choice d is incorrect. Also, the homology of their claws does not imply that the two species of animals use their claws for the same purpose. This rules out choice c. Choice e is a total distractor because having claws on forelimbs does not mean that the animals have to have claws on their hindlimbs.

8. This question asks students to analyze a cladogram. Birds and mammals are the only homeotherms and yet they are not sister taxa. The observation that homeothermy is not present in any of the ancestors to birds and mammals nor is it not present in crocodiles indicates that homeothermy is not a homologous trait (choices a and e). Therefore, homeothermy arouse

twice independently during evolutionary history and choice b is correct. Birds and mammals are a polyphyletic group—not a monophyletic group (choice d).

9. In this question students must continue to analyze the cladogram. It is clear that turtles diverged at point A. However, students need to figure out what the turtles diverged from. Even though turtles are the extant taxa, students should not just read across the tips. Rather, students need to analyze the branching patters to see that at point A the ancestor to turtles diverged from the ancestor to snakes, crocodiles, lizards, birds, and mammals. Therefore, choice a is the correct answer.

10. The goal of this question is to get students to distinguish between vertical and horizontal gene transfer. *Agro* transformation of *Arabidopsis* is horizontal gene transfer because the movement of DNA is not generational. Vertical gene transfer occurs when the plant passes the bacterial genes to future plant generations (choice a is correct). Students need to correctly apply these concepts to arrive at the correct answer. Choice e is an unrelated distractor.

11. Students have probably seen the stem of this question before; however, they are presented with an unexpected set of answer choices. Mitosis and meiosis can generate gametes and yet they are not among the choices. Students need to consider each process and realize that it does produce gametes. Therefore the answer is choice e.

12. The question requires students to understand germinating seeds are undergoing mitosis; therefore, blocking mitosis will kill germinating seeds. Germinating seedlings are not undergoing fertilization, meiosis, or gametogenesis so choices b, c, and d are incorrect. Choice e is an unrelated distractor because bacterial cell walls—not plant cell walls—contain peptidoglycan.

13. In this question students need to compare and contrast the characteristics of the new organism with the answer choices. Since the organism shows alternation of generations answer choices b and e must be incorrect. Choices a and d are incorrect because the organism in the stem lacks vascular tissue. A moss is the only given organism that can have all those characteristics (choice c).

14. This questions tests students understanding of photosynthesis and seeds. Students are told that photosynthesis requires light so they need to apply that knowledge to realize that without light, photosynthesis will not occur. Also, students need to understand that the seed contains enough nutrition to support a developing seed until it can carry out photosynthesis. A simplistic "plants need energy so they must be doing photosynthesis" will lead to an incorrect answer.

15. In this question students must differentiate between the central nervous system (CNS) and the peripheral nervous system (PNS). Students must also make inferences about the consequences of demyelination and of treating inflammation in MS patients. Choice a is incorrect because it refers to the PNS which is unaffected in MS. Choice b and c are correct because myelin serves to protect neurons and speed up nerve impulses by insulating the neurons. Therefore, the best answer is choice e. Students also need to infer that if inflammation is one of the causes of MS, then reducing the inflammation with anti-inflammatory agents will lessen the symptoms (as is the case). Therefore, choice d is incorrect.

16. To solve this item, students need to consider each potential mechanism of α-dendrotoxin action and infer the impact it would have on the frequency of action potentials. Choice d is the only mechanism that would increase the frequency of action potentials. Decreasing the magnitude of the undershoot lessens the length of time it takes for the neuron to

return to resting potential and less time between action potentials allows for an increased frequency of action potentials. All other mechanisms decrease the frequency of action potentials because: raising the threshold for an action potential makes it harder for an action potential to occur (choice a), slowing the time it takes for the K^+ channels to close would extend the time in between action potentials (choice b), slowing the opening of the voltage gated Na^+ delays the progression of an action potential (choice c), and decreasing the resting potential of the membrane increases the amount of depolarization needed to stimulate an action potential (choice e).

17. To answer this item students need to apply their knowledge of muscle contraction and make inferences. Muscle cells use ATP to maintain a lower intracellular calcium concentration by pumping calcium into the sarcoplasmic reticulum and out of the cell. Post-mortem depletion of ATP therefore results in a net flow of calcium from the sarcoplasmic reticulum into the cytoplasm of the skeletal muscle cells and choice a is correct. Upon entering the muscle cell calcium binds to troponin, tropomyosin shifts its conformation so that the actin can bind to myosin. Choices b and c are for students who do not understand this cascade. Choice d is incorrect because it implies that the intracellular [Na^+] is greater than the extracellular [Na^+], which is not true. Eliminating choice e requires students to realize that the muscles of a dead person will not increase their rate of glucose metabolism.

18. In this question students need to apply their knowledge of what a voltmeter is measuring as well as the intracellular and extracellular concentrations of the various ions. The question is essentially asking students to identify what would increase the charge difference across the membrane. The charge difference can be increased by moving positive ions from the neuron into the extracellular fluid or by moving negative ions from the extracellular fluid into the

neuron. Opening the potassium channels would cause a net flow of K^+ ions to diffuse out of the neuron and is the correct answer. Opening the Na^+ channels would cause positively charged sodium to enter the neuron, opening the Cl^- channels would cause chloride to exit the neuron—both of these would lessen the charge difference across the membrane (choices b and c). Adding glucose to the fluid would not affect the charge difference (choice e) and reversing the position of the electrodes would cause the voltmeter to register a potential of +70mV (choice a).

19. This question asks students to apply the definition of resting potential as well as their knowledge of neurons. In this term "potential" refers to voltage difference across a membrane; therefore, choice b is incorrect. Choices c and e are incorrect because even though the resting potential remains relatively constant, positive and negative ions are always being transported across the membrane. Choice d has the signs reversed and implies that the units of mV are a count of charges. Choice a is correct and students who understand the definition of resting potential may not even need to sort through the distractors. By convention, the membrane potential reflects the interior of the cell relative to the exterior of the cell. So a resting potential of -70mV means that the interior of the cell is -70mV more negative than the exterior environment.

20. In this question students need to infer which of the answer choices would increase the rate at which skeletal muscles reclaim K^+. Once again, students need to apply their knowledge of concentration gradients. The Na^+/K^+ pump imports K^+ into skeletal muscle fibers and exports Na^+. Therefore upregulation of this pump would increase the rate of K^+ entry into skeletal muscle fibers and choice c is correct. Even though it contains the terms "upregulation" and "K^+", choice a is incorrect. Upregulating the K^+ leak channels would make the situation worse by facilitating the efflux of K^+ of potassium from skeletal muscle fibers. Upregulating the

sodium leak channels would not help skeletal muscle fibers reclaim K^+ (choice b). Also, increasing intestinal absorption of K^+ would increase blood levels of potassium. In another part of the course students learned that nutrients are absorbed from the digestive track into the blood so they should be able to reason through distractor e.

21. In this question students need to consider signal transmission across a chemical synapse. Students who remember the steps in this process should be able to identify the processes of diffusion, facilitated diffusion, and exocytosis. Endocytosis is often part of signal attenuation; however, this question does not ask students to continue the chain of events that lead to attenuation. Therefore, the correct answer is choice a.

22. To answer this item students need to combine their knowledge of sarcomere contraction and the heartbeat. When the sarcomeres are contracted actin and myosin are crosslinked (a is incorrect), myosin is bound to ADP (b is incorrect), Ca^{2+} is bound to troponin (c is correct), and actin is not bound to ADP (d is incorrect). Also, the cardiac sarcomeres contract during the systolic phase of the heartbeat so choice e is incorrect.

23. This question asks students to decide which piece of data best supports a cause and effect relationship between Monamone Oxidase (MAO) and Norepinephrine (NE) in the autonomic nervous system (ANS). The ANS does innervate smooth muscle cells and smooth muscle cells do produce NE but choice a only represents "guilt by association" option. The classification of NE as a monoamine does not give any information about whether or not it is cleaved by MAO in the ANS (choice b). Choice c is incorrect because the nerves that stimulate the skeletal muscle contractions are not part of the ANS. Choice d is incorrect because correlation does not imply causation. Choice e is correct because the data given support the

hypothesis that MAO cleaves NE in the ANS (the textbook figure shows that some of the nerves from the ANS originate in the brainstem).

24. In this question students need to transfer what they learned from cartoon sarcomeres to human skeletal muscle sarcomeres. By recognizing that they are looking at sarcomeres, students can immediately eliminate choices c, d, and e. The image is too "zoomed in" for the answer to be any of those three options. Admittedly, if students can properly orient themselves to the image they have a 50/50 chance at getting the correct answer. Students who can transfer the cartoon sarcomere will realize that the thickest lines in the image are the lines between sarcomeres and choice e is correct.

25. In class students learned that skeletal muscle cells are syncytial and form through cell fusion events rather than through cell division. Applying this knowledge will allow students to recognize that strength training cannot increase the rate of cell division in muscle fibers (choice c is the answer). Students can also arrive at the correct answer by eliminating the distractors. All students should be able to eliminate e. Students who read the text can use recall to eliminate choice a. Students who listened to the lecture can eliminate choices b and d.

26. The textbook presents students with the EKG pattern, a corresponding diagram of the heart that highlights the stimulated region(s), and walks students though the cardiac cycle. However, the text does not link cardiac cycle to the commonly measured pulse. Therefore, this question requires students to reason that it is the SA node that starts each heartbeat and if a person's pulse is the number of heartbeats per minute, then the pulse corresponds to the number of times per minute that the SA node depolarizes and choice d is correct.

27. In this item students are asked to consider the consequences of packing collagen between cardiac cells (cardiac fibrosis). The only way for students to solve this item is to place

each answer option in the context of what they learned about the heart. Students learned that cardiac cells are connected by gap junctions and ion flow through the gap junctions allows for the spread of impulses. Students are expected to reason that packing collagen in between cardiac cells will disrupt the ion flow between the cells (choice a is correct). Students should also recognize that adding collagen to the heart will not make it smaller (choice b). In class students learned that, unlike skeletal muscle sarcomeres, cardiac sarcomeres are not aligned so choice c does not represent a pathological consequence of cardiac fibrosis. Lastly, choice d is incorrect because the mixing of oxygenated and deoxygenated blood would imply that the structural integrity of the heart was compromised—an unlikely consequence of depositing collagen in the heart. Since choices b, c, and d are incorrect, the answer cannot be all of the above (choice e).

28. This question asks students to place a function of the heart in the context of the heartbeat and the heart anatomy. The contracting of the left ventricle during systole pressurizes the blood sent to the systemic circuit so the answer is choice d. Students can arrive at this answer by thinking through the steps of the heartbeat, by first deciding whether it is systole or diastole, or by another algorithm. The distractors represent the various terms and concepts that students could be confused on.

29. This question deals with diffusion but does not focus extensively on concentration gradients. Students need to consider each distractor and determine if it would increase the rate of glucose diffusion out of the capillaries. The stem of the question implies that the concentration of glucose is higher in the blood than in the surrounding tissues. Therefore, choice a is incorrect because lowering the amount of glucose in the blood would decrease the concentration gradient across the blood. Choice b is correct because slowing the flow velocity of blood through the capillaries would give glucose more time to exit the capillary. Diffusion is a slow process so

extending the time for diffusion to occur would help with glucose disposal to tissues. Choice c is incorrect because the flow velocity through the veins is an after the fact issue. By the time the blood reaches the veins it has already passed through the capillary beds. Choice d is incorrect because it would result in a decrease in the surface area for diffusion. Active transport works against diffusion so students who select option e are completely confused or did not read the answer carefully.

30. In class students learned about of Fick's law of diffusion in the context of gas exchange in the lungs where Q is the volume of gas that diffuses per unit time. Under Fick's law: Q = (Area for diffusion/tissue thickness)(diffusion constant)(difference in pressure across the tissue). This question asks students to transfer Fick's law to the case of insulin flux across the capillary membrane. The question asks students to figure out why increasing the number of capillaries recruited to muscle tissue increases Q. Students may immediately recognize that the capillaries represent the surface area for diffusion. Therefore, recruiting more capillaries to the muscle increases S (choice b). Students may also understand that recruiting more capillaries to the muscle does not affect the capillary permeability (choice a) nor does it have an immediate effect on the insulin concentrations on either side of the membrane (Choices c and d).

31. This question presents students with a piece of actual data to interpret. The hard part about this graph is that students need to understand that the y-axis represents the rate of glucose disposal (how fast the tissues are uptaking glucose from the blood). The graph compares the rates of glucose disposal in lean and obese rats. If students understand the y-axis they can immediately eliminate choice e because the unfilled bars are taller than the filled bars. Thus, the lean rats had higher rates of glucose disposal. Once students understand that the y-axis is the rate of glucose disposal the next step is to determine the effect of glucose disposal on blood glucose

levels. Removing glucose from the blood serves to lower blood glucose levels. So if the obese rats had lower rates of glucose disposal at all time points than the lean rats, then the obese rats should have higher blood glucose levels than the lean ratsat time points and choice c is correct and choice a is incorrect. Choice b is incorrect because a constant rate of glucose disposal means that blood glucose levels are steadily declining. Choice b was written for students who think that the y-axis represents blood glucose levels. Students should be able to immediately eliminate choice d because the information on the graph makes no mention of exercise. The students who chose to comb through distractor e should recognize that if exercise was at all a factor, it would increase glucose disposal so the rats would have been more active at the 190 minute time point not the 40 minute time point.

32. In this item students are asked to map their knowledge of breathing onto a graph. Students need to first recall that when the sarcomeres of the diaphram contract, the diaphram flattens out and the lungs fill with air. As it pertains to the graph, if the lungs are filling with air then the lung volume is increasing and the answer is choice b. The important thing for students to recognize about the graph is that it is not the absolute lung volume that is important, rather it is the trend. The portion of the graph with the positive slope represents inhalation while the portion of the graph with the negative slope represents exhalation. Choice e was inspired by students' misconceptions to a question I asked last semester. I asked students to view a video of actin/myosin dynamics and tell me if the myosin head was in the high or low energy conformation at the end of the video. Too many students answered the question by writing: the myosin head is in the low energy configuration because it is not moving.

33. This is another item that requires students to interpret a graph. The stem of the question directs students to the trend that starts 10 minutes after the start of the experiment.

Students are expected to interpret the trend as in increase in blood glucose levels. Students who correctly interpret the graph then need to remember that Glucagon (choice d) has the effect of raising blood glucose levels. All other choices represent hormones that students learned about in class but do not act to raise blood glucose.

34. In class students learned about the role of Erythropoietin (EPO) in stimulating the production of red blood cells (RBCs). This question asks students to make inferences about the consequences of insufficient levels of EPO. Students need to reason that since EPO stimulates red blood cell development, lack of EPO will result in low RBC counts (choice a). To obtain the correct answer students need to infer that a sequela of reduced RBC counts is poor oxygen delivery to tissues (choice c). Thus the correct answer to this question is choice e. Choice b is incorrect because as it is a consequence of high RBC counts (students learned about the dangers of high RBC counts in class). Choice d is incorrect because low RBC counts would not affect the overall oxygen-hemoglobin binding dynamics (in class students learned about what would impact these dynamics).

35. This question asks students to apply their knowledge of the mechanisms by which the kidneys filter the blood and produce urine. The stem of the question informs students that the drug enters the filtrate and is not reabsorbed by the kidneys. To obtain the correct answer, students must infer that the drug moves from the blood remains in the urine. Therefore the answer is choice b. The stem of the question gives students the information they need to eliminate choice a. Since the drug enters and stays in the urine, it is efficiently removed from the blood. Choice c is incorrect because if the drug was larger than a Red Blood Cell it would not be able to enter the filtrate. Choice d is incorrect because it is the blood pressure that forces the movement of fluid and solutes from the blood into the filtrate. A decrease in blood pressure

would reduce the rate at which the kidneys filter out the drug from the blood. Choice e is incorrect because the downstream effect of rennin is an increase in blood pressure.

36. To answer this question, students need to recognize that they are viewing values of urine osmolality (something they learned about in class). Students then need to interpret the numbers given to infer that the organism's urine is more concentrated than its blood. Of the animals listed, only the mammal (cat—choice c) is capable of producing hyperosmotic urine.

37. Even though students learn about hormones one by one, it is important for them to understand that at any given moment there are a multitude of hormones flowing through a human. Students need to pick apart Joe's breakfast and deduce the effect it would have on his hormone status. Overall, the salt in Joe's breakfast would stimulate the secretion of ADH, the sugar in Joe's breakfast would stimulate the release of insulin, and the lack of calcium would stimulate the release of PTH. Thus the correct answer is choice e.

38. In this question students need to map the ovarian and hormone cycles onto the standard 28 day menstrual cycle. The corpus luteum forms after the follicle has released its egg at day 14 and begins to secrete progesterone at day 16. Therefore, choice c is correct and choices a and b must be incorrect. Choices d and e are incorrect because the corpus luteum begins to secrete progesterone before day 22.

39. This diagram is presented to students in the context of the male and female hormonal cascades. This question attempts to be an out of context application of the concept. Rather than thinking about the targets of LH and FSH in males and females, students need to place this diagram in the context of the human life cycle. Students need to apply their knowledge that this diagram describes, in part, the male and female reproductive cycles to infer that the sequence of events does not begin until puberty in males and females (choice e). Choices c and d select for

students who do not remember that this sequence of events is present in males and females. Choices a and b select for students who do not understand that these events initiate human reproductive capabilities.

40. Admittedly, this is a tricky question that asks students to apply the definition of mutualism. The stem of the question informs students that the Corkus causes itchy boils on the Spudnus. Since the Spudnus is harmed by the Corkus, this cannot be an example of mutualism. Therefore, the answer is choice e. All other choices are incorrect because they imply that this is a case of mutualism.

41. This question asks students to recall the role of LH in the female reproductive cycle and infer the consequences of low levels of LH. The surge in LH around day 14 of the female reproductive cycle triggers ovulation. Low levels of LH therefore result in a failure to ovulate (Choice b). Also, since a surge in LH triggers ovulation, low levels of LH would not lead to increased fertility (choice d). Choice e is incorrect because low levels of LH would reduce the amount of Inhibin produced. Choice a is incorrect because FSH regulates follicle development, not LH. Choice c is incorrect because low LH would lead to low estrogen and reduced endometrial development.

Appendix E: Item parameters, fit statistics, and estimates for the 2012 exam. The *b* values are the item parameters, the infit and outfit mean square values are the fit statistics, and the discrimination and asymptote values are the item estimates. Items 86 – 109, sans 97, were the 23 validated critical thinking items.

Item	*b* value	Error	Infit Mean Square	Outfit Mean Square	Discrimination	Lower Asymptote	Upper Asymptote
1	0.28	0.11	0.987	1.002	1.043	0.000	0.988
2	-0.36	0.12	1.094	1.118	0.766	0.366	0.948
3	0.28	0.11	0.922	0.905	1.397	0.000	1.000
4	-0.92	0.13	0.943	0.901	1.081	0.000	1.000
5	-0.4	0.12	0.950	0.924	1.122	0.000	1.000
6	-0.58	0.12	1.005	1.021	0.986	0.000	0.995
7	-1.04	0.14	1.000	0.912	1.020	0.657	1.000
8	2.53	0.16	1.090	1.224	0.898	0.021	1.000
9	-1.26	0.15	0.966	0.870	1.051	0.000	1.000
10	-1.13	0.14	0.926	0.871	1.090	0.000	1.000
11	-0.9	0.13	0.913	0.850	1.130	0.000	1.000
12	0.16	0.11	0.989	0.972	1.065	0.113	1.000
13	0.66	0.11	0.873	0.864	1.671	0.000	1.000
14	-2.58	0.24	0.978	0.736	1.029	0.000	1.000
15	0.32	0.11	1.101	1.137	0.454	0.180	0.893
16	0.96	0.11	1.040	1.048	0.827	0.049	0.994
17	-0.76	0.13	1.014	1.036	0.970	0.000	0.992
18	-0.06	0.11	0.968	0.941	1.129	0.000	1.000
19	-1.66	0.17	0.958	0.846	1.044	0.000	1.000
20	-0.71	0.13	0.977	0.972	1.040	0.022	1.000
21	0.18	0.11	0.892	0.876	1.495	0.000	1.000
22	0.87	0.11	1.005	1.013	0.968	0.013	1.000
23	-1.17	0.14	1.036	1.146	0.946	0.000	0.988
24	-1.09	0.14	0.935	0.818	1.100	0.818	1.000
25	0.67	0.11	1.031	1.040	0.827	0.060	0.990
26	0.55	0.11	1.055	1.055	0.709	0.052	0.918
27	-1.91	0.19	0.973	0.849	1.030	0.000	1.000
28	-0.02	0.11	0.979	0.982	1.073	0.000	1.000
29	-0.43	0.12	0.941	0.933	1.130	0.000	1.000
30	-1.98	0.19	1.016	1.201	0.972	0.000	0.995
31	-0.54	0.12	0.927	0.883	1.156	0.056	1.000
32	1.15	0.12	1.109	1.145	0.616	0.073	0.686
33	-0.52	0.12	0.883	0.797	1.264	0.000	1.000
34	-0.76	0.13	0.981	0.957	1.033	0.306	1.000
35	0.11	0.11	1.049	1.036	0.812	0.176	0.977
36	-0.92	0.13	0.956	0.867	1.078	0.000	1.000
37	1.17	0.12	0.946	0.927	1.188	0.000	1.000
38	1.64	0.12	1.056	1.188	0.843	0.039	0.929
39	1.64	0.12	1.034	1.043	0.930	0.022	1.000
40	1.3	0.12	1.023	1.001	0.954	0.004	0.863

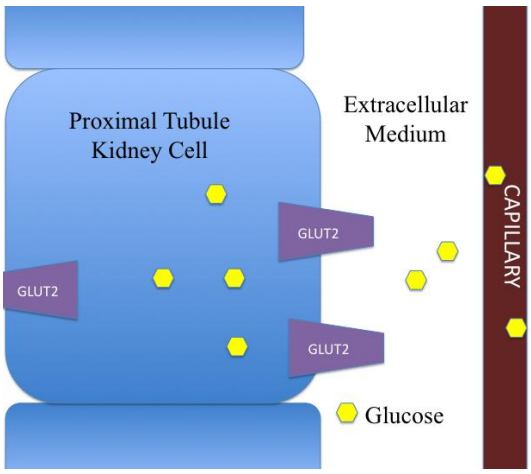
Item	*b* value	Error	Infit Mean Square	Outfit Mean Square	Discrimination	Lower Asymptote	Upper Asymptote
41	-0.32	0.12	0.977	1.000	1.047	0.000	1.000
42	-0.74	0.13	1.062	1.058	0.907	0.140	0.978
43	-0.31	0.12	0.970	0.961	1.080	0.107	1.000
44	-0.68	0.13	0.891	0.858	1.190	0.000	1.000
45	-0.22	0.12	0.906	0.858	1.290	0.000	1.000
46	-0.57	0.12	0.976	0.933	1.060	0.000	1.000
47	-0.74	0.13	0.994	1.023	1.002	0.000	1.000
48	0.21	0.11	0.970	0.964	1.147	0.000	1.000
49	0.92	0.11	1.041	1.045	0.820	0.044	0.960
50	-1.19	0.15	0.992	1.005	1.008	0.000	1.000
51	-0.26	0.12	0.995	0.999	1.011	0.021	1.000
52	-1.23	0.15	1.003	1.016	0.992	0.000	0.998
53	1.39	0.12	1.016	1.051	0.946	0.016	1.000
54	0.16	0.11	1.020	1.029	0.903	0.000	0.969
55	1.24	0.12	0.975	0.986	1.067	0.000	1.000
56	-0.56	0.12	0.920	0.849	1.176	0.241	1.000
57	-2.52	0.24	0.964	0.764	1.035	0.000	1.000
58	0.56	0.11	1.063	1.072	0.655	0.068	0.908
59	1.03	0.11	1.032	1.059	0.855	0.025	0.903
60	-0.18	0.12	1.055	1.032	0.854	0.396	0.981
61	0.62	0.11	1.023	1.035	0.860	0.049	0.989
62	1.07	0.11	1.053	1.073	0.787	0.058	0.987
63	0.94	0.11	0.957	0.945	1.193	0.000	0.941
64	-0.47	0.12	1.065	1.149	0.832	0.266	0.961
65	0.04	0.11	0.946	0.924	1.225	0.000	1.000
66	1.52	0.12	1.089	1.153	0.785	0.050	0.780
67	1.02	0.11	1.021	1.012	0.930	0.010	0.938
68	0.36	0.11	1.118	1.127	0.387	0.240	0.902
69	0.51	0.11	1.123	1.138	0.319	0.118	0.810
70	1.04	0.11	0.944	0.925	1.228	0.000	1.000
71	0.05	0.11	0.903	0.884	1.389	0.000	1.000
72	-1.15	0.14	1.016	1.038	0.974	0.000	0.994
73	0.92	0.11	0.966	0.950	1.162	0.000	1.000
74	-0.04	0.11	0.935	0.905	1.247	0.000	1.000
75	0.77	0.11	1.017	1.020	0.914	0.037	1.000
76	-0.11	0.12	0.952	0.951	1.154	0.000	1.000
77	-0.27	0.12	0.976	0.963	1.069	0.005	1.000
78	-0.78	0.13	0.908	0.900	1.139	0.000	1.000
79	0.74	0.11	1.031	1.039	0.835	0.049	0.982
80	-1.15	0.14	1.057	1.167	0.920	0.000	0.982

Item	*b* value	Error	Infit Mean Square	Outfit Mean Square	Discrimination	Lower Asymptote	Upper Asymptote
81	-0.95	0.14	1.002	0.962	1.006	0.000	1.000
82	-2.22	0.21	1.024	1.434	0.965	0.000	0.995
83	0.64	0.11	1.088	1.115	0.498	0.140	0.916
84	1.14	0.12	1.087	1.137	0.670	0.075	0.859
85	3.89	0.27	0.977	0.908	1.019	0.000	1.000
86	-0.37	0.12	1.040	1.050	0.902	0.306	0.981
87	0.61	0.11	0.976	0.971	1.132	0.004	1.000
88	0.44	0.11	0.970	0.970	1.158	0.000	1.000
89	1.13	0.11	1.118	1.168	0.559	0.118	0.972
90	-0.13	0.12	1.138	1.167	0.557	0.268	0.901
91	0.2	0.11	1.032	1.037	0.850	0.127	0.986
92	0.59	0.11	0.890	0.881	1.594	0.000	1.000
93	-0.43	0.12	1.156	1.252	0.622	0.373	0.911
94	1.37	0.12	0.916	0.907	1.213	0.000	1.000
95	0.06	0.11	0.982	0.961	1.090	0.027	1.000
96	0.04	0.11	0.939	0.939	1.231	0.000	1.000
97	0.07	0.11	0.999	1.007	0.998	0.000	0.996
98	-0.42	0.12	1.021	1.052	0.937	0.000	0.981
99	0.39	0.11	0.939	0.929	1.329	0.000	1.000
100	1.67	0.12	0.948	0.937	1.099	0.000	1.000
101	0.95	0.11	0.984	0.970	1.081	0.000	0.905
102	1.16	0.12	1.043	1.061	0.845	0.051	1.000
103	-0.5	0.12	0.951	0.884	1.127	0.113	1.000
104	1.16	0.12	1.039	1.076	0.841	0.046	1.000
105	0.91	0.11	0.988	0.999	1.039	0.015	1.000
106	0.55	0.11	1.053	1.074	0.686	0.118	0.969
107	-1.75	0.18	0.940	0.854	1.055	0.000	1.000
108	-0.95	0.14	1.081	1.158	0.877	0.000	0.971
109	-1.23	0.15	0.928	0.910	1.080	0.000	1.000
110	-1.11	0.14	1.028	1.249	0.934	0.821	0.985
111	-1.11	0.14	0.974	0.974	1.029	0.000	1.000
112	0.45	0.11	1.048	1.050	0.740	0.063	0.941
113	-0.67	0.13	1.048	1.032	0.929	0.755	0.985

Appendix F: The 23 validated, critical thinking items on the Spring 2012 exam. Correct answers are indicated in bold type. The number of students (n) who selected each answer choice is indicated in parentheses.

86. In the diagram below, GLUT2 is a facilitated diffusion transporter that transports glucose out of proximal tubule cells in the kidney into the extracellular medium. The glucose then moves from the extracellular medium into the capillary. Which of the following would increase the rate of GLUT2 activity:

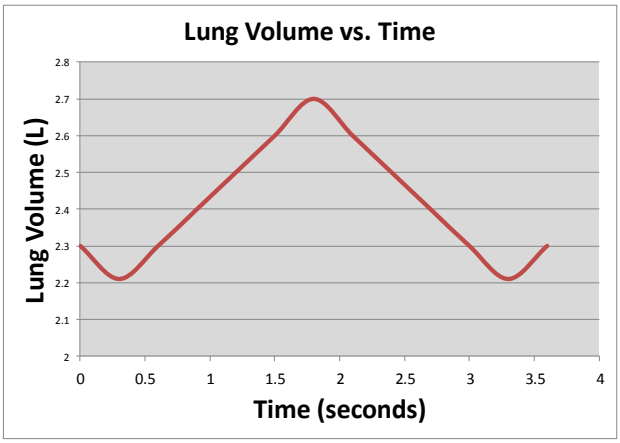
f. Increasing the fructose concentration gradient across the proximal tubule cell membrane. (n = 8)
g. **Increasing the concentration of glucose in the proximal tubule kidney cells**. (n = 249)
h. Increasing the rate at which kidney cells metabolize glucose. (n = 38)
i. Shuttling more ATP to GLUT2. (n = 55)
j. Decreasing the rate of glucose entry into the capillary. (n = 7)



87. Which of the following would result from reducing the surface area to volume ratio (SA/V) of a cell (same shape):

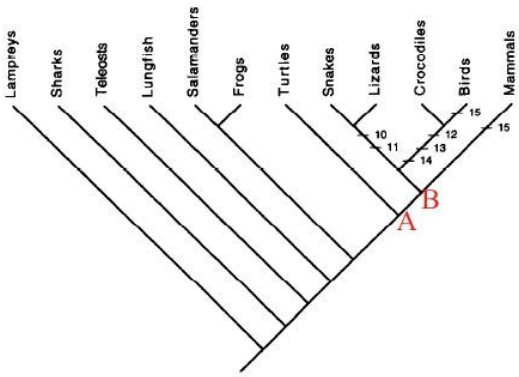
a. The time it takes O_2 to diffuse from the cell surface to the mitochondria would increase. (n = 174)
b. The amount of glucose needed to fuel the cell would decrease. (n = 133)
c. The amount of genomic DNA would increase. (n = 6)
d. The surface area would increase at a faster rate than the volume. (n = 37)
e. The cell membrane would become porous. (n = 8)

88. The graph below describes the typical breathing pattern for an adult male at rest. At which of the following time points would you expect the sarcomeres of the thoracic diaphragm to be contracting:
    a. 0 seconds (n = 14)
    b. 1 second (n = 188)
    c. 2.5 seconds (n = 52)
    d. 3.2 seconds (n = 92)
    e. Since the person is at rest, the sarcomeres of the thoracic diaphragm are not contracting. (n = 12)



89. The intracellular calcium concentration [Ca^{2+}] of cardiac muscle cells is lower than the extracellular [Ca^{2+}]. This is important for heart function. To maintain this lower intracellular [Ca^{2+}], cardiac muscle cells rely on a $3Na^{+}/1Ca^{2+}$ antiporter. Given that the import of Na^{+} drives the export of Ca^{2+} ions, which of the following is true (HINT: Draw a diagram:
    a. The antiporter transports Na^{+} down its concentration gradient. (n = 133)
    b. Downregulating the activity of the sodium/potassium pump will not affect the $3Na^{+}/1Ca^{2+}$ antiporter. (n = 19)
    c. The antiporter utilizes ATP for energy. (n = 130)
    d. The antiporter maintains equal intracellular [Na^{+}] and [Ca^{2+}]. (n = 19)
    e. The antiporter stops working when the intracellular [Na^{+}] drops below the intracellular [Ca^{2+}]. (n = 57)

90. Two different species of animals have the homologous trait of long curved claws on their forelimbs. Based on this information you can infer that:
    a. The two species of animals are the closest living relatives. (n = 17)
    b. The most recent common ancestor of the two species had long, curved claws. (n = 232)
    c. The two species of animals use their long, curved claws for similar functions. (n = 81)
    d. The gene that confers long, curved claws is identical between the two species. (n = 27)
    e. The two species of animals also have long curved claws on their hindlimbs. (n = 0)



91. What does the node at point A in the above figure represent?
    - **a. The point where the turtle ancestor diverged from the ancestor to snakes, lizards, crocodiles, birds and mammals. (n = 207)**
    - b. The point where the turtle ancestor diverged from the ancestor to snakes. (n = 26)
    - c. The point where the turtle ancestor diverged from the ancestor to mammals. (n = 21)
    - d. The point where the turtle ancestor diverged from the ancestor to frogs and salamanders. (n = 103)
    - e. The point where the turtle ancestor diverged from the ancestor to frogs. (n = 1)

92. A herbicide that kills germinating seeds most likely blocks the process of:
    - **a. Mitosis (n = 175)**
    - b. Fertilization (n = 88)
    - c. Meiosis (n = 41)
    - d. Gametogenesis (n = 38)
    - e. Peptidoglycan formation (n = 15)

93. Multiple Sclerosis (MS) is characterized by inflammation and loss of myelination of the neurons of the Central Nervous System. Which of the following do you predict would occur in patients with MS:

    a. Disruption of impulses along the somatic motor neurons. (n = 57)
    b. Disruption of impulses along neurons of the medulla. (n = 34)
    c. Spinal cord neurons are more susceptible to damage than sensory neurons. (n = 10)
    d. Patients with MS do not respond to anti-inflammatory drugs. (n = 3)
    e. **Both B and C**. (n = 254)

94. Rigor mortis (stiffness of death) is believed to result from the depletion of ATP in skeletal muscle cells. Lack of ATP in skeletal muscle cells after death would result in:
    a. A net flow of Ca^{2+} from the sarcoplasmic reticulum into the cytoplasm of the skeletal muscle cell. (n = 115)
    b. An increase in the rate of glucose metabolism. (n = 3)
    c. Tropomyosin blockage of myosin binding sites. (n = 108)
    d. The detachment of actin from myosin. (n = 105)
    e. A net flow of Na^+ from the cytoplasm of the skeletal muscle cell to the extracellular space. (n = 26)

95. The resting potential of a neuron is measured in the lab by a voltmeter. Which of the following would cause the voltmeter to register a membrane potential more negative than -70mV?
    a. Reversing the positions of the reference and measurement electrodes. (n = 11)
    b. Adding a chemical that opens Na^+ channels. (n = 98)
    c. Adding glucose to the extracellular fluid. (n = 5)
    d. Adding a chemical that opens K^+ channels. (n = 218)
    e. Adding a chemical that blocks Cl^- channels (n = 26)

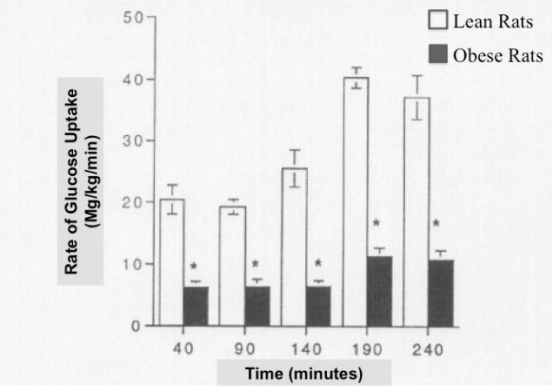
96. Which of the following pieces of evidence BEST supports the hypothesis that the Monoamine Oxidase (MAO) cleaves the neurotransmitter norepinephrine (NE) in the autonomic nervous system:
    a. Smooth muscle cells produce MAO. (n = 30)
    b. Norepinephrine is classified as a monoamine. (n = 24)
    c. Pharmacologic inhibitors of MAO do not increase the frequency of skeletal muscle contractions. (n = 54)
    d. Age related decreases in MAO activity correlate with cognitive defects. (n = 30)
    e. Mice deficient in MAO show increased levels of norepinephrine in the brainstem. (n = 219)

98. Cardiac fibrosis is marked by large, stiff collagen deposits between cardiac cells and is commonly seen in patients with chronic heart failure. Inserting collagen between cardiac cells can result in:
    a. **Decreased ion flow between cardiac cells**. **(n = 253)**
    b. Decreased heart size. (n = 4)
    c. Misalignment of cardiac sarcomeres. (n = 52)
    d. Mixing of oxygenated and deoxygenated blood. (n = 12)
    e. Increased ability of the heart to contract. (n = 37)

99. In a double circulatory system, the pressure that sends the blood to the systemic circuit is generated during _____ by the _____ :
    a. Diastole, Left Ventricle (n = 60)
    b. Systole, Right Atrium (n = 55)
    c. Diastole, Right Ventricle (n = 31)
    d. **Systole, Left Ventricle (n = 192)**
    e. Systole, Left and Right Atria (n = 20)

100. Which of the following would increase the rate of glucose diffusion from the capillaries to the surrounding cells:
    a. Decreasing the amount of glucose in the blood. (n = 24)
    b. **Decreasing the flow velocity of blood through the capillaries. (n = 95)**
    c. Increasing the flow velocity of blood though the veins. (n = 48)
    d. Decreasing the total cross sectional area of the capillaries. (n = 55)
    e. Increasing the number of glucose active transporters in the capillary membrane. (n = 136)

101. The graph below shows the rate of glucose uptake by tissues in lean (unfilled bars) and obese (filled bars) rats after being given equal amounts of glucose. The y-axis shows how fast glucose is taken up by the cells. Based on the graph, which of the following is true:



    a. The lean rats have higher blood glucose value at the 140 minute mark than at the 90 minute mark. (n = 110)
    b. The obese rats have the same blood glucose values at the 40 minute and 90 minute marks. (n = 80)
    c. **The obese rats have higher blood glucose values than the lean rats at all time points**. (n = 147)
    d. The lean rats were more active at the 40 minute mark than at the 190 minute mark. (n = 16)
    e. The obese rats had higher glucose uptakes than the lean rats. (n = 5)

102. The kidneys of healthy individuals secrete Erythropoietin. However, individuals with severe kidney disease are unable to produce Erythropoietin. Which of the following do you predict is associated with severe kidney disease?
    a. High amounts of Red Blood Cells. (n = 9)
    b. Increased chance of blood clots due to abnormally viscous blood. (n = 146)
    c. **Poor Oxygen delivery to tissues. (n = 131)**
    d. A right-shift of the Oxygen binding to hemoglobin curve. (n = 32)
    e. Low levels of waste products in the blood. (n = 40)

103. After being filtered into the filtrate, Drug XYZ is poorly reabsorbed by the kidneys. Based on this information, which of the following is most likely true:
    a. The kidneys are inefficient at clearing Drug XYZ from the blood. (n = 52)
    b. **Individuals taking Drug XYZ excrete large amounts of the drug in their urine. (n = 259)**
    c. The molecular size of Drug XYZ is larger than a Red Blood Cell. (n = 29)
    d. Drug XYZ is positively charged. (n = 5)
    e. Secretion of renin will lower the rate at which the kidneys filter Drug XYZ from the blood. (n = 13)

104. A new animal previously unknown to humans was just discovered! The identity of this animal is kept secret, but you hear about some lab results. You learn that the animal has blood values of 894 milli-osmoles/L and produces urine that is 493 milli-osmoles/L. Of the choices listed below, this newly discovered animal CANNOT be a:
    a. Turtle (n = 7)
    b. Freshwater Fish (n = 156)
    c. **Bald Eagle (n = 131)**
    d. Crocodile (n = 17)
    e. Cricket (n = 47)

105. Joe woke up in the morning and ate a full breakfast of salted bacon, an orange, and toast. Which combination of hormones do you predict is taking effect in Joe after this breakfast:
    a. Renin, Angiotensisn, Insulin (n = 74)
    b. Glucagon, Anti-diuretic Hormone, Calcitonin (n = 31)
    c. Insulin, Calcitonin, Renin (n = 81)
    d. Glucagon, Parathyroid Hormone, Angiotensin (n = 22)
    e. **Parathyroid Hormone, Insulin, Anti-diuretic Hormone (n = 150)**

106. It has been observed that secretion of progesterone by the corpus luteum causes a woman's body temperature to rise by 0.5°F. If the woman does not get pregnant, the body temperature drops by 0.5°F around the time of menstruation. Based on this information, at what point during a woman's 28 day cycle would you predict this increase in temperature occurs:
    a. Day 7 (n = 12)
    b. Day 10 (n = 46)
    c. **Day 16 (n = 179)**
    d. Day 22 (n = 81)
    e. Day 28 (n = 39)

107. The tiny Corkus organism lives on the skin of large, hairy Spudnus and causes the Spudnus to develop itchy boils. Which of the following would make this interaction an example of mutualism:
f. The Spudnus provides the Corkus with warmth. (n = 6)
g. The Corkus feeds off the Spudnus' secretions. (n = 4)
h. The hair of the Spudnus protects the Corkus from UV Rays. (n = 2)
i. All of the above (n = 26)
j. None of the above (n = 320)

108. The anterior pituitary of a female with hypogonadism secretes abnormally low levels of Luteinizing Hormone. Insufficient levels of Luteinizing Hormone can lead to:
a. Increased endometrial development (n = 17)
b. Development of multiple follicles at a time (n = 22)
c. **Failure to ovulate (n = 285)**
d. Increased fertility (n = 10)
e. Abnormally high levels of Inhibin (n = 23)

109. Fick's Law has been applied to gas exchange and insulin exchange. As it applies to insulin exchange between the plasma and the muscle interstitium: $Q = PS\ (C_p - C_I)$. In this equation Q = rate of insulin exchange, P = the permeability of the surface to insulin, S = the surface area for exchange, C_P = plasma insulin concentration, C_I = interstitium insulin concentration. Increasing the number of capillaries recruited to muscle tissue would enhance insulin delivery to muscle by:
a. Increasing P (n = 36)
b. **Increasing S (n = 300)**
c. Increasing C_P (n = 16)
d. Decreasing C_I (n = 5)
e. Decreasing PS (n = 1)