DR. BERT F. GREEN, JR.
Johns Hopkins University:[1]

Tailored testing has been talked about for many years in academic circles. In this conference we have heard firm plans for action. The promise of tailored testing is becoming real. Numberless simulated examinees have taken tailored tests and a substantial, though smaller, number of real people have also had the experience. The use of tailored tests will provide substantially improved efficiency and will have a number of beneficial side effects, as mentioned by McKillip and Weiss, among others. Testing conditions will be more nearly standardized, the test will hold the taker's interest because each item will be a challenge, possibly there will be less test anxiety, feedback may decrease racial bias. (Weiss; Johnson, 1973)[2]

There will also be some harmful side effects, that we may as well face. People will have trouble understanding the system, and complaints will be frequent. Two people with widely different abilities will both experience getting about half the items rights, yet get very different scores; one will be accepted, the other rejected. If these two people compare notes, they may be confused. The anti-testing forces are also for the most part anti-computer, so negative voices will be raised. Security is at least as difficult with a computer system as with a paper and pencil system. But these are operational problems, and now is not the time to worry about them. They will all be solved, somehow. I merely list them to counter the tendency to believe that the millenium is upon us.

Now let me make one thing perfectly clear. I am about to criticize aspects of the work reported at this conference. That is my job. But the one most important fact, that outweighs all criticism, is this: The operational use of tailored testing is a giant step forward in personnel evaluation. Evidence indicates as much as a 2 to 1 gain in efficiency, and possibly some very important side benefits. I am completely convinced that this is an important step to take. My comments are of two kinds—suggestions for clarifying and improving the theoretical basis for this big step, and impatience at our not yet having planned further giant steps. These steps should be justified not in terms of saving money, which Hansen claims, but in terms of doing a better job.

Let us now consider some of the technical problems in a computer-based system. We have heard two plans for item analysis "on-the-fly", as they say in the computer trade. A question arises about some of the item analysis procedures (Urry; Jensema) which still seem to be built on the biserial correlation of the item with the ability scale, and the overall proportion of correct answers. These raw data are reparameterized (to use an ugly word that should be banned from civilized discourse) but the basic data are $\rho_{\theta i}$ and P_i. Both of these indices depend on the notion of a population of test takers. Yet one purpose of tailored testing is to avoid the notion of population. What, for example, is the population for Lord's broad-range flexilevel test of verbal ability? Everyone from fifth grade to college? In tailored testing, it would seem that the item parameters must be based on the regression of the item on the ability scale. This sounds a little circular—perhaps it is. Some sort of iterative optimization process would be needed at the start, to ever get the ability scale in the first place. Cliff described one such procedure for his ordinal scale model; an equivalent procedure could easily be devised for the metric model.

Cliff's procedure also depends on a population. He goes so far as to say that the purpose of a test is to rank order the population of examinees. Sometimes it is, but often it is not. Often the purpose is to categorize the examinee as qualified or not qualified for a particular job. Or even better, to give a quantitative index of the degree of qualification. The only population we are really interested in is the population of successful job holders.

There are other technical problems with Cliff's scheme, which he promises to solve. For example, he did not describe what happens when a person's item responses have contradictory implications for other cells in his matrix. Indeed his system probably tries to avoid asking questions that might provide contradictory information.

The main reservation I have about the technical side of tailored testing is the commitment to latent trait theory. The concept of a latent ability scale is a great improvement over the concept of a true score. The true score model was never a very good idea; rather, it was a simple model that worked pretty well. But are we sure that the latent ability score is much better? Does the latent trait model fit the tests for which it is used? Is the assumption of local independence really tenable? Suppose, for example, that there are secondary factors in common among subsets of items. How much difference would that make? Nobody knows.

The point is that latent trait theory *is* a theory, just as any other behavioral theory, and it needs verification. Empirical work is needed to show that latent ability scores work as the theory predicts. Simulated examinees will not do—studies are needed with real people. Are the scores invariant over item selections, or over samples of individuals? Does the precision of measurement really work the way the information variable says it does? What about the relation of validity to test length or information? Empirical

[1] This work was done with support from Grant GB37520 from the National Science Foundation. The author is indebted to Warren S. Torgerson for many fruitful discussions of computer applications in testing and personnel decision.

[2] Throughout, references to other papers in this conference are by author only; other references are followed by publication year.

work has been presented by Waters and others, but whether it supports the theory is not clear.

Classical test theory has a curious status: most psychologists and educators believe that it is fact, not theory. Nowhere in Lord & Novick's treatise is there a section on empirical verification of the theory. Actually, test theory is a self-consistent, much-elaborated theory that seems to work pretty well. For example, the Spearman-Brown formula usually works. Some people look upon the Spearman-Brown formula as a fact. It is a fact only in the sense that it is a logical consequence of the basic assumptions of the theory. So far as I know, neither true score theory nor latent trait theory has been put to a critical test, as have most other mathematical theories of behavior.

One final theoretical issue needs clarification. The literature contains results (e.g., Lord, 1970) indicating that a tailored test is not much more effective than an ordinary test with a peaked item difficulty distribution. The advantage lies mainly in the extremes. But the theoretical and empirical results presented in this conference indicate that a tailored test is much better even in the mid-range. Work is needed to clarify when a tailored test will help and when it won't.

One final point about technical terminology. In the simulation studies of Jensema, Waters, McBride, and others, the estimated ability $\hat{\theta}$, which is the test score in tailored testing, is supposed to be nearly θ. The closeness of $\hat{\theta}$ to θ is measured both by $(\Sigma(\theta - \hat{\theta})^2/N)^{1/2}$, which was called the "standard error" and by $r_{\theta\hat{\theta}}$, which was called the "validity". In engineering, the former measure is commonly called the root-mean-square error, or R.M.S. error; it is not, after all, a standard error, since it's not a standard deviation. Mean square error includes both error variance and squared bias. Thus the measure is very appropriate; but it is misnamed. To call $r_{\theta\hat{\theta}}$ the "validity" is much worse; it is downright sinful. This use of the term goes back, I'm told, to Ledyard Tucker and Hubert Brogden, but that only proves that people in high places make mistakes. A different word must be used. "Validity" is seriously misleading, and has even been mis-interpreted at this conference. My own candidate for a name for $r_{\theta\hat{\theta}}$ is "fidelity". I hope the in-group either uses "fidelity" or finds another word.

Next Steps

Now that tailored testing is about to become operational, perhaps it is time to take a longer-range perspective. Do the present developments really exploit the power of an interactive computer? Many scientists, in their first encounter with a computer, use the computer mainly to do faster and neater what they were already doing before computers. It is as if the horse and buggy industry's reaction to internal combustion engines had been to build a mechanical horse. Statistical computation is a good case in point. To a very large extent, statistics is still at the mechanical horse stage in its use of computers. The statistical program packages are fast ways to do old things—analysis of variance, regression, factor analysis. Even the few new things, such as nonmetric scaling and clustering, had their roots in pre-computer ideas. Interactive statistical methods are still in their infancy. Mostly, interaction means replacing the control cards in an input deck by questions printed by the machine and answered by the user on the spot. No subtle interplay of human judgment and computer speed is implied.

The mechanical horse stage in computerized testing would be an automatic test production system. Given the characteristic of a population, the computer would select the most appropriate items from its item files and would print a suitable test. I naively thought testing had avoided this typical first stage, but apparently such systems were built, some years ago.

Tailored testing is one step beyond the mechanical horse stage. To be sure, the up-and-down method had seldom been used in mental testing, barring Binet, who didn't do it right, but the up-and-down method is an old stand-by in psychophysics, and in sensitivity testing generally, dating from World War II and earlier. Also, test theoreticians knew that measurement was best when the items were all sufficiently difficult that the examinee got about half of them correct. (Actually about 68% for 5-alternative items, Fred Lord reminds me, because of guessing.) This is one part of the theory that none of the operational people believed, but the theory was there. So the adaptive test was a natural next step in computer involvement in testing. Still, the only use of the computer in tailored testing, apart from the trivial use in presenting the items on a terminal, is in selecting the next item and computing the ability score. The same 5-choice items are being used, the item is scored either right or wrong, the same kinds of traits are being measured. Now is the time to move on, in research at any rate, to better things.

Many more opportunities exist. Some have been mentioned at this conference. Samejima proposes that we use the particular wrong choice of an item as partial information. Some wrong choices are better than others. Item response weighting has minimal utility in standard tests, primarily because of the test length. Weighting becomes more useful with fewer items, which is just what tailored testing provides. In addition to Samejima's proposal, even more information could be obtained, when the response is wrong, by asking for a second try. The procedure of trying alternatives until getting the right answer goes back to the 1940's or earlier. In those days, Science Research Associates sold a punch board on which answers were punched out. Instructions were to punch out alternatives until the red dot appeared, signalling the right choice. The item score was the number of unpunched choices, except that omits got a negative score. I am told that test scores based on these item scores were consistently more reliable and more valid than scores based on a 1-0 item scoring. The computer terminal is an elegant punch-board! Another possibility is

to have the examinee rank or rate the alternatives for suitability. The probability assignment proposal of Shuford et. al., (1966) now being tried by Weiss and his coworkers is equivalent; though the restriction that the ratings must add to one, like probabilities, is an unfortunate complication that is likely to have adverse operational consequences. Ratings or rankings would be better.

The computer permits the use of constructed responses--fill in the blanks--rather than multiple choice. Computer processing of constructed responses has been worked on in computer assisted instruction; these techniques could be adapted to the testing situation. Most of our present item types have evolved in a multiple choice environment, and constructed responses would be no help. For example, some verbal analogies items would not work as constructed responses – e.g., "Brick is to building as leather is to_____." Others would work: "Shoe is to foot as helmet is to_____." The difficulty of vocabulary items is controlled almost entirely by the distractors, so asking the examinee to construct a synonym would markedly alter the item. But there is no reason why new item types cannot evolve in the new context. Verbal fluency is a natural for the computer to test, and virtually impossible in the multiple choice context.

Of even more interest is the possibility of new types of items, and new types of traits. The GRIP tests of Cory are especially interesting, as are some of the items briefly mentioned by Weiss, such as his conceptual maze. Many of these types can be tried on present day alphanumeric terminals, others need graphic terminals, which are at present too costly, but which may soon be relatively inexpensive.

I am convinced that the potential for new styles of items, or contingent sets of items, is the next important contribution of the computer. After all, we already know how to measure verbal ability and quantitative ability. The computer merely gives us efficiency. What we need is more information.

The computer could also be immensely helpful if we placed less emphasis on measurement and more on the decision process. Instead of providing a test battery, we could provide a decision system. Many years ago Cronbach & Gleser (1965) argued for the necessity of coupling the decision process with the testing process. The computer, and computer assisted testing, have provided an unparalleled opportunity to do this. Hansen, McKillip, & Lord have mentioned this.

Consider the simple example of selecting among applicants for a particular job or for entry to a particular college. The test's job is to label each taker as qualified or not qualified. This implies a cut-off score, or at least a cut-off region. The very well qualified and the very poorly qualified persons can probably be identified relatively quickly; most of the effort should be spent on the borderline cases. To be sure, we must beware of Lord's lucky guesser, and Weiss' low consistency scorer, but with care, an efficient system can be devised that does not measure accurately at all levels, but only where it counts.

A one-dimensional case is only the beginning. Both Weiss and Hansen have suggested that additional savings can be made when there are several relevant dimensions. Here, progress requires that the decision process be coupled with the testing process to build a complete system.

There are many different approaches to a personnel decision system. One model would treat jobs as regions in a space whose dimensions are specific job requirements, specific abilities, or characteristics needed for the job. A person is a point in this space, the testing problem is to pinpoint the person's position sufficiently accurately to be able to list the jobs for which he is qualified, and possibly to list these in rank order from the ones for which he is most qualified to the ones for which he is barely qualified. The dimensions of the job space might be abilities, or they might not. And individual items might serve to locate a person on only one dimension, or items might help to locate a person in the total space. At least, there is no *a priori* reason for discarding impure multidimensional items. Indeed such items might be especially useful in a decision system.

Five years ago at a similar conference (Green, 1970) I said that the computer had a great future in testing. Today, happily, it has a present as well as a future. Operational versions of tailored tests represent a great technical achievement. Furthermore, the computer plays a central role in the enterprise. Still, the potential of the computer has barely been tapped. The future lies ahead.

REFERENCES

Cronbach, L. & Gleser, G. C. *Psychological tests and personnel decisions.* 2nd ed. Urbana, Illinois: University of Illinois Press, 1965.

Green, B. F. Comments on tailored testing. In Holtzman, W. H. (ed.) *Computer-assisted instruction, testing, and guidance.* New York: Harper & Row, 1970.

Johnson, D. F. and Mihal, W. L. Performance of blacks and whites in computerized vs. manual testing environment. *American Psychologist,* 1973, *28,* 694-699.

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

Shuford, E. H., Albert, A., and Massengill, H. E. Admissable probability measurements procedures. *Psychometrika,* 1966, *31,* 125-145.

ANNOUNCEMENTS

Dr. Robert J. Gettelfinger of Educational Testing Service announced that organization's willingness to edit a newsletter on the subject of computer-assisted testing. He asked for suggestions as to the content of the newsletter, and for

the opinions of the conferees as to what subject matter should be covered and as to whether contributions should be entirely voluntary or should be obtained by assigning papers.

Dr. David J. Weiss of the University of Minnesota announced that he will edit a new journal, *Applied Psychological Measurement*, that will publish empirical research on the application of techniques of psychological measurement to substantive problems in all areas of psychology and related disciplines such as sociology and political science. He invited conference participants to submit their papers and promised to send further details to all participants.