ANALYZING CALCULUS CONCEPT INVENTORY GAINS IN INTRODUCTORY CALCULUS

Matthew Thomas, University of Arizona
Guadalupe Lozano, University of Arizona

Research in science education, particularly physics education, indicates that students in Interactively-Engaged classrooms are more successful on tests of basic conceptual knowledge. Despite this, undergraduate mathematics courses are dominated by lectures in which students take a passive role. Given the value of such tests in assessing students' conceptual knowledge, the method for measuring such change is largely unexplored. In our study, students were given one such inventory, the Calculus Concept Inventory, in introductory Calculus classes as a pretest and posttest. We address issues of how gains might be measured on this instrument using two techniques, and the implications of using each of these measures.

Keywords: Calculus, Measurement, Concept Inventories

A recent report in the MAA Notices stated that almost two thirds of college and university instructors surveyed agreed with the statement that "Calculus students learn best from lectures, provided they are clear and well-prepared" (Bressoud, 2011, p. 1). Strong support for traditional lecture as the primary means of undergraduate mathematics instruction remains in spite of a growing body of research supporting the claim that students learn best when they are interactively and cognitively engaged with subject matter in mathematics and other sciences (Epstein, 2007; Hake, 1998a; Prather, Rudolph, Brissenden, & Schlingman, 2009; Rhea, n.d.; Smith et al., 2005). These studies often use instruments called concept inventories to measure conceptual knowledge gains by giving the instrument as a pretest and posttest. We discuss two types of measures of gain on one such instrument, the Calculus Concept Inventory, given to introductory calculus students at a large southwestern university, and the differences observed by using each of the measures of gain.

Background

Conceptual understanding may be measured through instruments called concept inventories. The first concept inventory, the Force Concept Inventory (FCI), is a test in introductory mechanics which paved the way for analyzing student conceptual understanding of the basic ideas in a subject area (Hake, 1998a, 2007; Hestenes & Wells, 1992; Hestenes, Wells, & Swackhamer, 1992). Since then, many concept inventories have been written in various subject areas (Allen, 2006; Anderson, Fisher, & Norman, 2002; Carlson, Madison, & West, 2010; Carlson, Oehrtman, & Engelke, 2010; Garvin-Doxas, Klymkowsky, & Elrod, 2007; Libarkin, 2008; Marbach-Ad et al., 2009, 2010; Mulford & Robinson, 2002; Prather et al., 2009; Rhoads & Roedel, 1999). There has been active discussion about how to interpret the results of the FCI (Heller & Huffman, 1995; Henderson, 2002; Hestenes & Halloun, 1995; Huffman & Heller, 1995). The concept inventory discussed in this paper is the Calculus Concept Inventory (CCI), developed by Epstein (2007).

One of the most well-known uses of the FCI was Hake's (1998a) comparison of classrooms which utilized Interactive-Engagement (IE) methods with those which were described as "traditional lecture" (TL). In his study, Hake defined IE teaching as a collection of teaching

methods which are "designed at least in part to promote conceptual understanding through interactive engagement of students in heads-on (always) and hands-on (usually) activities which yield immediate feedback through discussion with peers and/or instructors" and found differences between the two types of classes of almost two standard deviations (Hake, 1998a, p. 65). The concept of an IE classroom has been further explored in physics by Hake (1998b) and in mathematics by Epstein (2007). IE teaching styles share features with Peer Instruction (Mazur, 1997) including ConcepTests (Pilzer, 2001), and pure discovery learning (Paris & Paris, 2001).

Normalized Gain

Normalized gain is a measure first used by Hake (1998) in his study of Interactively-Engaged teaching styles with the Force Concept Inventory to measure how much material has been learned by students during a course. This measure is almost always used in concept inventory studies for measuring gains. In particular, this is the gain score that Epstein (2007) and Rhea (n.d.) used to report their findings on the Calculus Concept Inventory. The normalized gain score is defined as

$$\langle g \rangle = \frac{Final - Initial}{Total - Initial}$$

and measures the fraction of unknown material learned throughout the course. For example, if a student correctly answered 50% of the questions on the pretest, and 75% on the posttest, the normalized gain would be <g> = 0.5, meaning that student correctly answered half of the 50% of the material they did not know at the beginning of the class. Normalized gain is often calculated using section averages of pretest and posttest scores, so each section of a course will be assigned a single normalized gain score. Many studies compare the effects of instructional practices on student learning, so the effect of interest is at the section level: normalized gains calculated at the section level allow one to analyze the effect of instructional practices on the entire class. One can also create an individual normalized gain score by using the pretest and posttest score for each student. The effect of computing individual normalized gains for each student has been investigated and compared to using section-level normalized gain scores (Bao, 2006; Coletta & Phillips, 2005). The two methods produce close, though not numerically identical, results. The advantage of considering individual-level normalized gains is that class-level variables can be considered along with student-level variables such as demographics or previous mathematics courses.

Item Response Theory

Item Response Theory (IRT) is a modern approach to analyzing instruments like tests or surveys (Embretson & Reise, 2000). IRT is based on the idea that an instrument measures a single latent trait or ability, such as conceptual knowledge of calculus. While this trait cannot be directly observed, the effects can be observed through answers to questions. In an email to Hake, Mislevy (n.d.) pointed out that IRT has some benefits over the use of normalized gains such as handling floor and ceiling effects (e.g., students who obtain a perfect score obtain a normalized gain of 1 regardless of initial ability levels). IRT also provides the opportunity to analyze individual questions as opposed to a single test score for each individual.

IRT is a methodology for predicting ability levels based on an instrument and can be leveraged to determine gain scores by measuring ability levels on the pretest and on the posttest. The difference between these two ability levels is the change over the course, and so measures gain (Wallace & Bailey, 2010). Despite the benefits of using IRT to create gains scores instead of using normalized gains, as Mislevy suggested to Hake for analyzing the FCI, IRT analysis is very rarely used in science and mathematics education research (Wallace & Bailey, 2010). The

FCI has been analyzed using IRT (Wang & Bao, 2010), and comparisons of IRT and normalized gain methods have been made in astronomy (Wallace & Bailey, 2010). An IRT analysis of the CCI has not been published, nor have multiple measures of gain been studied for this instrument. Our study contributes to the existing literature in both of these areas.

Purpose of Study

Our study builds on previous studies by considering the implications of using both normalized gain and IRT to measure gains on CCI. Previous studies have investigated the effects of instructional practices on student learning by using concept inventory pretests and posttests. In order for the connections between instructional choices and learning gains to be studied, a method for measuring learning gains needs to be established. It is also important that these gain scores accurately reflect learning since they may be used for practical decisions, such as administrative decisions involving the careers of teachers. If a teacher's career is affected by how much their students' scores improve on an exam, it is worth considering that there are multiple ways to measuring these gains which may produce different results.

Methods & Analysis

The subjects in this study were Calculus I students at a large southwestern university. All Fall 2010 Calculus I students at the university were required to take the Calculus Concept Inventory as a pretest and posttest as part of the course, and consenting students had their scores collected along with demographic information. The pretest was graded for course credit on completion, and the posttest scores were factored into students' final grades. Of the 880 students who took the pretest and 668 students who took the posttest, 507 students took both tests and consented to participate in the study. Of these students, 482 had non-zero scores on the CCI pretest, CCI posttest, and the final exam, meaning they did not miss any of the tests. There were 26 sections, with a maximum capacity of 35 in each section. On average, 18.5 students per section participated in the study, ranging from 10 to 26.

Analysis was done using a combination of the software tools BILOG-MG and R. These are commonly used software tools for conducting IRT analysis (BILOG-MG) and general purpose statistical analysis (R).

Results and Implications

Normalized Gain Scores

In his 1998a study, Hake grouped sections by their normalized gain, <g>, scores: “low-g” sections were defined as those with <g> values less than 0.3, “medium-g” as those between 0.3 and 0.7, and “high-g” as those above 0.7 (p. 65). In his study, Epstein (2007) gave the CCI to 1100 students at 12 American universities and 1 university in Finland. He found <g> values largely clustered between 0.15 and 0.23, similar to the scores in traditional lecture physics classes on the FCI. A large midwestern research university with a department-wide IE focused teaching style reported an average <g> score of 0.35 among their 51 sections, with a range of 0.21 to 0.44 (Rhea, n.d.). Ten of the sections had <g> scores above 0.40.

The mean normalized gain for the entire participant group at the large, southwestern university where our study was conducted was 0.25, meaning that 25% of the previously unknown material was learned during the course. Normalized gain scores ranged from 0.14 to 0.36. By Hake's definition, 4 of the 26 sections had medium-g scores, and the remaining 22 had low-g scores.

Individual normalized gain scores were then computed so that a comparison with IRT gains could be made, since IRT gains are computed at the individual level. A histogram is given in

Figure 1. The difference between using individual normalized gains averaged by class and class average normalized gains makes almost no difference in this data set, as the mean was slightly higher, but still 0.25 when rounded to two decimal places.

IRT Gains

Item Response Theory allows for a variety of models to be created. The Rasch model is the simplest of these, estimating a single parameter for each item. In the Rasch model, the probability of an individual, *i,* correctly answering question *p* is given by

$$P(X_{pi}=1|\theta_p,b_i)=\frac{\exp[\theta_p-b_i]}{1+\exp[\theta_p-b_i]}$$

which results in a logistic curve for each item (Embretson & Reise, 2000). For the CCI pretest under the Rasch model, the plot of curves for each item is given in Figure 2. The interpretation is that for any level of conceptual knowledge of calculus (θ), there is a corresponding item-specific probability of correctly answering that item. The difficulty of item *i*, b_i, is the ability which is required for a 50% chance of correctly answering the question. For the Rasch model, the difficulty of each item uniquely determines this curve, called the item characteristic curve.

One can introduce a different model by introducing a new parameter, α, called the discrimination. This model is given by the formula

$$P(X_{pi}=1|\alpha,\theta_p,b_i)=\frac{\exp[\alpha(\theta_p-b_i)]}{1+\exp[\alpha(\theta_p-b_i)]}$$

The effect of the discrimination parameter is to change the slopes of all the item characteristic curves. The larger the value of α, the steeper the slope of the curve, and the more discriminating the item. This model is known as the one parameter logistic model (1PL). A plot of the item characteristic curves for the CCI pretest is given in Figure 3.

The two parameter logistic model (2PL) relaxes the condition that the discrimination parameter must be the same for all items, resulting in the following model.

$$P(X_{pi}=1|\alpha_i,\theta_p,b_i)=\frac{\exp[\alpha_i(\theta_p-b_i)]}{1+\exp[\alpha_i(\theta_p-b_i)]}$$

The graphs of item characteristic curves for the CCI under the 2PL model is given in Figure 4.

Looking at Figure 4, it is apparent that the behavior of the last item on the test is counter to what one would expect for a test measuring a single construct: the item characteristic curve for item 22 is decreasing, indicating that as one's conceptual knowledge of calculus increases, the likelihood of answering the item correctly decreases. The same item was also a poor fit on the posttest, where the questions were reordered. This is seen in Figure 5, where item 20 does not fit the pattern of the other items well. Since no explanation for this behavior was apparent, the item was removed from future analysis. Using the method proposed by Bejar (1980), it was determined that the remaining items were assessing the same construct.

To ensure that the pretest and posttest scores were comparable, item parameters were estimated from the pretest, and these item parameters were used to estimate individual ability levels for both pretest and posttest students, following the method used by Wallace and Bailey (2010). Therefore, a score of 0 was a common score, interpretable as the ability of an average student taking the pretest. A student who is estimated to have an ability level of 0 on the posttest would then have an ability comparable to an average student at the beginning of the course.

Comparison of the Two Gain Models

A comparison of the section average normalized gains and IRT gains averaged by section is displayed in Figure 6. The two measures of gain are strongly correlated, $r(21) = 0.92$, $p < 0.01$,

so 86% of the variation in one measure is explained by the other. This is low if the two quantities are actually measuring the same trait – the knowledge gained over the course. At the student level, the correlation between the normalized gains and IRT gains is similarly correlated, $r(480) = 0.92$, $p < 0.01$, suggesting that 15% of variation of one measure is still unaccounted for by the other measure. This relationship is shown in Figure 7. If one were to interpret one of the two gain measures (IRT or normalized gains) as "correct," then using the other measure would introduce error. This suggests that, unless one measure can be objectively preferred to the other, the two measures provide different information, and answer different questions.

Practically, one of the fundamental differences between normalized gain scores and IRT gain scores is the dependence of the normalized gain score on the pretest score. Given two classes with the same difference between pretest and posttest scores, the normalized gain will be larger for the class with the higher pretest score (Wallace & Bailey, 2010). Consider a concept inventory with 22 questions. Suppose a class with a pretest average of 18 points achieves a posttest average of 20 points, and another class with a pretest average of 11 points achieves a posttest average of 13 points. The normalized gain score would be higher for the first class more than the second. Those 2 points were likely more difficult to achieve than the 2 points achieved by the second class since the pool of available questions to improve upon is smaller for the first class. Additionally, if the difficulties of the questions are distributed roughly normally, a class with a medium pretest is likely to encounter questions which are not far beyond their ability, while a high achieving class may encounter questions which are farther from their (current) ability. In this way, the normalized gains are rewarding classes in a reasonable way.

From a theoretical point of view, the normalized gain is a test-specific measure of gain. The normalized gain is the total learned out of the total that could be learned, implying a maximum knowledge level. Once a student has correctly answered all of the questions on the instrument, the total knowledge has been achieved for that construct. IRT does not have this type of frame of reference of total knowledge. In this way, normalized gains are a measure much more closely tied to a particular instrument than IRT gains. Consider the following example. Suppose version 1 of a concept inventory has 2 questions which are correctly answered by everyone who takes the test, and version 2 of the concept inventory replaces those 2 items with 2 items which are answered incorrectly by everyone on the test. In an IRT analysis, this change would not make any difference since these 2 items provide no information that allows individuals to be compared. In a normalized gain analysis of the concept inventories, however, the normalized gains will be different. If the concept inventory had 22 questions, a change from 14 to 16 ($\langle g \rangle = .25$) on version 1 would become a change from 12 to 14 ($\langle g \rangle = 0.2$) on version 2. These scores are still worth considering and can reveal a great deal about gains during a course, but the dependence on the instrument itself should not be ignored. This is also noteworthy when comparing normalized gains on the FCI to normalized gains on the CCI. These are two completely different instruments, and so comparing normalized gains on one with the other may not be reasonable. In particular, useful cutoffs for high-, medium-, and low-$\langle g \rangle$ scores may not transfer from one test to the other.

Conclusions

Measuring gains using a pretest and posttest appears like a simple task, but the two measures investigated here produce quite different results. Both IRT gains and normalized gains aim to determine the amount of learning that has taken place during a course, but they are measuring this quantity in different ways, producing different results. A priori, there is no objective way to chose one measure as preferred to the other, as each have advantages. IRT produces measures

which are test and population independent, but are more difficult to interpret than normalized gain scores. IRT's test independence is an advantage. With IRT, if a different test were created which measures the same construct as the CCI and were given to the same population, the individual ability estimates would remain unchanged. This is not the case for a normalized gain analysis. The maximum is achieved when the student has successfully mastered the material that is on the test. While much can be learned from studies of other concept inventories, he CCI is measuring mathematics-specific ability, and so needs to be studied further. The difference between normalized gains and IRT gains demonstrates that we can assess gains on the CCI in different ways and achieve different results, highlighting the need for attention to the type of gain score used.

Plans for Future Research

The Calculus Concept Inventory plays an important role in our study as an externally created and validated measure of student understanding of the concepts of introductory calculus. To build upon the analysis of gains presented here, hierarchical linear models will be created to incorporate student-level variables such as demographic information and mathematics background, and final exam scores will be used to consider the relationship with potentially different types of knowledge.

References

Allen, K. (2006). *The Statistics Concept Inventory: Development and analysis of a cognitive assessment instrument in statistics*. (Unpublished doctoral dissertation). University of Oklahoma.

Anderson, D. L., Fisher, K. M., & Norman, G. J. (2002). Development and evaluation of the conceptual inventory of natural selection. *Journal of Research in Science Teaching*, *39*(10), 952–978. doi:10.1002/tea.10053

Bao, L. (2006). Theoretical comparisons of average normalized gain calculations. *American Journal of Physics*, *74*(10), 917. doi:10.1119/1.2213632

Bejar, I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement*, *17*(4), 283–296.

Bressoud, D. M. (2011). The worst way to teach. *MAA Launchings*, *July*.

Carlson, M., Madison, B., & West, R. (2010). The Calculus Concept Readiness (CCR) Instrument: Assessing student readiness for calculus. History and Overview.

Carlson, M., Oehrtman, M., & Engelke, N. (2010). The Precalculus Concept Assessment: A tool for assessing students' reasoning abilities and understandings. *Cognition and Instruction*, *28*(2), 113–145. doi:10.1080/07370001003676587

Coletta, V. P., & Phillips, J. a. (2005). Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability. *American Journal of Physics*, *73*(12), 1172. doi:10.1119/1.2117109

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

Epstein, J. (2007). Development and validation of the Calculus Concept Inventory. *Proceedings of the Ninth International Conference on Mathematics Education in a Global Community* (pp. 165–170).

Garvin-Doxas, K., Klymkowsky, M., & Elrod, S. (2007). Building, using, and maximizing the impact of concept inventories in the biological sciences: Report on a National Science Foundation sponsored conference on the construction of concept inventories in the biological sciences. *CBE—Life Sciences Education*, *6*(4), 277. doi:10.1187/cbe.07

Hake, R. R. (1998a). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, *66*(1), 64–74. doi:10.1119/1.18809

Hake, R. R. (1998b). Interactive-engagement methods in introductory mechanics courses. *Physics Education Research*, *74*, 64–74.

Hake, R. R. (2007). Six lessons from the physics education reform effort. *Lat. Am. J. Phys. Educ. Vol*, *1*(1), 24.

Heller, P., & Huffman, D. (1995). Interpreting the force concept inventory: A reply to Hestneses and Halloun. *The Physics Teacher*, *33*, 507–511.

Henderson, C. (2002). Common concerns about the force concept inventory. *Physics Teacher*, *40*(9), 542–542.

Hestenes, D., & Halloun, I. (1995). Interpreting the Force Concept Inventory: A response to March 1995 critique by Huffman and Heller. *Physics Teacher*, *33*(8), 502–504.

Hestenes, D., & Wells, M. (1992). A mechanics baseline test. *The Physics Teacher*, *30*(3), 159. doi:10.1119/1.2343498

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, *30*(3), 141–158. doi:10.1119/1.2343497

Huffman, D., & Heller, P. (1995). What Does the Force Concept Inventory Actually Measure? *Physics Teacher*, *33*(3), 138–43.

Libarkin, J. (2008). Concept inventories in higher education science. *National Research Council Promising Practices in Undergraduate STEM Education Workshop 2*. Washington, D. C.

Marbach-Ad, G., Briken, V., El-Sayed, N. M., Frauwirth, K., Fredericksen, B., Hutcheson, S., Gao, L.-Y., et al. (2009). Assessing student understanding of host pathogen interactions using a concept inventory. *Journal of Microbiology & Biology Education*, *10*(1), 43–50. doi:10.1128/jmbe.v10.98

Marbach-Ad, G., McAdams, K. C., Benson, S., Briken, V., Cathcart, L., Chase, M., El-Sayed, N. M., et al. (2010). A model for using a concept inventory as a tool for students' assessment and faculty professional development. *Life Sciences Education*, *9*(4), 408. doi:10.1187/cbe.10

Mazur, E. (1997). *Peer Instruction: A user's manual*. Upper Saddle River, NJ: Prentice–Hall.

Mislevy, B. (n.d.). Mislevy Hake Emails. Retrieved from http://www.education.umd.edu/EDMS/mislevy/papers/Gain/

Mulford, D. R., & Robinson, W. R. (2002). An inventory for alternate conceptions among first-semester general chemistry students. *Journal of Chemical Education*, *79*(6), 739. doi:10.1021/ed079p739

Paris, S. G., & Paris, A. H. (2001). Classroom applications of research on self-regulated learning. *Educational Psychologist*, *36*(2), 89–101. doi:10.1207/S15326985EP3602

Pilzer, S. (2001). Peer instruction in physics and mathematics. *Primus*, *11*(2), 185–192. doi:10.1080/10511970108965987

Prather, E., Rudolph, A. L., Brissenden, G., & Schlingman, W. M. (2009). A national study assessing the teaching and learning of introductory astronomy. Part I. The effect of interactive instruction. *American Journal of Physics*, *77*(4), 320. doi:10.1119/1.3065023

Rhea, K. (n.d.). The Calculus Concept Inventory at a large research university. *Unpublished manuscript*.

Rhoads, T. R., & Roedel, R. J. (1999). The Wave Concept Inventory - A cognitive instrument based on Bloom's Taxonomy. *Proceedings of the 29th ASEE/IEEE Frontiers in Education Conference* (Vol. 3, pp. 14–18). IEEE.

Smith, A. C., Stewart, R., Shields, P., Hayes-Klosteridis, J., Robinson, P., & Yuan, R. T. (2005). Introductory biology courses: A framework to support active learning in large enrollment introductory science courses. *Cell Biology Education*, *4*(2), 143–56. doi:10.1187/cbe.04-08-0048

Wallace, C. S., & Bailey, J. M. (2010). Do concept inventories actually measure anything? *Astronomy Education Review*, *9*(1), 010116. doi:10.3847/AER2010024

Wang, J., & Bao, L. (2010). Analyzing force concept inventory with item response theory. *American Journal of Physics*, *78*(10), 1064. doi:10.1119/1.3443565

Figures

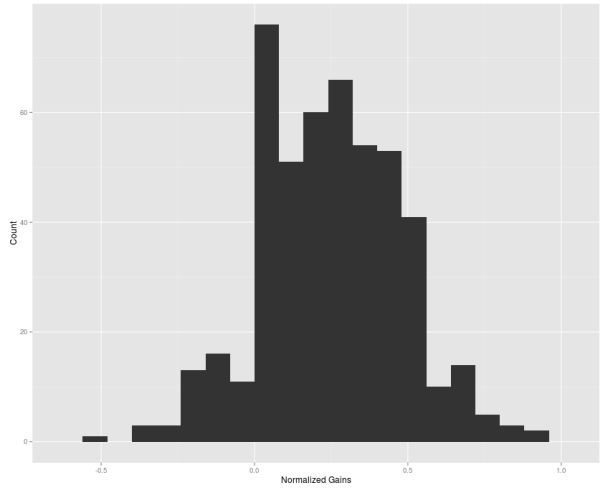

Figure 1: Histogram of Normalized Gains

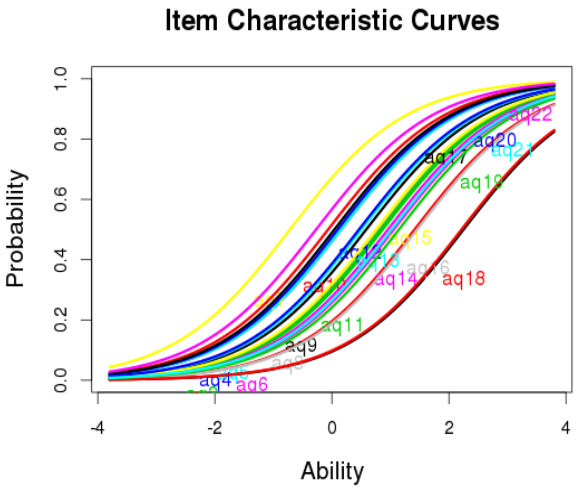

Figure 2: Rasch Model for CCI Pretest

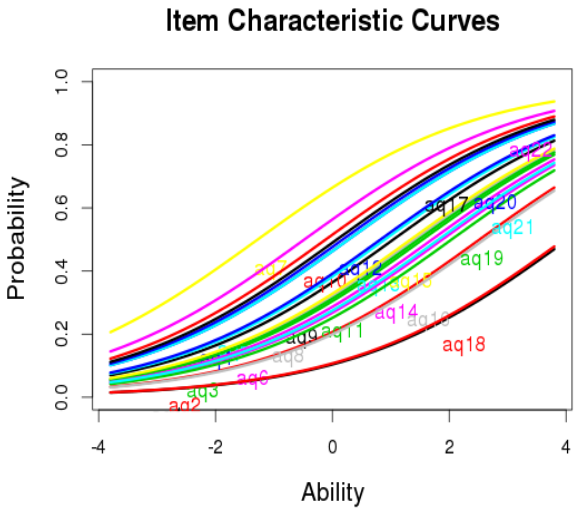

Figure 3: 1PL Model for CCI Pretest

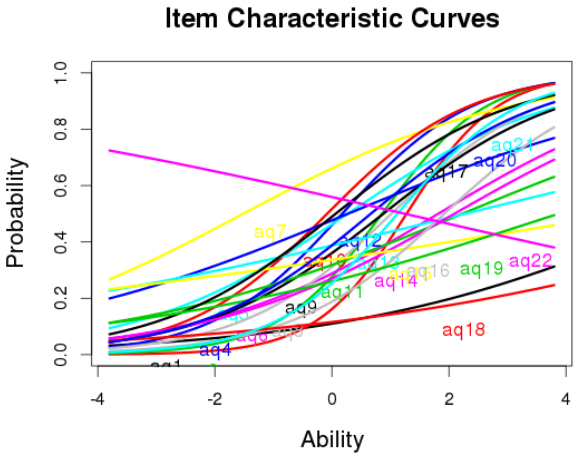

Figure 4: 2PL Model for CCI Pretest

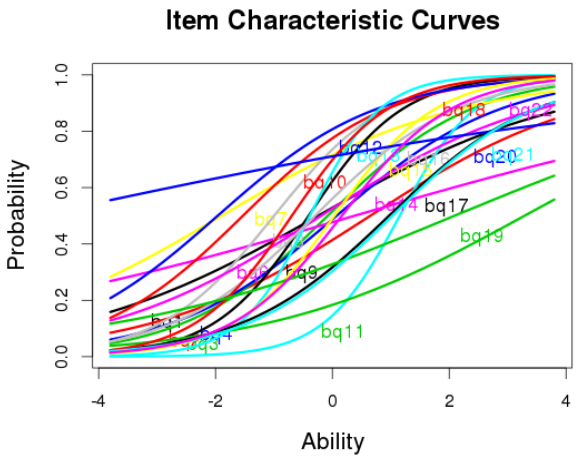

Figure 5: 2PL Model for CCI Posttest

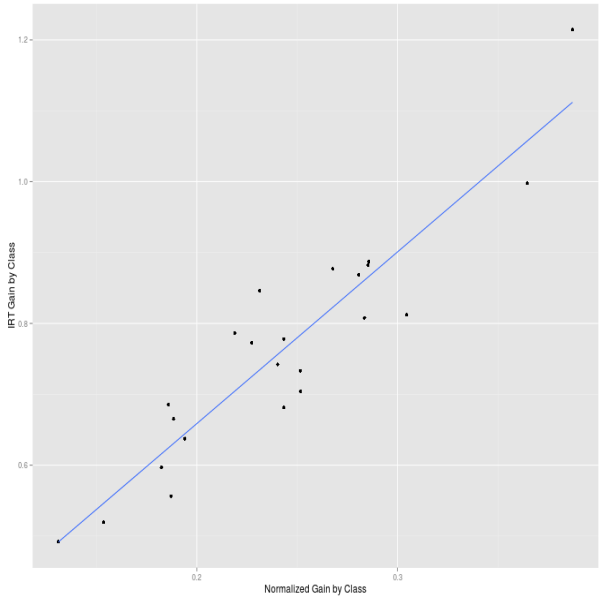

Figure 6: Normalized Gain vs. IRT Gain by Section

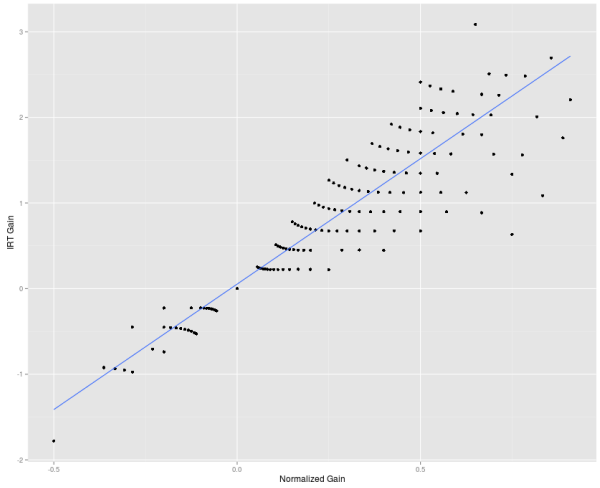

Figure 7: Individual Normalized Gains vs. IRT Gains