Psicológica (2000) 121-133.

CATs: Whither and whence

Howard Wainer*
Educational Testing Service

In this essay I sketch the background that gave rise to adaptive testing and frame a discussion of CAT's progress around Bert Green's expectations of the advantages of this technology. Data from the first decade of operational CATs are used to compare what has happened to what was hoped for. I find that some of the goals for CAT that Green expressed are close to being accomplished, but that most of them remain in the future.

Key words: adaptive testing, CAT's progress, CAT's advantages

Throughout its entire history there has always been the tradeoff between individual testing and group testing. An individually administered test does not contain too many inappropriately chosen items and, furthermore, we are assured that the examinee understands the task. A group-administered test has the advantage of uniformity of situation for all examinees, as well as a vastly reduced cost of testing. Throughout the first 90 years of the 20th century, the choice has almost always been in favor of the mass-administered test.

A critical problem facing a mass-administered test is that it must be assumed that there is a relatively broad range of ability to be tested. To effectively measure everyone, the test must contain items whose difficulties match this range (i.e., some easy items for the less proficient, some difficult ones for the more proficient). If the test did not have difficult items, we might not, for example, be able to distinguish among the proficient examinees who got all the easy items correct. Similarly, if there were no very easy items on the test, we might not be able to distinguish among the less proficient examinees who got the more moderate items all wrong. If making these kinds of discriminations is important, the test must contain as broad a

* This research was supported by the ETS research allocation and I am pleased to have the opportunity to acknowledge it. In addition, I would like to thank Karen Copper, Penni Davis-Tantum, Valerie Folk, Alan Nicewander and Pam Rice for providing me with the CAT volumes for TOEFL, GRE, ASVAB and GMAT respectively. These data were used to construct Figure 1. My gratitude also to the many other ETSers the results of whose unpublished work on the details of CAT provided the facts behind the judgments I made.

range of item difficulties as the ability range of the population to be tested. The accuracy with which a test measures at any particular proficiency level is (roughly) proportional to the number of items whose difficulties match that level.

Fortunately for mass-administered testing, Lincoln's observation that "the good Lord must have loved the common man because he made so many of them" remains valid. Most examinees' abilities seem to lie in the middle of the continuum. Thus, mass-administered tests match this by having most of their items of moderate difficulty with fewer items at the extremes.

The consequence of this test structure has historically been that the most proficient examinees have had to wade through substantial numbers of too easy items before reaching any that provided substantial amounts of information about their ability. This was wasteful of time and effort as well as introducing possibly extraneous variables into the measurement process, for instance, the chance of careless errors induced by boredom. Less proficient examinees face a different problem. For them, the easy items provide a reasonable test of ability, whereas the difficult ones yield little information to the examiner. They can, however, cause confusion, bewilderment, and frustration to the examinee. They also add the possibility of guessing, which injects extraneous noise into the measurement process.

In the early 1970s, the possibility of a flexible mass-administered test that would alleviate these problems began to suggest itself. The pioneering work of Frederic Lord (1970, 1971a,b,c,d) is of particular importance. He worked out both the theoretical structure of a mass-administered, but individually tailored test, as well as many of the practical details.

The basic notion of an adaptive test is to mimic automatically what a wise examiner would do. Specifically, if an examiner asked a question that turned out to be too difficult for the examinee, the next question asked would be considerably easier. This stems from the observation that we learn little about an individual's ability if we persist in asking questions that are far too difficult or far too easy for that individual. We learn the most when we accurately direct our questions at the same level as the examinee's ability. An adaptive test first asks a question in the middle of the prospective ability range. If it is answered correctly, the next question asked is more difficult. If it is incorrectly answered, the next one is easier. This continues until we have established the examinee's ability to within some predetermined level of accuracy.

Early attempts to implement adaptive tests were clumsy and/or expensive. The military, through various agents (e.g., Office of Naval Research; Navy Personnel Research and Development Center; Air Force

Human Resources Laboratory; Army Research Institute) recognized early on the potential benefits of adaptive testing and supported extensive theoretical research efforts. Through this process much of the psychometric machinery needed for adaptive testing was built. Nevertheless, the first real opportunity to try this out in a serious way awaited the availability of cheap, high-powered computing. The 1980s saw this and the program to develop and implement a computerized adaptive test (CAT) began in earnest (see Sands, et al, 1997, for a detailed description of the development of the CAT-ASVAB, and Wainer et al, 2000 for a reasonably up-to-date textbook on CAT).

This work was aimed at improving the entire measurement process. In addition to the increased efficiency of testing the other advantages expected of a CAT (from Green, 1983) were:

1. **Improved test security**, to the extent that a test is safer in a computer than in a desk drawer. Moreover, because what is contained in the computer is the item pool, rather than merely those specific items that will make up the examinee‘s test, it is more difficult to artificially boost one’s score by merely learning a few items. This is analogous to making available a dictionary to a student prior to a spelling test and saying, “All the items of the test are in here.” If the student can learn all of the items, the student’s score is well earned.

2. **Individuals can work at their own pace**, and the speed of response can be used as additional information in assessing proficiency. Aside from the practical necessity of having rough limits on the time of testing (even testing centers must close up and clean the floors occasionally), we can allow for a much wider range of response styles than is practical with traditional standardized tests.

3. **Each individual stays busy productively** — everyone is challenged but not discouraged. Most items are focused at an appropriate range of difficulty for each individual examinee.

4. **The physical problems of answer sheets are solved**. No longer would a person's score be compromised because the truck carrying the answer sheets overturned in a flash flood — or other such calamity. There is no ambiguity about erasures, no problems with response alternatives being marked unwittingly.

5. **The test can be scored immediately**, providing immediate feedback for the student. This has profound implications for using tests diagnostically.

6. **Pretesting items can be easily accomplished** by having the computer slip new items unobtrusively into the sequence. Methods for doing this effectively are still under development.

7. **Faulty items can be immediately expunged**, and an allowance for examinee questioning can be made.

8. **A greater variety of questions** can be included in the test builder's kit. The multiple-choice format need not be adhered to completely — numerical answers to arithmetic problems can just be typed in. Memory can be tested by use of successive frames. With voice synthesizers, we can include a spelling test, as well as aural comprehension of spoken language. Video disks showing situations can replace long-winded explanations on police or firefighter exams.

THE PRESENT

With such convincing cheerleading, it is no wonder that the actual use of computerized testing for operational tests took off in the decade of the 1990s. In Figure 1 are shown (on a logarithmic scale) the number of CATs given in four testing programs: the Graduate Record Exam (GRE), Graduate Management Admissions Test (GMAT), the Test Of English as a Foreign Language (TOEFL), and the Armed Services Vocational Aptitude Battery (ASVAB). These four tests constitute four of the largest operational testing programs that have "gone CAT." We see that in 1990 only a few hundred CATs were administered, but by 1999 this figure had grown to more than a million. The growth over this decade was exponential and while it is hard to predict how much longer it will remain so, it is clear that CAT utilization is a long way from leveling off.

At the same time that CAT utilization has been booming there has been a movement toward "distance learning."[1] The idea of using internet technology to reach distant students is the latest attempt to spread the scarce resource of first class education more broadly than is possible within the bounds of face-to-face instruction. On-line internet instruction is the 21st century version of a 20th century correspondence course. But when the student is at-a-distance how can we measure the efficacy of the instruction? How much has the student learned? Correspondence courses would often include extensive written exercises and exams that would be mailed in for

[1] Until secure and valid "distance assessment" is operational, it is probably more accurately called "distance teaching," or more honestly, "distant students."

teacher evaluation. It is natural to think that if the course was provided electronically, over the internet, so too would be the evaluation. And, if a computer is administering the test, efficiency would suggest that it might as well be made adaptive. With this scenario of utilization looming on the near horizon, how could anyone doubt the bright future of CATs? The issue yet to be resolved is "how can we know at-a-distance who is answering the questions?"

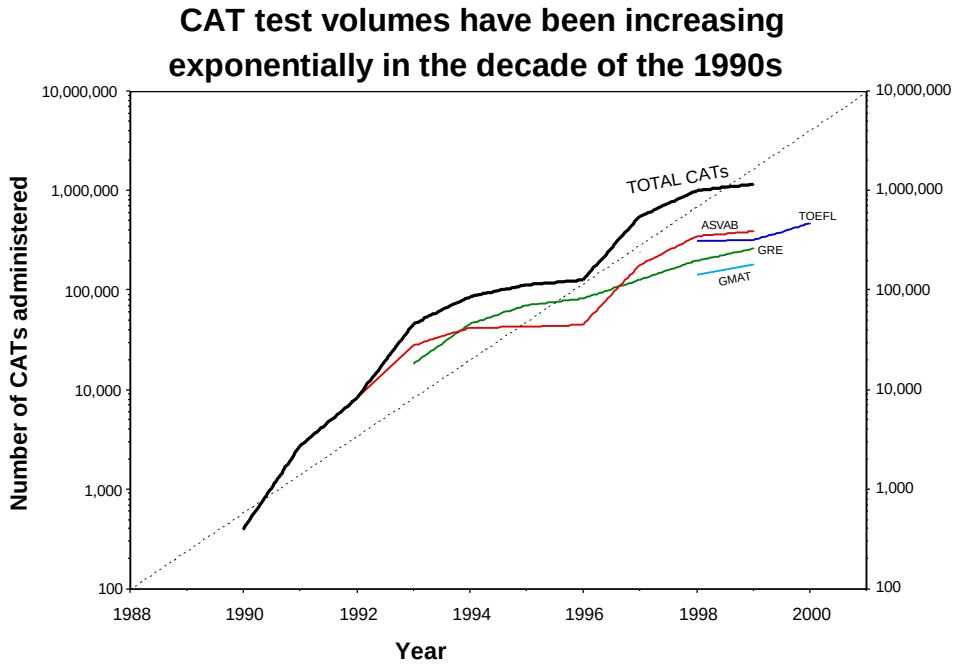

Figure 1. The total number of computer administered GRE, GMAT, TOEFL and ASVAB tests given annually since 1990. The exponential growth of CAT shows up as looking linear on the log scale.

The administration of more than a million CATs a year becomes even more impressive when one considers the circumstances under which a CAT is administered. Most typically it is done in a small room with no more than 8 to 10 testing stations, each in a separate cubicle, overseen by a test administrator. The administrator has a monitor that allows him/her to see what each examinee is doing. Compare the cost of such a set-up with the more familiar situation for mass administration of tests in which a gymnasium is filled with desks and a couple of proctors roam the room keeping an eye out for improper behavior. Typically a measure of test

security is added through the use of two or three different forms of the same test that are "spiraled[2]" throughout the examinees in the room.

"A pessimist is an optimist with data."
Linda Steinberg, 1999

With the administration of more than three million operational CATs, the decade of the 1990s has provided us with an enormous amount of testing information. The importance of the enterprise also has had the effect of increasing the closeness with which those data were scrutinized. This examination revealed practical limitations to the technology that were not apparent earlier. As the glow of initial enthusiasm faded and as our eyes became accustomed to the darker reality, previously unsuspected problems emerged. With our increasing awareness of practical limitations has come the requirement that we reevaluate old assumptions.

The questions we now must address deal less with "how to use it?" but more often "under what circumstances and for what purposes should we use it?" The future surely holds a promise for the possibilities of testing that are hard to foresee, but tests will still need to fulfill the age old canons of validity that characterize good practice. Test security remains an essential element for the validity of most tests, and how to maintain security at-a-distance remains an unsolved problem.

Let us reconsider Green's eight points with the wisdom of both data and hindsight.

1. **Test security**. Current economic realities mean that CATs are given continuously. Thus the item pool is constantly being exposed. In addition, the CAT item selection algorithm does not choose all items with equal likelihood. In fact, a very small proportion of the item pool accounts for a large proportion of the items administered (Wainer, 2000); a common finding is that between 15 and 20 percent of the item pool accounts for more than 50% of the test items. Thus, although we might provide a dictionary as the corpus of items for a spelling test, the item selection algorithm would choose some words much more often than others (Zipf, 1949). Hence the effective size of the item pool is much smaller than the actual size. This is an enormous problem since test security seems to increase logarithmically with item pool size. Since item writing costs are linear with pool size, this means

[2] "Spiraled" is the term that is often used to describe the process of interleaving different test forms in the shipping box so that when they are passed out to examinees people sitting next to one another do not have the same test form. This makes copying from your neighbor's test futile.

that costs increase exponentially with linear increases in test volume. This contrasts sharply with the economics of mass administered, paper and pencil tests, in which costs decline with increased volume; indeed the marginal cost of a paper and pencil test goes almost to zero.

Some help in this is hoped for through the development of methods for automated item generation (Irvine & Kyllonen, 2000). These seem promising for some areas of testing (e.g. verbal reasoning, spelling, arithmetic), but are more problematic in others (e.g. history, chemistry, Spanish literature).

2. **Individuals can work at their own pace**. But seat time at a computer is expensive and so time limits have not been overly generous. In 1997 CAT testing at one ETS program found that so many examinees did not finish the exam that "the rule of 80%" was instituted. This rule stated that if the examinee finished at least 80% of the items that were to have been administered, their score would be computed. Less than 80% and they would have to retake the exam. Subsequent research showed that examinees that finished the test would, on average, have had a higher score if only the first 80% of the items were scored. Because it did not seem fair to penalize students who actually finished the test, the "rule of 80%" was rescinded. It was replaced by a version of what happens on a fixed-format test when you don't finish; the items unanswered are counted as wrong. Nevertheless, while theoretically examinees can have as much time as they might need, the substantial cost of testing time means that compromises must be made.

3. **Each individual stays busy productively.** This appears to be somewhat true. However the limitations of finite item pools, when coupled with the requirement that the test must span a broad set of content specifications, means that psychometric optimality must too often be sacrificed. Sometimes the only items available for a particular topic are inappropriately easy or hard.

4. **The physical problems of answer sheets are solved**. Substituted instead are the problems of electronic transmission and storage. Trucks filled with answer sheets do not often overturn in the mud, similarly, networks and machines do not often go down, but even in the best of worlds, such situations occur. Without hard data I suspect that machines crash more often than trucks. However the problems associated with incomplete erasures and inadvertent marks are indeed solved.

5. **The test can be scored immediately**. This remains true, and perhaps, from the point of view of the examinee, is one of the strongest practical justifications for computerized testing. Although it should be remembered that small, portable answer sheet scanners are available that would also allow examinees to scan their answer sheets on the way out of an

exam and get an immediate score. This is certainly a bit clumsier than a CAT, but it is a viable alternative if immediate scoring is the principal benefit desired of CAT. Immediate scoring is critically important for diagnostic tests used within the context of adaptive instruction. However, although widespread use of CATs within an instructional environment are almost certainly inevitable, such uses remains in the future.

6. **Pretesting items can be easily accomplished.** This is accomplished in essentially the same way that pretesting items is done with traditional testing. The key difference is that with traditional testing one must wait until there is a test administration date, whereas with a CAT one must wait until enough examinees have taken the new items to yield statistically stable estimates of the parameters.

7. **Faulty items can be immediately expunged**. Indeed they can, although the sort of two-way communication that Green envisaged has yet to be made operational

8. **A greater variety of questions can be included.** A glance at the tests that have been made operational reveals no items that could not have been administered in paper and pencil format. So the possibility of increased variety has not yet been realized.

At this point it seems sensible say a few kind words about five of the qualities of paper and pencil testing that, lamentably, are lost when a test is made adaptive. These lamentations are based on the current situation. I do not mean to imply that things must remain the way they are, but neither do I mean that remediation will be either quick or easy.

Accessibility in place. There are many fewer places where computer based tests can actually be administered. So instead of the friendly confines of the local high school's gymnasium where you previously could have taken the test along with a cast of thousands, now you must journey 10 or 20 or more miles to a specially designed testing center. This limitation also has differential consequences on inner city and extreme rural examinees. For the latter situation many must factor transportation, hotel and restaurants into the test’s cost.

Accessibility in time. We have only recently rediscovered the wisdom of our predecessors in the testing business who decided to offer an administration of a college admissions test on a Saturday morning in December and in January (the two most popular administration dates for the SAT). When students are given their choice of when to take such a test -- not limited by just those two Saturday mornings -- they typically choose one of those Saturday mornings anyway (although their preference is a little later than the 9AM start time currently in use). There is no mystery why this is as

it is. Students would like to postpone taking the test as long as possible (presumably under the questionable assumption that to do so maximizes the amount they will have learned), but need it included with their college admissions dossier that must be complete by December or January. Add to this that students are busy during the week and typically have other activities scheduled for Sunday, and out jumps the time that most would like to take the test. However facilities to administer computerized tests do not currently have the capacity to accommodate all who would like to take them at this time. This yields the seemingly anomalous result that when tests are offered in P&P format at certain fixed dates, more students can take them exactly when they want.

Accessibility in price. As shown in Figure 2, the cost of taking a computerized test is considerably greater than a paper and pencil one. If one adds on the additional travel expenses associated with having to take a test at a more remote site, this cost differential increases. For some examinees an increased fee can become a serious impediment and thus must be an important issue for all those concerned about building a true meritocracy.

Switching testing formats to CAT usually requires a change in test-taking strategy. Some common test-taking practices on linearly administered tests are generally thought of favorably from an didactic viewpoint, yet cannot be easily implemented with CATs while maintaining the efficiency of evaluation that is CAT's *raison d'etre*. Two such common test-taking strategies lost are:

Response review. Every year throughout their schooling, students are instructed to use the test time constructively. When they have completed the exam they are told to use any remaining time to go back and check over their answers. It is hoped that by doing this the frequency of careless errors can be reduced. In a CAT no such review is possible. Once the answer is given, the next item is chosen on the basis of that response. If a response was to be subsequently changed, the entire string of items that had been selected for that individual may no longer be optimal. Hence review is usually not allowed.

Skip and return. If a student is unsure of the answer to an item on a linearly administered test they are often coached to not waste time on it, but rather to skip it and come back to it later. On most CATs such behavior is not possible.

CONCLUSIONS

CAT, like Marxism and Christianity, seems to work better in theory than in practice. We are a decade into operational CAT, and many of its promises have yet to be fulfilled. Examinees have seen the benefit of immediate scoring, but at a substantially increased cost. In Figure 2 are shown the costs of three large operational tests, the GRE, TOEFL and the SAT. Both the GRE and TOEFL are now administered in CAT forms and have had exponential increases in cost. The SAT is still mass administered in its traditional format and has had only a modest linear increase in its cost over the past half century.

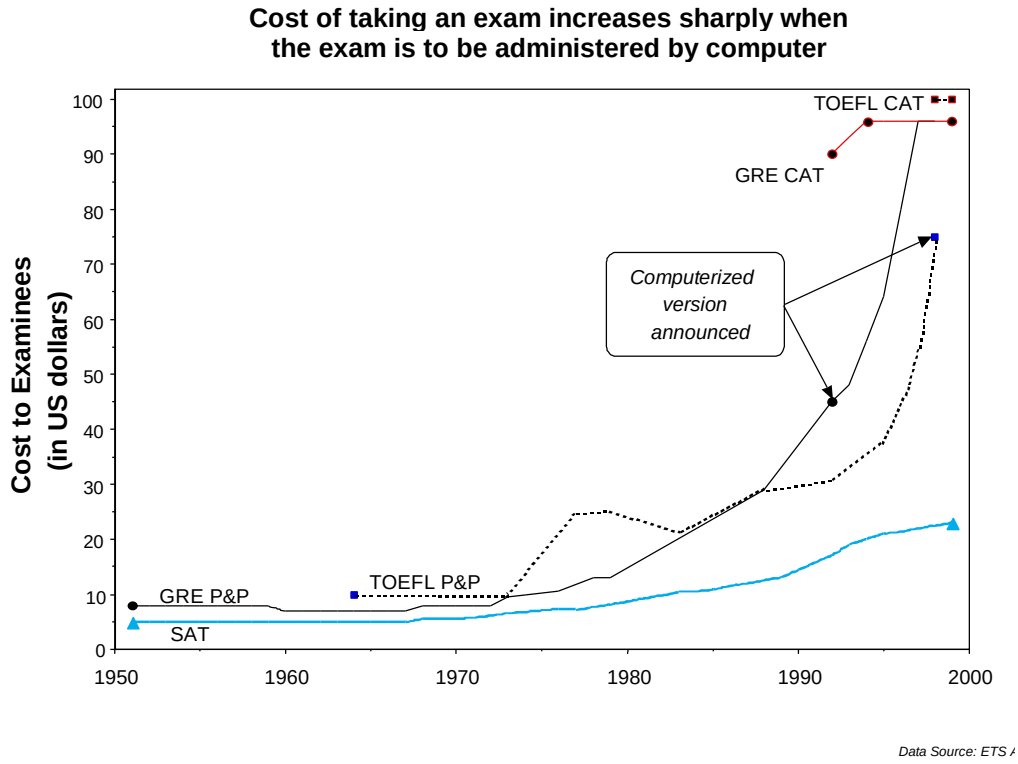

Figure 2. The fees charged to examinees for three large scale testing programs over the past fifty years shows rapid increases for the GRE and TOEFL in the years just before they were computerized. Such exponential increases are absent for the paper and pencil SAT.

Our experience, derived from giving more than three million CATs, is equivocal: it seems like a good idea for some applications and not such a

good idea in others. Wisdom lies in being able to tell one situation from the other. Let me take a crack at providing some guidelines.

Tests should be computerized if the constructs they are trying to measure cannot be assessed easily without the computer; one example might be tests of architectural design that requires a simulation task embedded within a CAD-CAM environment.

Tests can be computerized if it is important to offer the test continuously in time; examples are licensing tests, where a delay means a loss of income for the successful candidate, and the ASVAB, which historically has been offered continuously.

It is currently impractical to offer a computerized test in a mass administration a few times a year. Current economic constraints mean that a computerized test must be offered continuously. Continuous testing offers an enormous security challenge when the tests have high stakes for the examinee. This challenge is difficult to meet even with all of the power and flexibility of CAT; it is nigh onto impossible in paper and pencil format. We must be sure that we need continuous testing before venturing onto this particular minefield. But if we decide that continuous testing is an important feature (and not an annoying consequence) CAT emerges as a sensible option.

Tests can be computerized if it is important for everyone involved to get the right answer; no sane person would cheat on an eye test. Into this third category falls both diagnostic and placement tests. Moreover, the flexibility of CAT fits very well with the aims of both of these kinds of tests. In diagnostic testing a CAT can efficiently zero in on exactly what areas are weak. This diagnosis can help guide instruction; when combined with a matched program of instruction it is called a placement test.

High stakes tests whose results are required only once or twice a year are poor candidates for computerized testing; final exams, Advanced Placement exams, entrance exams all fall into this category.

I believe that the principal reason that the full promise of CAT has yet to be fulfilled is because it has been adopted by large programs that do not need it. Why? Computerizing a test requires a substantial infrastructure. New, innovative tests do not have the volume to support the extensive infrastructure necessary to provide broad access to all examinees who might want to take it. Old, established tests have the volume to support such infrastructure, but usually don’t need to computerize. It is a real chicken and egg problem. Why develop a test that needs to be administered by computer if there is no way to give it? Why build a system to administer computerized tests, when they don't yet exist? A business strategy adopted jump-start the

entire enterprise has been to computerize large tests that don't need to be, and, when the infrastructure is complete use it for new innovative tests yet to be developed. So far, the costs of building the infrastructure have been so great that there have been few resources left over to build the innovative tests that would most benefit from computerization. I have hopes that the next decade will see a change in this[3].

RESUMEN

TAI: Hacia dónde y de dónde. En este trabajo esbozo el estado de la cuestión que dio lugar a los test adaptativos, centrando la discusión en el avance de los TAI y en particular en las expectativas que expresara Bert Green sobre las ventajas de esta tecnología. Se usan datos de la primera década donde los TAI fueron operativos, para comparar lo que ocurrió con aquello que se esperaba que ocurriera. Encuentro que algunas de las metas de TAI expresadas por Green están próximas a cumplirse, aunque la mayoría de ellas quedan para el futuro.

Palabras clave: test adaptativos informatizados, avances en TAI, ventajas de TAI

[3] Initially, to clearly express my pessimism about the future of CAT, I had toyed with the notion of making the title a pun ("Wither CAT?"). But in the end I have convinced myself that there is indeed a rich future for CATs, but only in those areas for which they are well suited. I suspect that in the not very distant future, testing organizations can expect every student to bring a lap top computer into the testing situation, as they now expect #2 pencils. When that happens we can turn away from continuous testing and the security problems that it causes and return to the more efficient mass administrations. Before this can occur testing software must be developed that can use insecure machines without compromising the validity of the testing instrument. I suspect that although such software can be written, determined hackers eventually will be able to break it, so that there will be a continuing effort in this task. But most testing will eventually be done by computer, and if it is, there is no good reason why the tests ought not be adaptive. The future is certainly better described by "whither," not "wither."

REFERENCES

Green, B. F. Jr. (1983). Notes on the efficacy of tailored tests. In *Principals of Modern Psychological Measurement*, (Eds.) H. Wainer & S. Messick , Hillsdale, N.J. : Lawrence Erlbaum Associates.

Irvine, S. & P. Kyllonen, P. (Eds.) (2000). *Item Generation for Test Development.* Hillsdale, N.J. : Lawrence Erlbaum Associates.

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance*, pp. 139-183. New York: Harper and Row.

Lord, F. M. (1971a). The theoretical study of the measurement effectiveness of flexilevel tests. *Educational and Psychological Measurement, 31*, 805-813.

Lord, F. M. (1971b). The self-scoring flexilevel test. *Journal of Educational Measurement, 8*, 147-151.

Lord, F. M. (1971c). Tailored testing, an application of stochastic approximation. *Journal of the American Statistical Association, 66*, 707-711.

Lord, F. M. (1971d). Robbins-Monro procedures for tailored testing. *Educational and Psychological Measurement, 31*, 3-31.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Sands, W. A., Waters, B. K. & McBride, J. R. . (Eds.) (1997). *Computerized Adaptive Testing: From Inquiry to Operation.* Washington, DC: American Psychological Association.

Wainer, H. (2000). Rescuing Computerized Testing by Breaking Zipf's Law. *Journal of Educational and Behavioral Statistics, 25*, xxx-xxx.

Wainer, H., , Dorans, D. J. , Eignor, D., Flaugher, R. , Green, B. F. , Mislevy, R. J. , Steinberg, L. , & Thissen, D. (2000). *Computerized Adaptive Testing: A Primer (2nd edition).* Hillsdale, N.J. : Lawrence Erlbaum Associates.

Zipf, G. K. (1949). *Human Behavior and the Principle of least effort.* Cambridge, MA: Addison-Wesley

.