Some Item Analysis and Test Theory for a System of Computer-Assisted Test Construction for Individualized Instruction

Frederic M. Lord
Educational Testing Service

Under given conditions, conventional testing and computer-generated repeatable testing *(CGRT)* are equally effective for estimating examinee ability; *CGRT* is more effective than conventional testing for estimating the mean ability level of a group; and *CGRT* is less effective for estimating ability differences among individuals. These conclusions are drawn from domain-referenced test theory as distinguished from norm-referenced test theory.

There are various ways in which computers are used to aid in the construction of tests (see Lippey, 1974). Under a procedure used for individualized instruction in certain colleges and universities, a student contacts the computer (or its designated agent) when he/she wishes to take a test. The computer selects a limited number of test items to be administered to this student from a large, stored pool of items. In the case considered here, item selection is random or random within a stratum. Prosser and Jensen (1971) and Emerson (1974), among others, have discussed this system of testing and labeled it computer-generated repeatable testing *(CGRT)*.

In general, there must be little overlap between the items administered to any two students or to the same student on two different testings. This is necessary in order to prevent a student from obtaining a high score simply by memorizing the scoring key for a test taken by a friend. This safeguard makes it necessary for the computer to have a pool available which contains many items—perhaps 10 to 40 items for each item chosen for administration to a given student at a given time. In this situation, any one item is administered to only one-tenth to one-fortieth of the examinees. For any given pair of items, relatively few students will likely be presented with both items.

Baker (1974) discusses the need for item analysis and item statistics to be used with computer-assisted test construction. He points out the difficulties arising when two different examinees, for the most part, take different sets of items and mentions the possibility of applying the theory of randomly parallel tests (Lord & Novick, 1968, chap. 11) to this problem. The present paper uses the theory of randomly parallel tests (or theory of item sampling) to evaluate an obvious, but not obviously effective, method of scoring and item analysis in the *CGRT* situation already described.

As ordinarily used, item-sampling theory is a domain-referenced theory, as distinct from a norm-referenced theory. If some of the conclusions reached are unexpected, it is because we are accustomed to thinking in terms of norm-referenced test theory.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 1, No. 3 Summer 1977 pp. 447–455

Item Difficulty

It will be convenient to discuss first a simple and familiar problem—the estimation of item difficulty. In conventional testing and item analysis, the same number N examinees answer each item. If these N examinees are a random sample from some population of examinees, the observed proportion p_i of correct answers to dichotomously scored item i is an unbiased estimate of the corresponding proportion π_i, the *item difficulty* in the population of examinees.

Over random samples from an infinite population of examinees, Np_i has the familiar binomial distribution. The expected value of p_i is π_i and the standard error of p_i is

$$S.E.(p_i) = \sqrt{[\pi_i(1 - \pi_i)/N]} \quad . \qquad [1]$$

This is true even though different examinees have different chances of success on the item. Since most critical readers believe this last statement to be false, an illustrative example is offered. Suppose there are just three different types of examinees, denoted by *A, B, C.* In the infinite population, these three types occur with equal frequency. On a given item *i,* the chance of success is .1 for examinees of Type *A,* .5 for Type *B,* and .6 for Type *C.* Samples of just $N = 2$ examinees are drawn at random. In any sample, the estimated item difficulty p_i equals the proportion of correct answers. Since $N = 2$, $p_i = 0$, .5 or 1. The left column of Table 1 shows the nine equally likely possible samples of $N = 2$ examinees. For each sample of examinees, the three columns to the right show the probability that $p_i =$ 0, .5, and 1. The bottom row of the table is obtained by averaging the numbers above it. Thus, it shows the probability across all samples that $p_i = 0, .5, 1$. It is now seen that these overall probabilities follow the binomial distribution: $(1 - \pi_i)^2$, $2\pi_i(1 - \pi_i)$, and π_i^2, with $\pi_i = (.1 + .5 + .6)/3 = .4$.

Suppose, instead, that there are three times as many of Type *A* examinees as there are of Type *B* or *C* in the population. This means that the rows of Table 1 containing *A* are three times as frequent as the others—except for the top row, which is nine times as frequent. When the probabilities in the body of the table are weighted by these new frequencies before being averaged, the overall probabilities (last line of the table) are found to be .5184, .4032, and .0784. This is the binomial distribution with $\pi_i = (3 \times .1 + .5 + .6)/5 = .28$.

The reader may use other numerical examples, which in this case may be more convincing than mathematical arguments. A logical proof is as follows: It will make no difference in the final sampling distribution whether we first choose N examinees at random from an infinite population of examinees and then secure the responses of these N examinees to the given item or whether we first secure the response of each examinee in the population to the given item and then draw N responses at random from the resulting infinite population of responses. The former procedure is what we usually think of doing; the latter, effectively identical procedure clearly produces a binomial distribution of successes.

In *CGRT,* each item is administered to a random sample of examinees. The mean, standard error and distribution of p_i are thus formally the same as in conventional testing. However, the number N of people who take a given item is likely to be much smaller in *CGRT* than in conventional testing.

In theory, one could try to adjust the p_i for different items to take into account the fact that the examinees who happen to take item i differ slightly in ability from those who take item j. The adjustment in p_i would be of order $1/\sqrt{N}$. Since $S.E.(p_i)$ is also of order $1/\sqrt{N}$, the improvement in $S.E.(p_i)$ would only be of order $1/N$. Such small adjustments will not be considered here.

Item Sampling

A pool of items is available to the computer. Statistical analysis here could be carried through for a finite pool, but for simplicity we will assume the pool to be infinite.

Table 1

Illustrative Distribution of Sample Item Difficulty Across Samples

Sample of Examinees	$p_i = 0$	.5	1
(A,A)	.81	.18	.01
(A,B)	.45	.50	.05
(A,C)	.36	.58	.06
(B,A)	.45	.50	.05
(B,B)	.25	.50	.25
(B,C)	.20	.50	.30
(C,A)	.36	.58	.06
(C,B)	.20	.50	.30
(C,C)	.16	.48	.36
Overall Probability	.36	.48	.16

Large pools are often constructed by writing 10 to 40 items on each piece of information or instructional goal to be tested. Usually, the computer constructs a test by selecting (at random) one and only one item on each piece of information or instructional goal. This is known as stratified random sampling, and it reduces sampling fluctuations below what would occur under simple random sampling. For simplicity, the present report deals only with the simple case, where each item is selected at random from the entire pool. Under stratified sampling of items, the sampling errors would be smaller than found here.

We will want to compare the errors of measurement (sampling errors) that we find for *CGRT* with those found in more conventional testing. If the same random sample of n items is administered to all examinees, we will follow Cronbach, Gleser, Nanda, and Rajaratnam (1972), borrowing a term from analysis of variance and referring to this situation as the *crossed* case. If a different random sample of n items is administered to each examinee, we will refer to this as the *uncrossed* (or *nested*) case. Cronbach et al. give a thorough discussion of the application of analysis of variance to crossed and to uncrossed data of the kind considered here.

It may be helpful to locate our uncrossed case within the general field of item and matrix sampling. Our uncrossed case differs from simple item sampling and from simple matrix sampling because in our case, each person gets a different sample of items. Our uncrossed case is subsumed under multiple-matrix sampling (Lord & Novick, 1968, Section 11.12), in which different samples of rows (items) are administered to different samples of columns (examinees). Furthermore, our case is a special case of multiple-matrix sampling in that for us, the number of examinees in each sample of examinees is just one, so that our matrix sample is really a vector sample.

Test Score

How useful will number-right scores be for measuring examinees when, as assumed here, each examinee takes a different random sample of n items? No further assumptions are needed to conclude that x_a, the number-right score of examinee a, has the binomial distribution

$$f(x_a) = \binom{n}{x_a} \zeta_a^{x_a} (1 - \zeta_a)^{n-x_a} \qquad [2]$$

where ζ_a is the *proportion-correct true score* of examinee a. (This result is discussed in detail by Lord & Novick, 1968, Section 11.9.) The true score ζ_a may be thought of as the proportion of items the examinee would answer correctly in the entire infinite population of items. Cronbach et al. call it the universe score. Tryon (1957), Millman (1974), and workers in criterion-referenced testing call it the domain score.

The mean of this distribution is

$$\xi_a \equiv n\zeta_a \qquad [3]$$

and the variance is

$$\mathrm{Var}_I x_a \equiv \mathrm{Var}_I(x_a \mid \zeta_a) = n\zeta_a(1 - \zeta_a) \quad , \qquad [4]$$

where the subscript I denotes that the variance is taken over all random samples of items. Again, all this is true regardless of the fact that different items are of different difficulty for examinee a. Table 1 will provide a numerical example, provided we interchange people and items.

Let $e_a \equiv x_a - \xi_a$ denote the *error of measurement* for examinee a. Since ξ_a is constant for examinee a, $Var_I e_a$ (the squared *standard error of measurement* for examinee a) is the same as $Var_I x_a$

$$\mathrm{Var}_I e_a \equiv \mathrm{Var}_I(x_a \mid \zeta_a) = n\zeta_a(1 - \zeta_a) \quad . \qquad [5]$$

The standard error of measurement (see Equation 4) for an individual is the same whether all individuals are tested with the same random sample of n items (crossed case) or with a different random sample of n items (uncrossed case). Neither crossed nor uncrossed testing has any advantage for the measurement of single individuals.

The following section nevertheless demonstrates the counterintuitive result: for determining the mean score of a group of individuals, one method is distinctly better than the other.

Group Mean Test Score

We are concerned with the sampling variance of the mean score for a fixed group of N examinees across all samples of items (note that we are sampling items, not people):

$$\mathrm{Var}_I \bar{x} = \frac{1}{N^2} \mathrm{Var}_I \sum_{a=1}^{N} x_a = \frac{1}{N^2} \sum_{a=1}^{N} \sum_{b=1}^{N} \mathrm{Cov}_I(x_a, x_b) \quad . \qquad [6]$$

In the uncrossed case, the scores of examinees a and b are independent, since each takes an independent sample of items, so that for $a \neq b$, $Cov_I(x_a, x_b) = 0$. Thus, for the uncrossed case

$$\mathrm{Var}_I \bar{x} = \frac{1}{N^2} \sum_{a=1}^{N} \mathrm{Var}_I x_a = \frac{n}{N^2} \sum_{a=1}^{N} \zeta_a(1 - \zeta_a) \quad . \qquad [7]$$

In the crossed case, if the n items in the test are difficult items, x_a and x_b are likely to be low; if they are easy items, x_a and x_b are likely to be high. Thus, for random sampling of items in the crossed case, typically $Cov_I(x_a,x_b) > 0$, unless all items in the infinite pool are of equal difficulty.

In the crossed case, $Var_I\bar{x}$ is larger than in the uncrossed case, unless all items are of the same difficulty. In the uncrossed case (Equation 7), $Var_I\bar{x}$ vanishes for large N. This does not happen in the crossed case when the items are of unequal difficulty, because the right side of Equation 7 is simply the average of the N^2 terms in the summation, and these are typically greater than zero. The uncrossed *CGRT* method of testing discussed here is better for estimating the average performance of a large group than is the more conventional method of giving the same test to all examinees.

In the uncrossed case, $x \equiv \Sigma_a x_a/N$ is a function of N independently distributed random variables, x_a. As is usual in such cases, the sampling variance of $\bar{x}$ is of order $1/N$. In the crossed case, x_a is *not* distributed independently; if n difficult items are chosen, all x_a tend to be lower than if n easy items are chosen. It is this lack of independence in the crossed case that prevents the sampling variance of $\bar{x}$ from vanishing for large N.

The variance for the crossed case is easily written down directly from the fact that

$$\bar{x} = \sum_{i=1}^{n} p_i = n\bar{p}_i \quad . \qquad [8]$$

The variance is seen to be

$$\mathrm{Var}_I\bar{x} = n\,\mathrm{Var}_I p \quad , \qquad [9]$$

where $Var_I p$ is the variance across all items in the pool of the item difficulty statistic p_i (the proportion of correct answers to item i). There is no simple expression for the difference between the variances represented by Equations 7 and 9, except as already discussed in connection with Equation 6.[1]

The advantage of *CGRT* over conventional testing for estimating the mean level of performance of a group is a result of the following fact: the total pool of items is much better represented in the data when each examinee takes a different sample of n items than when all examinees take the same sample of n items.

Individual Differences

Suppose we do not need to measure individuals on an absolute scale, but only to compare examinees. How accurately can we estimate differences between individual true scores?

Clearly, $x_a - x_b$ is an unbiased estimator of $\xi_a - \xi_b$. Now consider the sampling variance

$$\mathrm{Var}_I(x_a - x_b) = \mathrm{Var}_I x_a + \mathrm{Var}_I x_b - 2\mathrm{Cov}_I(x_a, x_b) \quad . \qquad [10]$$

As already noted, $Cov_I(x_a,x_b)$ is zero in the uncrossed case, so for that case:

$$\mathrm{Var}_I(x_a - x_b) = n(\zeta_a - \zeta_a^2 + \zeta_b - \zeta_b^2) \quad . \qquad [11]$$

---

[1]Some perspective on our problem is obtained by noting that Equation 7 for the standard error of a group mean could be obtained from the general formula for the standard error of a mean under multiple-matrix sampling given by Lord and Novick (1968, Equation 11.12.3). Our uncrossed case is obtained by replacing their N by 1, M by N, $\bar{N}$ by N, and letting their $\bar{n} \to \infty$. Our crossed case (Equation 9) can be obtained by replacing their M by 1, $\bar{N}$ by N, and letting their $\bar{n} \to \infty$.

If the items in the pool are not all of equal difficulty, $Cov_I(x_a,x_b) > 0$ in the crossed case, so that $Var_I(x_a - x_b)$ is *smaller* than in the uncrossed case.

So far, we have the following conclusions: (1) crossed and uncrossed measurement are equally effective for measuring a single individual; (2) crossed measurement is better for measuring differences between individuals; (3) uncrossed measurement is better for determining the absolute level of performance of a group.

Item-Test Correlation

The main products of an item analysis are often an item difficulty statistic and an item-test correlation for each item. The latter, to be denoted by r_{gx}, is the correlation between u_g, the score on item g, and x, the number-right test score. To avoid a spurious relationship, item g will be excluded from the items used to find x.

By a standard formula

$$r_{gx} \equiv s_{gx}/s_g s_x \quad , \qquad [12]$$

where s_{gx} is a covariance, s_g and s_x are standard deviations. In the conventional (crossed) case, each examinee's number-right score is based on the same set of n items. What is the effect on r_{gx} if each examinee takes a different set of n items (uncrossed case)? Does this greatly lower the item-test correlation or increase its standard error?

Let us answer these questions for the numerator of Equation 12, the covariance,

$$s_{gx} = \frac{1}{N} \sum_{a=1}^{N} (u_{ga} - p_g)(x_a - \bar{x}) \quad , \qquad [13]$$

where u_{ga} denotes the score of examinee a on item g. Since $\Sigma_a(u_{ga} - p_g) = 0$, Equation 13 can also be written

$$s_{gx} = \frac{1}{N} \sum_{a=1}^{N} (u_{ga} - p_g)x_a \quad . \qquad [14]$$

The responses $u_{ga}(a = 1, 2, \ldots, N)$ to item g are considered given and fixed in this case. The only sampling fluctuations arise from the random sampling of the items used to find x_a. The effects of sampling examinees are not considered here. Expectations, to be denoted by ε_I, are to be taken over all possible random samples of items, excluding item g. Thus,

$$\varepsilon_I s_{gx} = \frac{1}{N} \sum_{a=1}^{N} (u_{ga} - p_g)\xi_a = s_{g\xi} \quad . \qquad [15]$$

The expectation of the item-test covariance is equal to the covariance between the item and the true score. (Note that in an infinite population of items, ξ is not changed by removing item g from the population.)

This result is equally true for the crossed case and for the uncrossed case. We see that neither method of testing introduces any bias into the item-test covariance.

Our next concern is with the sampling variance of s_{gx}:

$$\mathrm{Var}_I s_{gx} = \mathrm{Var}_I\left[\frac{1}{N}\sum_{a=1}^{N}(u_{ga} - p_g)x_a\right]$$

$$= \frac{1}{N^2}\sum_{a=1}^{N}\sum_{b=1}^{N}(u_{ga} - p_g)(u_{gb} - p_g)\mathrm{Cov}_I(x_a, x_b) \quad . \qquad [16]$$

Since $Cov_I(x_a,x_b) = 0$ whenever $a \neq b$ for the uncrossed case, by Equations 16 and 4

$$\mathrm{Var}_I s_{gx} = \frac{1}{N^2}\sum_{a=1}^{N}(u_{ga} - p_g)^2 \mathrm{Var}_I x_a$$

$$= \frac{n}{N^2}\sum_{a=1}^{N}(u_{ga} - p_g)^2 \zeta_a(1 - \zeta_a) \quad . \qquad [17]$$

Note that this variance decreases as N increases.

For the crossed case,

$$\mathrm{Var}_I s_{gx} = \mathrm{Var}_I s\left(u_g, \sum_i u_i\right) = \mathrm{Var}_I \sum_{i=1}^{n} s_{gi}$$

$$= \sum_{i=1}^{n} \mathrm{Var}_I s_{gi} = n\,\mathrm{Var}_I s_{gi} \quad , \qquad [18]$$

where s_{gi} denotes the inter-item covariance over N examinees between u_{ga} and u_{ia}—the responses to items g and i respectively. $Var_I s_{gi}$ is the variance of this covariance taken over all items in the infinite pool, item g being fixed.

The main point to note is that in the uncrossed case (Equation 17) the sampling variance of s_{gx} becomes small as N becomes large. In the crossed case (Equation 18), this is not true.

It can be shown that in the uncrossed case, the denominator as well as the numerator of Equation 12 has a sampling variance that vanishes for large N. Thus, assuming the score variance s_x to be nonzero, the sampling variance of the item-test correlation vanishes for large N in the uncrossed case but not in the crossed case.

When N is given and large enough, the *CGRT* procedure estimates item-domain correlation better than the crossed procedure. Offsetting this in practice, however, is the fact that the available N for the *CGRT* procedure is likely to be much smaller than in conventional testing.

Sample Estimates of Sampling Variances

The main concern so far has been to compare statistics obtained from the unconventional, uncrossed *CGRT* testing procedure with statistics obtained from conventional or "crossed" testing. The formulas developed for various sampling variances are not in a form useful for practical workers. The formulas below, presented without proof, give sample estimators of the sampling variances of interest. These estimators are unbiased in item sampling, as indicated by the symbol $\hat{=}$, to be read "is estimated without bias by."

In both crossed and uncrossed testing,

$$\mathrm{Var}_I x_a = \mathrm{Var}_I e_a \hat{=} \frac{x_a(n - x_a)}{n - 1} \quad . \qquad [19]$$

In uncrossed testing,

$$\mathrm{Var}_I \bar{x} \hat{=} \frac{1}{N^2(n - 1)} \sum_{a=1}^{N} x_a(n - x_a) \quad ; \qquad [20]$$

$$\mathrm{Var}_I s_{gx} \hat{=} \frac{1}{N^2(n - 1)} \sum_{a=1}^{N} (u_{ga} - p_g)^2 x_a(n - x_a) \quad . \qquad [21]$$

In crossed testing,

$$\mathrm{Var}_I \bar{x} \hat{=} \frac{n^2 s_p^2}{n - 1} \quad ; \qquad [22]$$

$$\mathrm{Var}_I s_{gx} \hat{=} \frac{n^2 s^2(s_{gi})}{n - 1} \qquad [23]$$

where $s_p^2 \equiv \Sigma_i (p_i^2/n) - \bar{p}^2$ and

$$s^2(s_{gi}) \equiv \frac{1}{n} \sum_{i=1}^{n} s_{gi}^2 - \left(\frac{1}{n} \sum_{i=1}^{n} s_{gi} \right)^2 \quad . \qquad [24]$$

These formulas do not assume that n or N is large. Items are assumed to be sampled from an infinite pool. The N examinees are *not* considered as sampled but rather as fixed.

Numerical Example

For one set of test data at hand, $N = 1521$, $n = 41$, $\bar{x} = 30.9$, $s_x = 5.7$, $s_p = .167$. For the crossed case, using Equation 22, the estimated standard error of x is 1.1. If each set of n items were a random sample from an infinite pool of items, the estimated standard error of $\bar{x}$ for the uncrossed case would, by applying Equation 20, be .07.

The difference between 1.1 and .07 illustrates the strong advantage of item sampling for estimating a mean. This advantage arises from the fact that the pool of items is much better represented when each examinee takes a different set of 41 items than when only one set of 41 items is used for everyone.

As discussed earlier, the formulas derived here assume an infinite pool of items. Thus, the uncrossed case would require administering $41 \times 1{,}521 = 62{,}361$ different items. For the data at hand, only 1,033 different items were actually available for administration. Since $n = 41$ is small compared to 1,033, little is lost by assuming an infinite pool of items.

The estimated standard error of the item-test covariance was computed for five different items for the crossed case (Equation 23) and for the uncrossed case (Equation 21). The results are given in Table 2. For these data, the uncrossed procedure gives smaller sampling variances than the crossed procedure.

Table 2

Estimated Standard Errors of Item-Test Covariance

Item Designation	Standard Errors	
	Crossed	Uncrossed
68	.39	.24
69	.31	.19
70	.37	.33
71	.35	.32
72	.56	.45

In the uncrossed case, the standard error depends on the number of examinees taking the item. About 190 examinees took each item in Table 2. If the number had been smaller, the uncrossed standard errors would have been larger.

References

Baker, F. B. The role of statistics. In G. Lippey (Ed.), *Computer-assisted test construction.* Englewood Cliffs, NJ: Educational Technology Publications, 1974.

Cronbach, L. J., Gleser, G.C., Nanda, H., & Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley, 1972.

Emerson, P. L. Experience with computer generation and scoring of tests for a large class. *Educational and Psychological Measurement,* 1974, *34,* 703–709.

Lippey, G. (Ed.). *Computer-assisted test construction.* Englewood Cliffs, N.J: Educational Technology Publications, 1974.

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

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), *Evaluation in education.* Berkeley, CA: McCutchan, 1974.

Prosser, F., & Jensen, D. D. Computer generated repeatable tests. *AFIPS Conference Proceedings,* 1971, *38,* 295–301.

Tryon, R. C. Reliability and behavior domain validity: Reformulation and historical critique. *Psychological Bulletin,* 1957, *54,* 229–249.

Acknowledgment

Research reported in this paper has been supported by grant GB-41999 from National Science Foundation. The author wishes to thank Frank B. Baker, Gerald Lippey, Frederick R. Kling, and David M. Shoemaker for their helpful comments on a draft of this manuscript; and Robert Eastman for the data used for the numerical example.

Author's Address

Frederic M. Lord, Educational Testing Service, Princeton, NJ 08540