Testing Testing

Item Response Theory and Computer-based Testing

Rainier Raymond Robles
Seminar/Proseminar: Intelligente Lehr-/Lernsysteme
WS 17/18

The widespread availability and relative affordability of the personal computer, along with an alternative theory to testing, have opened up possibilities for creating shorter, individualized tests to measure students' ability. But will this new technique be widely accepted by educators?

Have you ever taken a test that was too simple for you? How about one that you felt was too difficult? More to the point, have you ever experienced taking a test or exam that didn't truly measure or reflect how much you actually learned in that particular subject?

Back in the day when the ratio of students per teacher was more manageable, educators were able to personally examine each student in order to determine their level of knowledge. Teachers could adjust their questions depending on the test taker's previous responses. If they sensed that the student was at a higher ability level, they could skip the simpler questions and delve deeper into the topics. On the other hand, if they felt that the student was at a lower ability level, then they could start off by testing how well they understood the basics of the topic.

Due to much larger numbers of students per teacher these days, it is no longer practical to conduct individual testing. Standardized tests that can be administered to many students at once have become the norm. This method of testing may be efficient in terms of time and money, but it is not without its downsides. Because the same questions are asked regardless of the examinee's level of ability, those on a higher level could feel they are unchallenged by the easier questions, and those on the lower level could feel that the exam is too difficult and not doable. The end result is the same for both of these extremes: They are less motivated to complete the exam, as they don't feel that their true ability levels are being measured.

Computer Adaptive Tests

In order to address this issue, Computer Adaptive Tests (CATs) were conceptualized in the 1970s. The idea, according to Howard Wainer and Robert J. Mislevy (qtd. in Conejo et al. 2004), "is to mimic automatically what a wise examiner would do." CATs are tests administered by a computer, where the selection of the items as well as the decision to terminate the test are based on the examinee's responses. Such a test starts with an initial estimation of the examinee's abilities. A question that would best confirm this estimation is then selected, and after the test taker answers, their ability level would be recalculated, and another question is selected based on this new estimation. This process repeats until a predetermined termination criterion is met.

There are several advantages to CATs in comparison to more traditional pen-and-paper tests. As mentioned before, CATs allow the examiner to administer individualized tests according to knowledge level, which then increases motivation in test takers. In addition, CATs lead to significantly shorter tests, which in turn leads to significantly shorter testing times. CATs also provide a more accurate estimation of test takers' proficiency levels.

The rise of the personal computer has made computer adaptive testing a viable option nowadays. In addition to the ability to store larger pools of questions as well as take advantage of multimedia content within tests, more powerful computers mean more efficient ways of implementing how CATs calculate and recalculate a test taker's level as well as selecting the next question.

Item Response Theory

Item Response Theory (IRT) is one of the more commonly used means of calculating examinee abilities in CATs. In order to define what IRT is, it would help to do so by contrasting it with Classical Test Theory (CTT). In CTT, one's proficiency level is determined by an aggregation of test responses. This is basically what happens in most tests in school or at a university level, where the total number of points or the percentage of correct answers determines your grade. On the other hand, individual responses to each item are more important in IRT, as these determine both the next question asked as well as the ability level of the test taker. Another difference between CTT and IRT is in how proficiency levels are portrayed. In CTT, a standardized achievement level is measured and produced, whereas a *probability* of one's proficiency level is measured and produced in IRT. This distinction will be important later on, when the wider usability and acceptability of CATs and IRT

is discussed.

"(T)he basic building block of item response theory" (Baker 2001) is the Item Characteristic Curve (ICC). The ICC is basically a graph which plots the probability $P(\theta)$ (a value between 0 and 1) that a test taker on a particular ability level θ can answer the question correctly. Each question on a test has its own ICC. Generally, IRT (and thus the ICC) takes into consideration an endless continuum of values, but for purposes of clarity, ability levels will be limited to a range between -3 and +3 in the sample graphs below.

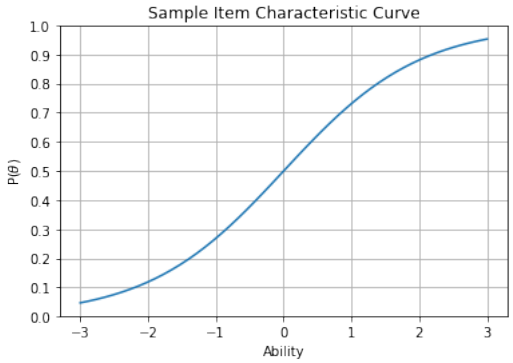


Figure 1: Sample ICC.

There are two main technical features of the ICC. The first is item difficulty, which describes the point at which an item is answerable. It generally corresponds to which proficiency level has a 50% chance of answering the question correctly.

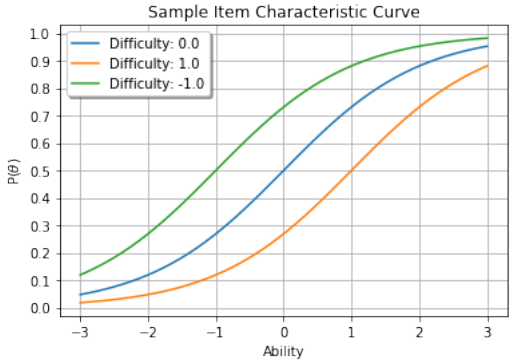


Figure 2: ICC for different difficulty levels.

The second technical feature is discrimination, which describes how well an item differentiates among test takers with abilities above and below the item's difficulty. Visually, changing the discrimination of an item affects the slope of the ICC. When the discrimination is decreased, then the slope of the curve is less steep, almost appearing as a straight line. On the other hand, when the discrimination is increased, then the slope becomes steeper.

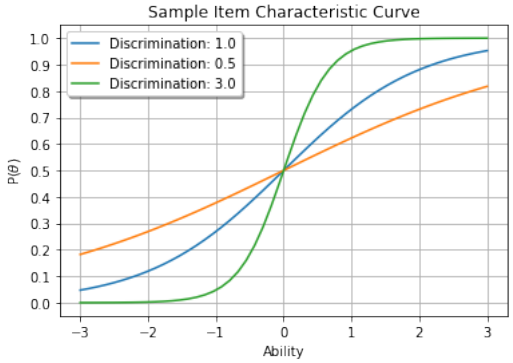


Figure 3: ICC for different discrimination levels.

How does one determine the graph for the ICC? More accurately, how does one compute for the probability of a correct answer given an ability level? One of the more common equations used to make these calculations is

called the three-parameter logistic model. This has the form

$$P(\theta) = c + (1 - c)\frac{1}{1 + e^{-a(\theta - b)}}$$

where θ is the ability level, $P(\theta)$ is the probability of someone at ability level θ answering the question correctly, b is the difficulty level, a is the discrimination, and c is the guessing parameter. The guessing parameter corresponds to the probability that someone can correctly answer the question purely by chance regardless of their actual proficiency level. One simple way of determining this guessing parameter in multiple choice questions is dividing the number of correct answers by the total number of options presented. For a true-or-false question, the guessing parameter would be 1/2 or 0.5, while for a multiple choice question with four options and one correct answer, this would be 1/4 or 0.25. (Note that with the addition of this parameter, the probability that a test taker at the same ability level as the difficulty of the item answers it correctly is no longer 0.5, but halfway between c and 1.)

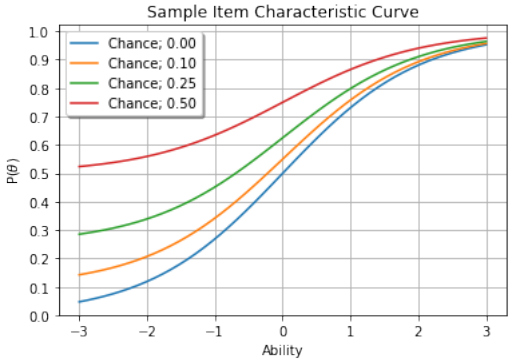
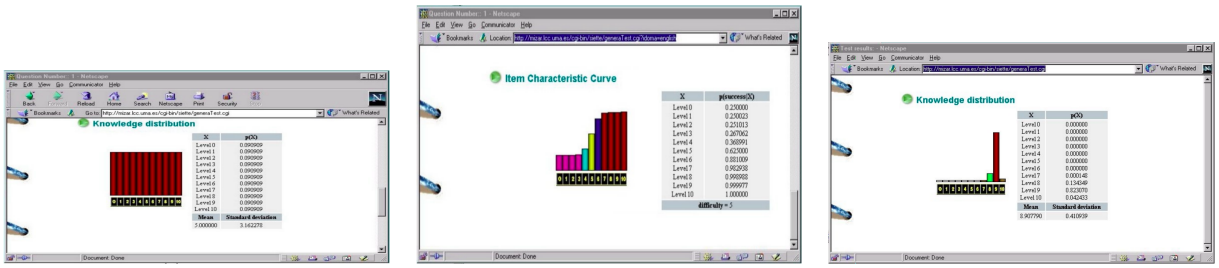


Figure 4: ICC for different guessing parameters.

Some other models use similar equations to calculate the probability but get rid of some parameters; for example, the two-parameter logistic model gets rid of the guessing factor completely.

SIETTE: A Case Study

One of the first tools to utilize CATs with IRT as its response model is SIETTE. SIETTE, which stands for SIsteme de Evaluación de Tests para la TeleEducación (in English, Intelligent Evaluation System using Tests for TeleEducation), was initially developed in 1998 by Antonia Rios as part of her master thesis at the Department of Computer Science at the University of Malaga, Spain. The tool started as an Intelligent Tutoring System (ITS) module as part of the TREE (TRaining of European Environmental Trainers and Technicians) project, whose goal is identification and classification of various species of vegetables in Europe. However, SIETTE was designed as a reusable, domain-independent module; indeed, later versions of the module have been integrated with the ActiveMath system as well as the learning management system (LMS) Moodle.



(a) Before the first question. (b) Sample item ICC. (c) After the test.

Figure 5: SIETTE at various states during a test. (Conejo, et al. 2004)

With SIETTE, teachers can design and develop tests as well as items for those tests. They can then set various parameters for them, including the ICC parameters for the test items. The system then generates adaptive tests automatically which the students can then take. One notable aspect of SIETTE is that it uses a discrete, finite range of ability levels, as opposed to the endless continuum of proficiency levels in regular IRT.

Ricardo Conejo and associates authored a study published in 2004 about evaluations they performed on the system. The goal was to test the validity of using discrete ability levels in IRT, as well as the usefulness of the

system itself. There were two evaluations, one with simulated students and one with real students.

For the evaluation with simulated students, N students were randomly generated for each of the K discrete ability levels (this K would depend on the experiment being conducted). Under the assumption that the test items were correctly calibrated, adaptive tests for each student were simulated on the system. These tests would terminate once the probability that one belonged to a particular ability level reached a confidence level ρ. If this ability level matched the ability level of the simulated student, then this would be considered a correctly classified student. In the experiments, the percentage of correctly classified students as well as the average number of questions asked were measured.

In the first experiment, the authors evaluated the number of ability levels K and the confidence level ρ. First of all, they found that it was impossible to correctly classify all of the simulated students. This was due to the nature of IRT itself, which says that even at high proficiency levels, there is still a probability that a test taker answers a question incorrectly. Second, they found that the percentage of correctly classified students was dependent on ρ; as this confidence factor was increased, so did this percentage. Finally, the average number of questions in a test was strongly correlated to the number of ability levels K; fewer levels meant fewer questions, and more levels meant more questions.

	Confidence factor $\rho = 0.75$		*Confidence factor $\rho = 0.90$*		*Confidence factor $\rho = 0.99$*	
Number of classes K	*% of correctly classified students*	*Average number of questions posed T*	*% of correctly classified students*	*Average number of questions posed T*	*% of correctly classified students*	*Average number of questions posed T*
3	84.05	2.00	95.82	3.58	99.46	5.65
5	81.61	6.23	92.76	10.38	99.37	19.27
7	80.96	11.11	92.85	18.16	99.38	33.12
9	80.86	16.15	92.93	26.39	99.42	47.27
11	80.52	21.19	92.92	34.54	99.26	60.85

Figure 6: Experiment 1 results. (Conejo, et al. 2004)

In the second experiment, the authors evaluated the discrimination a and the guessing factor c, fixing the number of proficiency levels at 7 and the confidence level at 0.9. Unsurprisingly, they found that they got the best results from items with a higher discrimination (meaning they could better differentiate students from different levels) and lower guessing factor (meaning a lower likelihood that students answered correctly purely by chance). While increasing a didn't lead to a significant decrease in the number of questions asked, a low a corresponded to an extremely high number of questions; for example, a discrimination of 0.20 required an average of almost 175 questions per test!

Guessing factor c	*% of correctly classified students*	*Average number of questions posed T*
0.00	92.85	18.16
0.10	92.37	25.34
0.25	92.11	36.05
0.33	91.73	43.37
0.50	91.49	63.37

Discrimination index a	*% of correctly classified students*	*Average number of questions posed*
0.20	90.4	174.9
0.50	91.5	35.2
0.70	91.9	26.3
1.20	92.8	18.1
1.70	93.8	15.3
2.20	95.4	14.8

Figure 7: Experiment 2 results. (Conejo, et al. 2004)

In the third and final experiment, the authors evaluated three item selection methods, two of which were adaptive and one which was completely random. They found that the adaptive methods performed better than random item selection, a result that was more evident when the number of levels K was increased. (Different discrimination and guessing parameters did not have a significant effect in this experiment.) In the end, the difficulty-based method, where an item that most closely matched the predicted ability level of the test takers was selected, was chosen between the two adaptive item selection methods because it was simpler to implement.

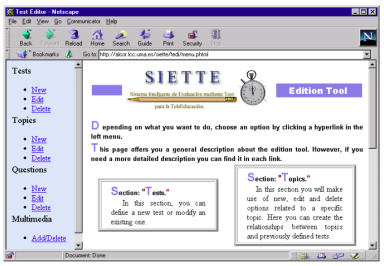
The evaluation on real students was conducted in June 2000 on 24 computer science students of the subject

	Bayesian		*Difficulty-based*		*Random*	
Number of classes K	*% of correctly classified students*	*Average number of questions posed T*	*% of correctly classified students*	*Average number of questions posed T*	*% of correctly classified students*	*Average number of questions posed T*
3	96.06	3.58	95.62	3.58	95.82	3.58
5	93.31	6.87	94.67	7.37	92.76	10.38
7	92.75	8.70	94.43	9.03	92.85	18.16
9	92.53	9.85	94.23	10.14	92.93	26.39
11	92.10	10.71	94.14	11.02	92.92	34.54

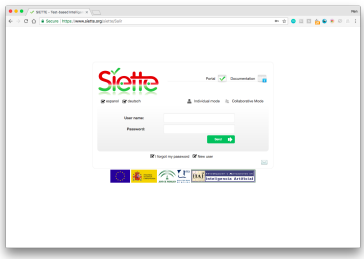
Figure 8: Experiment 3 results. (Conejo, et al. 2004)

Artificial Intelligence at the University of Malaga. These students had an exam in four weeks and were asked to take an online test about LISP on the SIETTE system then filled out a questionnaire. Participation in this evaluation was voluntary and anonymous.

According to the informal evaluation, a majority of the students needed less than five minutes to learn how to use the system, which meant that it was easy to learn. A majority also considered the system useful, as they considered it practical and a good complement to their lessons, and would recommend it to other students. Other positives were that the grade they received was considered accurate, and that the difficulty of the test was considered normal.



(a) In 2000. (Conejo, et al. 2004)



(b) SIETTE today.

Figure 9: SIETTE then and now.

The SIETTE system has evolved since the initial evaluations. In addition to integration with other systems, SIETTE is now not just a test-based system but an automatic assessment environment. This means that tests no longer need to be created on the system itself; tests created on Moodle, for example, can take advantage of SIETTE's assessment system. In addition, other assessment methods apart from IRT have been added, including more traditional percentage scoring.

Issues and Challenges

Three of the authors of the original study published a follow-up paper in 2015 about the current usage of SIETTE. They found that only 18 tests used IRT as the assessment method, compared to 642 which used percentage scoring and 581 which used item scoring. This meant that, despite the advantages of IRT such as shorter tests, a majority of tests still used Classical Test Theory for student assessment. Why is this the case?

Conejo and his colleagues put forward several hypotheses as to why IRT isn't as widely used. First, they conceded that there was a difficulty in understanding and ultimately accepting some of the underlying concepts behind CATs and IRT. Both teachers and students need to be educated but this may take considerable time and effort. Among the biggest hurdles is the idea that different students receive different tests with different questions and, ultimately, different levels of difficulty, despite studies that show that the estimated proficiency level highly correlates with the actual proficiency level.

Related to this is a general resistance to change on every level of the education system. Even if students, educators, and policy makers took the time to learn the ins and outs of CATs and IRTs, many would not be comfortable with innovation and changing the traditional way of doing things. Apart from the adaptiveness of the tests themselves, another challenging aspect to change would be reporting a student's proficiency level as a probability as opposed to being expressed proficiency as a certainty. In David Thissen's 2016 essay *Bad*

Questions: An Essay Involving Item Response Theory, while he considers this as an opportunity to educate teachers and students on the probabilistic nature of measuring proficiency, he acknowledges the danger that policy makers would be less flexible and instead keep turning to methods that provide them the certainty they seek.

Finally, one of the biggest challenges to the usage of Item Response Theory is item calibration. In the 2000 evaluation of SIETTE with simulated students, it was assumed that the items were calibrated correctly, meaning that the ICC parameters (difficulty and discrimination) set by the teacher were valid and accurate. However, this is not necessarily the case. While Frank B. Baker suggests a method for estimating item parameters in his book *The Basics of Item Response Theory*, Thissen notes in his essay that asking if the IRT model for a question fits is a "bad question" because "(n)o model ever fits." In addition, accurately modeling and calibrating an item requires a lot of data; at least 200-1000 data points would be required to be able to start accurate calibration of an item, whereas most tests on SIETTE are administered to classes smaller than 40. And while CATs do provide an opportunity to collect and analyze this data, the nature of CATs and IRT makes it possible that some questions, particularly those on the extremes of the ability scale, would be asked less frequently in tests and would therefore mean less data for analysis of the item.

The Future of Testing

Computer Adaptive Tests and Item Response Theory provide a valid alternative to more traditional pen-and-paper tests and Classical Test Theory. Indeed, CATs and IRT can reduce test length and test time, while still providing results that are accurate and comparable to standardized tests. However, before such methods reach widespread acceptance, members of every level of the education system need to be educated about it. Furthermore, more effective and efficient ways of item difficulty estimation and item calibration should be studied and implemented. Then we can return to the days of the "wise examiner" by utilizing modern methods and technologies.

References

Baker, Frank B. *The Basics of Item Response Theory.* 2nd ed., ERIC Clearinghouse on Assessment and Evaluation, 2001, 1-57.

Barla, Michal, et al. "On the impact of adaptive test question selection for learning efficiency." *Computers & Education* 55.2 (2010): 846-857.

Conejo, Ricardo, et al. "SIETTE: A web-based tool for adaptive testing." *International Journal of Artificial Intelligence in Education* 14.1 (2004): 29-61.

Conejo, Ricardo, Eduardo Guzmán, and Monica Trella. "The SIETTE automatic assessment environment." *International Journal of Artificial Intelligence in Education* 26.1 (2016): 270-292.

Thissen, David. "Bad questions: An essay involving item response theory." *Journal of Educational and Behavioral Statistics* 41.1 (2016): 81-89.

Wauters, Kelly, Piet Desmet, and Wim Van Den Noortgate. "Adaptive item-based learning environments based on the item response theory: possibilities and challenges." *Journal of Computer Assisted Learning* 26.6 (2010): 549-562.