What is Adaptive Testing?

A conventional (often mass-administered) test must be designed to assess a broad range of ability and contain items for those students of lower-ability, those of higher ability and everyone in between. Because we know that there are more students in the middle of the ability range (the average examinee), a conventional test tends to be designed to better measure the ability of those examinees as it contains more items of moderate difficulty. This makes sense as most students fall into this range and so it is an efficient way to assess a large number of students. As a result, a conventional test contains few items that are truly appropriate for students who fall outside of the middle range and most of the items are of moderate difficulty and are appropriate for the average examinee (Wainer, 2000).

A student with an ability level at one extreme or another will not be measured in an efficient or useful way by a conventional test designed to better assess the average student. Moreover, higher-ability students will have to trudge through many easy items which provide no information on their ability, while lower-ability students may become frustrated with items far beyond their range of competency. With the latter group, there is also the possibility of guessing which can make the analysis and interpretation of data more messy (Wainer, 2000).

Computerized adaptive testing (CAT) solves these problems by selecting items that directly target the observed level of ability of each student. An effective CAT can quickly identify and administer items that are the most informative at a particular ability level, and each CAT is specifically matched to test taker. This is achieved by first asking a student a question at a moderate difficulty level and then if the student answers this question correctly, the next question is more difficult. If that question is answered incorrectly, then the next question is easier. This pattern continues until the computer is able to make an estimate on the proficiency of the examinee based on a predetermined level of accuracy (Wainer, 2000). This results in a much more efficient testing environment and one that yields much more precise measurement information, and also may allow one to administer a shorter test (Davey & Pitoniak, 2006; Way, Davis, & Fitzpatrick, 2006; Wainer, 2000).

There are two basic types of CAT:

a) Fixed length: each examinee sees the same number of items. These tests more closely mirror a more traditional test as examinees who are best targeted by the items in the test, are measured in a more precise way than those students who are less well-targeted by the items in the pool.

b) Variable length: these tests keep serving up items to examinees until a pre-determined level of precision is achieved—examples of variable length CATs are the NCLEX and GRE.
Critical to adaptive testing is the fact that if the underlying algorithms are working correctly, then students see questions which are targeted to their skill and knowledge level and one would expect students to get about half of the questions correct—regardless of ability. Higher-ability students should get about half of a more difficult set of questions correct, and the lower-ability students about half of a less difficult set of questions correct. In either case, the percentage alone is not informative when looking at student proficiency. A score on an adaptive test depends both on how many questions the student answered correctly and the difficulty of the items. And the difficulty of the items can only be predetermined if there is a pilot testing period during which data on items is collected (the section on IRT below expands on these ideas).

**Components of a CAT**

In order to be truly adaptive, a CAT requires some basic components:

a) Item pool—a collection of calibrated items which is large enough that any one examinee only sees a small fraction of the items.

b) Decision rule for selecting the first item (or item set)—answers to the first item (or set) are used to determine the initial estimate of the test taker’s ability level.

c) Methods for selecting additional items or item sets—responses to each set help refine the performance estimate in an ongoing way. Items are neither too difficult nor too easy for the test taker.

d) Items are selected to maximize efficiency, in this case meaning that a test taker can be measured with a high level of precision on a small number of test items.

e) Test items are balanced in terms of content—test specifications, similar to those used to create a conventional test are adhered to in creating the pool of items.

f) Termination criteria—the constraints which are used to decide when the test taker can end the test (Davey and Pitoniak, 2006; Parshall et al., 2002).

While there is no simple equation for the number of items needed for a CAT system, the most commonly cited rule of thumb for the number of items required is anywhere from the equivalent of 5-10 conventional test forms (Davey and Pitoniak, 2006; Parshall et al., 2002; Stocking, 1994). These numbers are based on a fixed-length test. So if you were designing a fixed length CAT with 30 items, you would need a pool of around 300 items (upper limit).

Other factors that influence this number are:

a) coverage of content: the larger the breadth of content, the larger the number of items required.

b) examinee population: a larger range of proficiency will require a larger item pool that a more homogeneous population. There must be enough items at each difficulty level, and so you need more items to effectively cover a wider range of proficiency levels.

c) test security: the smaller the item pool, the more frequently one item will be used. In some cases, having the same items appearing more frequently may be a security concern.

**Advantages of CAT**

Some advantages of adaptive tests are:

a) Fewer items are needed for each student which can save testing time

b) Test security is increased as each test taker sees different sets of items and in different sequences.

c) Because the item pool is large, there is less chance of a student memorizing or learning a few items and this providing them a benefit on the actual test.

d) Students can work at their own pace and if desired, the speed at which a student completes a test, or answers a question can be used as additional data. Clearly there has to be some limit to the time allowed, but there is more room for flexibility.

e) Students are productive and hopefully not bored or discouraged as the items are focused appropriately. Fewer difficult items for low-ability students may serve to increase motivation as test takers feel as if they are doing well. The same is true for higher-ability students—they will see fewer easy items which will make them feel as if their time is being used wisely

f) The removal of a paper and pencil test and/or bubble sheet removes any possible ambiguities based on misread marks, or erasures. Also, the tests cannot get lost or damaged.

g) A reduction in testing time can result in increased student motivation.

h) Results are typically available immediately which gives instructors the opportunity to use results to inform instruction and remediation more quickly.

i) New items can be pretested/calibrated easily by incorporating them into the test.

j) Poor items can be identified and edited/removed easily.

(Thompson, 2010; Wainer, 2000; Weiss & Kingsbury, 1984; Green, 1983)

**Item Response Theory**

Item Response Theory, or IRT, is the psychometric theory which essentially characterizes what happens when items and individuals meet (Wainer, 2000). In a chapter on IRT, calibration and estimation, Wainer and Mislevy (2000) use the analogy of a high-jumper to describe how item selection in adaptive test functions. In a high-jump
In adaptive testing, item selection suffers from the same issues we see in the high-jump situation: We want to give each person items that will be the most informative and that will allow us to make the most precise measurement of ability given the amount of time given for the testing. We don’t want to waste the really good high-jumpers time having them jump 4 ½ foot bars. But characterizing the difficulty of the items is complicated. The challenges that face adaptive testing are:

a) Variations in items must be characterized in a useful way.
b) Rules for selecting items must be established.
c) Scores must be on a common scale, even though individuals have taken different tests at different times.

In adaptive testing, item selection suffers from the same issues we see in the high-jump situation: We want to give each person items that will be the most informative and that will allow us to make the most precise measurement of ability given the amount of time given for the testing. We don’t want to waste the really good high-jumpers time having them jump 4 ½ foot bars. But characterizing the difficulty of the items is complicated. The challenges that face adaptive testing are:

a) Variations in items must be characterized in a useful way.
b) Rules for selecting items must be established.
c) Scores must be on a common scale, even though individuals have taken different tests at different times.

Item Response Theory (IRT) is able to address all of these issues in adaptive testing. IRT is a model that allows for examinee A, to be shown a particular item taking into account parameters for that item and also the examinee.

IRT can address the challenges above in the following ways:

a) Variations in items must be characterized in a useful way.

IRT: Responses to items are collected and an IRT model fit to these responses. This process is called calibration. Once this has been done, the performance characteristics of each item are specified by that item’s parameters.

b) Rules for selecting items must be established.

IRT: once you have a rough estimate of the ability of an individual, the IRT model can help one determine which test item would be the most informative. If someone jumps 7ft, you wouldn’t then want to lower the bar to 3 ft. You wouldn’t learn anything useful from that exercise, but you might want to raise the bar to 8ft and see what happens then.

c) Scores must be on a common scale, even though individuals have taken different tests at different times.

IRT: once the items have been calibrated, then the proficiency of an examinee can be characterized through the parameters of the questions the examinee answers correctly. Using the high-jumper analogy, we could characterize the ability of a jumper based on the greatest number of feet he could jump.

(Wainer and Mislevy, 2000)

IRT helps us describe what happens when an examinee sees a test item. All items ought to measure the same thing, and IRT makes this formal in the sense that it explicitly posits a single dimension of knowledge or trait on which all of the test items rely, to some extent, for their correct response. A trait, in this case, might be jumping ability, or verbal reasoning skill, or memory.

Mastery-Based Learning

Mastery learning (in existence since the 1920s) refers to instructional methods which, broadly speaking, involve the establishment of a criterion performance level which must be met in order to demonstrate proficiency or mastery of a concept, skill or educational objective (Bloom, 1968; Slavin, 1987). Frequent assessment and corrective instruction are also common components of mastery-learning methods. In a mastery-learning context, students may be presented with units of instruction and are required to demonstrate mastery on an outcome measure before they can proceed to the next step, or unit. The idea is that all students should be able to master content given enough time, resources and the right conditions for learning. And so progression through a course, or unit depends on the students’ demonstration of proficiency rather than the amount of time spent on the work (Slavin, 1987; Kulik, Kulik and Bangert-Drowns, 1990).

Mastery learning became more visible in the educational field in the 1960s with two main approaches emerging as influential early on: a) Bloom’s Learning for Mastery (BLM or LFM), b) Keller’s Personalized System of Instruction (PSI). Under both approaches, students learn material divided into units and then take assessments (intended to be used formatively) on the material in each unit (Bloom, 1968; Keller, 1968).

The BLM approach is most often seen in elementary and secondary schools and involves students being led through the material by a teacher, and so the rate is controlled by the pace of the teacher. At some point students are given a formative assessment with a pre-established mastery criteria (usually 80-90% correct). If students do not meet the criteria they receive additional help (either individually or in a group) before moving on. Key to the approach is that the
corrective instruction should not be the same as the original instruction. The extra help is followed by another test (parallel to the original one) and then the class as a whole moves on—even if there are some students still not at the mastery threshold, or students receive more instruction (Slavin, 1987).

Under the PSI approach, students use written materials for the most part and their pace is determined by their own progress. This approach is more often seen at the post-secondary level of instruction and is sued along with group work and teacher lectures. Students can take tests multiple times until they have achieved mastery. Tests may be the same, or parallel forms. Students do not progress to the next unit until they have demonstrated mastery of the current one (Kulik, Kulik and Bangert-Drowns, 1990).

**Does Mastery Learning Work? What is the Evidence Efficacy?**

The literature is mixed on the efficacy and impact on mastery learning models. Kulik, Kulik and Bangert-Drowns (1990) published a meta-analysis of the research on mastery learning. They concluded that mastery-learning programs did have a positive impact on students in college, high school and upper elementary classes. They also concluded that the impact was greater for the lower-performing students in the class. Another finding was that while mastery programs seemed to have positive effects on student attitudes towards coursework and instruction, the amount of time spent in class increased, and in some college-courses this impacted course completion (see Kulik, et al., 1990). This meta-analysis (and an earlier article by the same authors—Kulik, Kulik and Bangert-Drowns, 1986) was critiqued by Slavin (1990) who pointed out that the large effect sizes reported came from studies where experimenter-made tests (rather than standardized tests were used). This choice of outcome potentially biases results as one would expect an experimental group to perform better on an outcomes measure which is well-aligned to the educational objectives of a course. Thus, if tests used in studies are created by the experimenter, there is a clear potential bias toward the objectives taught in the experimental group and not the control group (which may be unknown). This issue can be overcome by using a standardized test as an outcome measure. Based on the research, however, when standardized tests are used, most studies found no impact of mastery-learning programs (Slavin, 1990).

Slavin conducted his own meta-analysis on mastery learning programs using more stringent inclusion criteria than some of the early analyses. Inclusion criteria were:

- a) Students were tested on the mastery of instruction objectives at least once in every four-week period.
- b) Prior to each test students were taught as a group (this excluded studies using PSI (described above).
- c) Mastery learning was the only (or the most salient) intervention.
- d) Intervention classrooms were compared to other classrooms using traditional group-paced instruction.
- e) Evidence that treatment and control groups were essentially equivalent at the outset (demonstrated, for example, by performance on a pretest)
- f) Study duration had to be at least four weeks
- g) At least two experimental and two control classes had to be involved.
- h) The outcome measure had to be an assessments of objectives taught in both conditions.

Based on a review of studies which satisfied the inclusion criteria, Slavin (1987) found no impact of “group-based mastery learning” in studies using standardized achievement measures as outcomes. When experimenter-created measures were used, effect sizes tended to be positive and moderate in size. There are many possible explanations for these findings.

First of all, it might be that teachers do not have enough materials, or even the skills and experience, to provide effective corrective instruction for their students. Or, the amount of extra instruction may not be enough for some students—particularly low-achieving students. This explanation is supported by findings from a formative assessment-based math intervention project led by the author of this review (Phelan, et al., 2011). In that study we saw benefits of an intervention designed to help teachers use data from short formative assessment measures to determine which students needed more instruction, and also support teacher provision of that instruction. The intervention, however, had more impact on higher performing students than lower performing students. It may be that the time allotted to our intervention was not sufficient to effect change in lower performing students, even thought we did see an impact with higher performing students (Phelan et al., 2011). Many other issues may exist when it comes to giving some students in the class corrective instruction: for some students, the instruction may not be enough for some students—particularly low-achieving students. This explanation is supported by findings from a formative assessment-based math intervention project led by the author of this review (Phelan, et al., 2011). In that study we saw benefits of an intervention designed to help teachers use data from short formative assessment measures to determine which students needed more instruction, and also support teacher provision of that instruction. The intervention, however, had more impact on higher performing students than lower performing students. It may be that the time allotted to our intervention was not sufficient to effect change in lower performing students, even thought we did see an impact with higher performing students (Phelan et al., 2011). Many other issues may exist when it comes to giving some students in the class corrective instruction: for some students, the instruction may not go deeply enough to remedy underlying causes of poor performance and the timing and/or frequency of the corrective instruction may have an impact on student outcomes as well (Slavin, 1987). In summary, there is little evidence that large learning gains are present when using mastery-based approaches in a classroom context. This is likely because it is incredibly difficulty to deal effectively with the inherent differences present within a mixed-ability group in terms of how quickly they achieve mastery.
Post-secondary uses of mastery-based learning principles: CAI

As mentioned above, the mastery-based learning approach most common in the post-secondary context is Keller’s Personalized System of Instruction (PSI). Different from the group-focused BLM approach, this approach focuses on the individual progressing at their own pace—albeit with classroom lectures and/or other instructional components as part of the process.

An application of mastery-learning concepts which is becoming more and more popular is individualized computer-assisted learning (CAL) also known as CAI (computer-assisted instruction). In CAI, a computer is used to both deliver instruction and assessment to a student who progresses through them at their own pace. Studies on the efficacy of CAI have been conducted on students exposed to CAI in myriad areas (science, math, pre-service teaching and so on). A review of literature on individual studies, and other meta-analyses involving college age students suggest that CAI is an effective mode of instruction (Jenks & Springer, 2002). No studies, however, make the unequivocal claim that CAI is necessarily better than typical instruction, rather it seems to be at least as effective. In studies where the same teacher either delivered traditional instruction, or designed some CAI, there were no differences in student performance between the two conditions. Furthermore, the literature suggests that CAI is best delivered as an augmentation to traditional instruction rather than as a replacement (Lowe, 2001). This is typically how it CAI is used and the combination of both types of instruction is known as blended learning. It is perhaps intuitive, but the evidence suggests that the quality and design of instruction is likely more important than how that material is ultimately delivered.

Incorporating Computer-Based Activities (Using Assessment as a Learning Activity)

As described above, although mastery-based educational approaches are appealing and hold hopes for leveling the educational playing field for all students (regardless of ability) implementing these approaches can be extremely difficult. Practical limitations tend to render these practices less feasible for some teachers and in many cases not scalable or appropriate for all students or classroom situations. The more individualized approach seems to hold more promise—particularly with post-secondary-aged students. CAI is one application in which traditional instruction can potentially be augmented =by a mastery-based approach, which can afford some of the benefits of self-paced learning to students in need.

With the advent of more web-based educational technologies, course management systems etc. the incorporation of computer-based activities also becomes more feasible on a broader scale. One such activity is the online quiz which has been used in multiple disciplines with varying types of implementation. A review of the literature on online quizzing reveals different considerations which can be made, or parameters which can be set, when implementing online quizzes. These differences which potentially can impact how beneficial these quizzes include:

a) Time limits
b) Credit and the amount of credit
c) Is credit based on completion or achievement of some mastery criteria
d) Are the exercises optional?

The goals of many online practice questions or quizzes are to:
• encourage and enhance preparation for class meetings
• provide teachers and students with information on content mastery
• increase scores on summative exams
• provide instructors access to data on student usage patterns (frequency and number of questions answered, for example).
• Provide students information on their strengths and weaknesses related to course concepts

Online assessment tools have been used and studied with an eye to determining how they can best be utilized in out of class environments. Researchers have investigated levels of student usage and also ways to encourage usage of these tools when available to students. Participation rates ranging from between 40-and 52% are found when online quizzes were simply made available to students (Grabe & Sigler, 2001; Olson & McDonald, 2004; Kibble, 2007). In these studies, use of the online quizzes also resulted in an increase in student performance on the relevant summative class exams (Grabe & Sigler, 2001; Olson & McDonald, 2004; Kibble, 2007).

Clearly, providing students access to use online quizzing tools is not motivating for everyone. Not all students take advantage of these tools or see the potential benefits of use (Brothen, Daniel & Finley, 2004). Some have investigated making online quizzing a course requirement, or offering incentives such as extra credit for participating in online quiz-taking to encourage more students to become familiar with and begin to use these tools. Results indicate:

• offering small amounts of course credit and introducing performance criteria to earn credit increased student participation (Kibble, 2007), possibly by increasing the incentive to overcome the time and effort costs associated with learning how to access and use a new electronic tool.
• offering credit for taking weekly quizzes based on reading material and number of quizzes completed predicted higher course grades and also exam performance (Johnson & Kiviniemi, 2009)

Most of the extant research on online quizzing recognizes both the potential benefits to student learning and the possibilities for students to attempt “gaming” the system in order to get the correct answers. If the effectiveness of online quizzing tools is increased by helping students to use them in the right way, and ensuring that the tools actually encourage usage towards mastery.
References


